ALIGNING MULTIMODAL PATIENT EVIDENCE WITH BIOMEDICAL KNOWLEDGE GRAPHS FOR CLINICAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical questions often depend on linking a patient's multimodal evidence to external biomedical knowledge, yet existing predictive systems rarely represent such links explicitly, so they can neither be traced to their evidence sources nor removed to measure their contributions. We present MM-KG (**M**ulti**m**odal **K**nowledge **G**raph), which represents heterogeneous, multimodal patient observations and biomedical concepts as separate layers in one typed graph, joined by explicit alignment edges. First, modality-specific harmonizers convert EHR text, imaging, genomic, and biospecimen data into typed observations mapped to UMLS concepts, which a route-prioritized aligner links to a biomedical knowledge graph. Query-conditioned retrieval then selects a compact subgraph for downstream use by a large language model or a graph neural network. We build MM-KGs for MIMIC-IV and the Alzheimer's Disease Neuroimaging Initiative (ADNI) and evaluate them with a 2×2 design that separates patient evidence, biomedical knowledge, and their interaction. On questions that require both sources, neither source alone performs far above chance, whereas their combination yields a drug-controlled AUROC interaction of $+0.194$ on MIMIC and $+0.299$ on ADNI. On held-out five-candidate ranking, MM-KG outperforms MindMap by $+0.131$ Hits@1 and leads an adapted GraphCare on the items that require consulting the patient, and deleting the single answer-bearing relation from the retrieved packet returns Hits@1 to the no-knowledge baseline. Finally, query-conditioned retrieval reaches 0.731 AUROC with $6.8\times$ less context than the strongest generic policy, whereas static knowledge graph context gives no consistent gain on ordinary outcome prediction. Knowledge graphs thus benefit clinical LLMs not as background context but as explicit links between multimodal patient evidence and the relation a question requires, and MM-KG makes these links retrievable, traceable, and testable.

1 INTRODUCTION

Reasoning about an individual clinical case requires connecting heterogeneous patient observations to general biomedical knowledge. Patient records describe *observations*, eg, laboratory measurements, images, reports, medications, longitudinal events (Johnson et al., 2023; 2019b); whereas biomedical knowledge bases describe general relations among diseases, drugs, genes, and phenotypes (Chandak et al., 2023; Bodenreider, 2004; Stear et al., 2024). Neither is sufficient alone when a question depends on a link between the two, and such questions are common: across 28,582 MIMIC-IV ICU stays, 29.9% exhibit a laboratory-evident organ state that never appears in the coded diagnoses while the patient receives a medication that PrimeKG lists as contraindicated for that state (App. B). Answering such questions correctly requires reasoning about the current measurement *and* the drug-specific relation; either source in isolation leaves the question undecidable.

Existing works model each of these sides, but rarely their connection. Multimodal clinical models fuse EHR data, text, and images into predictive representations, but discard the links from individual observations to biomedical concepts (Hayat et al., 2022; Yao et al., 2024b;a). Knowledge-graph methods supply external facts or build patient-specific graphs for prediction and language-model reasoning (Jiang et al., 2024; 2025; Wen et al., 2024; Wu et al., 2025). Their patient side, however,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is typically derived from structured codes or clinical text alone; imaging, genomic, and biospecimen evidence is either excluded or fused outside the graph, so no single representation carries all modalities through to the biomedical knowledge backbone. Where multiple modalities are represented graphically, they are generally not linked to standardized biomedical concepts (Wu et al., 2024). Consequently, no explicit path runs from a source observation, through its normalized concept, to the biomedical relation used to interpret it, and the connection cannot be retrieved or ablated.

We introduce MM-KG (**M**ulti**m**odal **K**nowledge **G**raph), which represents multimodal patient evidence and a biomedical knowledge backbone in a single typed graph. Each modality passes through its own harmonizer, converting raw input into typed observation nodes rather than a fused vector. A vision-language model turns chest radiographs into calibrated textual findings; brain MRIs are segmented into regional abnormality findings under a normative model; and codes, measurements, and report text are resolved by vocabulary lookup, value-aware rules, and entity linking. Provenance edges retain each observation's source and timestamp, and a route-prioritized aligner links the resulting concepts to PrimeKG (Chandak et al., 2023), preferring exact identifier routes over synonym and parent hops (§3.3). Because instance and ontology nodes keep separate identities, each layer can be traced or removed independently. Query-conditioned retrieval then assembles a compact subgraph per question, joining the relevant patient neighborhood to candidate-conditioned domain relations under deterministic rules with no access to the label. The result can then be serialized for a downstream LLM or used as input to a GNN, exposing a path from source observations, to aligned concepts, to the relation supporting an answer (Fig. 1). This explicit structure makes the patient evidence, biomedical knowledge, and their alignment separately traceable and testable.

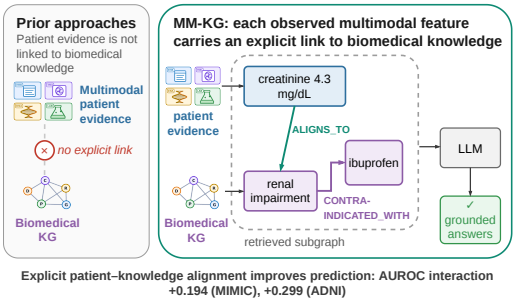


Figure 1: **Motivation of MM-KG.**

We build two independently constructed graphs: 51,844 MIMIC-IV patients with chest radiographs and reports, and 1,000 ADNI participants spanning longitudinal imaging, history, cognition, genetics, and biospecimens (Veitch et al., 2022). Because instance and ontology layers remain separately addressable, a 2×2 design can isolate patient evidence, KG relations, and their interaction (§3.5), a decomposition which fused representations do not support. Our primary evaluation targets questions in which neither source is sufficient alone. On 3,200 controlled MIMIC questions that link a condition documented in laboratory values or radiology reports to a drug relation, each source alone leaves performance near chance, whereas together they yield a drug-controlled AUROC interaction of $+0.194$ ($p<0.001$), reproduced on the 70-item ADNI test set at $+0.299$ across a different disease domain and evidence channel. On held-out five-candidate ranking, MM-KG exceeds MindMap by $+0.131$ Hits@1 and leads an adapted GraphCare on the items a candidate-only prior cannot resolve (0.582 against 0.375), while deleting the answer-bearing relation, replacing or shuffling packets reduces Hits@1 from 0.575 to 0.185–0.299 against 0.200 chance. The effect therefore requires the specific correspondence that alignment makes explicit, which query-conditioned retrieval supplies at AUROC 0.731 with 1,206 input tokens, against 8,157 for the strongest generic policy. Where one source suffices, each behaves as expected: patient measurements drive mortality prediction for both consumers while static PrimeKG adds nothing measurable (App. I), and the same relations improve biomedical fact-checking by $+0.063$ to $+0.420$ (App. Table 7).

Our contributions are summarized as follows. **(1)** A typed instance–knowledge graph construction framework in which multimodal observations, their provenance, and their alignment to a biomedical backbone remain separately addressable, so that each layer can be traced or ablated rather than fused away. **(2)** Query-conditioned retrieval that joins patient evidence to candidate-conditioned relations, recovering the answer path for 99.7% of positive items and reaching AUROC 0.731 with $6.8\times$ less context than the strongest vector, full-context, or typed one-hop policy. **(3)** A factorial evaluation showing that patient evidence and KG relations combine superadditively on questions that require both, yielding interactions of $+0.194$ (MIMIC) and $+0.299$ (ADNI) AUROC, and that removing or corrupting the answer-bearing alignment returns held-out ranking to the no-

knowledge level or below, whereas static KG context alone does not improve outcome prediction. MM-KG construction, retrieval, benchmark, and evaluation code will be released upon acceptance.[1]

2 Related work

Multimodal clinical representation. Clinical representation learning has progressed from longitudinal EHR models such as RETAIN (Choi et al., 2016) and GRAM (Choi et al., 2017) to multimodal architectures that combine structured records, text, and medical images. MedFuse (Hayat et al., 2022), DrFuse (Yao et al., 2024b), and DDL-CXR study fusion (Yao et al., 2024a) under missing or asynchronous modalities. MedGTX (Park et al., 2022) jointly encodes structured EHR graphs and clinical text, while MUSE (Wu et al., 2024) represents patient–modality relationships with a bipartite graph and evaluates on both MIMIC and ADNI. MedMod (Elsharief et al., 2025) provides a benchmark for multimodal prediction using MIMIC-IV and MIMIC-CXR. These methods are important prediction baselines, but they do not explicitly preserve the paths from source observations through normalized biomedical concepts to external biomedical relations.

Patient representations and biomedical KGs. Multimodal KGs associate symbolic entities with heterogeneous attributes or evidence (Chen et al., 2024). Biomedical resources such as PrimeKG and PetaGraph (Chandak et al., 2023; Stear et al., 2024) provide large concept-level backbones but do not by themselves represent a patient's multimodal clinical history. Several methods enrich patient or code representations with biomedical knowledge. MMUGL (Burger et al., 2023) maps structured codes and concepts extracted from clinical text into a shared UMLS-derived representation; ClinVec and MedTok combine standardized codes with textual or relational knowledge (Johnson et al., 2026; Su et al., 2025);and GraphCare constructs patient-specific graphs using external biomedical resources (Jiang et al., 2024). KG4Diagnosis proposes a hierarchical multi-agent framework that constructs a graph from unstructured medical text for diagnostic reasoning (Zuo et al., 2025). These studies show the value of combining patient records with biomedical knowledge; MM-KG builds on this further by retaining source-level multimodal observations, recording how each was derived, and linking the resulting concepts explicitly to a shared biomedical backbone.

Knowledge retrieval for clinical prediction. Retrieval-augmented generation (RAG) (Lewis et al., 2020) retrieves external evidence before producing an answer. In clinical prediction, RAM-EHR retrieves relevant information from external knowledge sources (Xu et al., 2024), while KARE organizes knowledge into communities and retrieves community-level evidence for healthcare prediction (Jiang et al., 2025). RAD retrieves disease-centered guideline knowledge and uses it to align multimodal features for clinical diagnosis (Li et al., 2025). KARE and RAD are especially relevant comparators, but they use retrieved knowledge primarily to construct summaries or guide latent representations. In contrast, MM-KG retrieves an explicit typed subgraph spanning both patient evidence and biomedical relations, which can be consumed by an LLM or a heterogeneous GNN.

KG retrieval and verification for medical LLMs. Several methods use biomedical graphs to support medical language models. MindMap (Wen et al., 2024) retrieves graph evidence for diagnostic reasoning, and DR.KNOWS (Gao et al., 2023) incorporates UMLS knowledge into diagnosis prediction. DOSSIER checks claims against structured patient records and biomedical knowledge graphs while limiting the exposure of private records to the LLM (Zhang et al., 2024). FAITH instead evaluates generated medical claims by decomposing responses into atomic statements and scoring their support using paths in a biomedical graph (Zhou et al., 2026). MedGraphRAG (Wu et al., 2025) is the closest prior system: it builds a hierarchical graph over MIMIC-IV, anchors entities to standardized vocabularies, retains source attribution, and conditions retrieval on the query. It operates over document chunks, so attribution is recorded per passage rather than per observation, and its patient side carries no laboratory values, imaging, or genomics. MM-KG records provenance for each typed observation across modalities and makes each observation–concept link a separate, removable edge. Our experiments then address a complementary question: whether separately controlled patient evidence, external biomedical relations, and their explicit alignment improve prediction and grounding before an answer is generated. Table 1 summarizes these distinctions across representative multimodal, knowledge-enhanced, and graph-retrieval methods.

[1] The KGs themselves are reproducible from credentialed, non-redistributable MIMIC-IV and ADNI data.

Table 1: **Capabilities of related methods compared to MM-KG.** "∼": partial. "Link": patient observations linked to a standard vocabulary; "Prov": per-observation provenance; "Q-Ret": question-conditioned retrieval.

	Patient evidence in the graph					Representation			Consumer	
Method	Codes	Labs	Imaging	Text	Genetic	Link	Prov	Q-Ret	LLM	GNN
Multimodal clinical representation										
MedFuse (Hayat et al., 2022)	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
MUSE (Wu et al., 2024)	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓
MedGTX (Park et al., 2022)	✓	✗	✗	✓	✗	✗	✗	∼	✗	✓
Knowledge-enhanced patient representation										
MMUGL (Burger et al., 2023)	✓	✗	✗	✓	✗	✓	✗	✗	✗	✓
GraphCare (Jiang et al., 2024)	✓	✗	✗	✗	✗	✓	∼	∼	✗	✓
Graph retrieval for clinical prediction and medical LLMs										
KARE (Jiang et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✓	✓	✗
MindMap (Wen et al., 2024)	✗	✗	✗	✓	✗	∼	∼	✓	✓	✗
MedGraphRAG (Wu et al., 2025)	∼	✗	✗	✓	✗	✓	✓	✓	✓	✗
MM-KG (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

3 METHOD: MULTIMODAL INSTANCE–KNOWLEDGE GRAPHS

Overview. MM-KG represents patient evidence and biomedical knowledge within a single graph while keeping their node sets disjoint, connected only by explicit alignment edges. This separation is the central design decision. Because each alignment edge is an individually addressable object rather than a component of a fused representation, it can be retrieved, traced, and removed, and the contribution of a given patient–relation correspondence to a downstream prediction can be measured directly rather than inferred. Modalities are accordingly not fused into a shared embedding, and the benefit comes from the correspondence rather than from added graph text: length-matched graph text with the correspondence disrupted reduces held-out ranking to near chance (§5.3).

3.1 PROBLEM SETTING AND GRAPH REPRESENTATION

Let $\mathcal{D} = \{x_{p,e,m}\}$ be a multimodal corpus, where p indexes an instance, e an event/encounter, and $m \in \mathcal{M}$ a modality. Let $\mathcal{B} = (V_O, E_O, R_O)$ be an external ontology or KG. Given a query q about instance p, the goal is to answer from a compact subgraph while retaining a trace to the supporting evidence in $\mathcal{D}$ or $\mathcal{B}$. Figure 2 summarizes the general architecture from multimodal inputs to query-conditioned model consumption. We define a multimodal instance–knowledge graph as

$$\mathcal{G} = (V, E, R, T, \tau, a, \pi), \qquad V = V_I \cup V_O, \qquad R = R_{\text{proc}} \cup R_{\text{align}} \cup R_O.$$

Here, V_I and V_O are the disjoint sets of instance and ontology nodes, T is the set of node types, each edge $(u, r, v) \in E$ carries a type $r \in R$, and the functions τ, a, and π store node types, attributes, and provenance. The partition of R is what allows each layer to be traced or removed independently: R_{proc} edges record how each observation was derived, R_{align} edges link observations to concepts, and R_O edges hold the external knowledge. Equivalently, $\mathcal{G} = \mathcal{G}_I \cup \mathcal{B} \cup \mathcal{G}_A$, where $\mathcal{G}_I$ is the instance graph and $\mathcal{G}_A$ contains alignment edges. App. A gives the fragment-level composition.

3.2 PROVENANCE-PRESERVING MULTIMODAL CONSTRUCTION

MM-KG construction preserves the derivation path of every observation: each observation node retains a pointer to its source record and to the time of recording, so that any prediction can be traced to a specific measurement. For every instance, we first build a local event graph: $p \xrightarrow{\text{HAS-EVENT}} e \xrightarrow{\text{HAS-RECORD}} d \xrightarrow{\text{YIELDS}} z$, where d is a source record and z is a modality-derived observation. The source pointer and timestamp are retained in π. Each modality-specific harmonizer produces a graph fragment: $\mathcal{G}_{p,e,m} = H_m(x_{p,e,m}; \phi_m), V_{p,e,m} = \{p, e, d, z_1, \ldots, z_k\}$. H_m converts raw input into typed observations, edges, attributes, and provenance using fixed parameters ϕ_m. Examples include vocabulary lookup for codes, entity linking for text, vision models for imaging, or value-aware rules for measurements. Dataset-specific harmonizers are summarized in §4 and detailed in Apps. B, H, and K.

Observations from different modalities remain distinct even when aligned to the same concept, preserving their source paths and supporting modality auditing and ablation. All derived observations retain provenance, instance and ontology nodes remain disjoint, and harmonization, alignment, and retrieval are frozen before test labels are revealed (App. A).

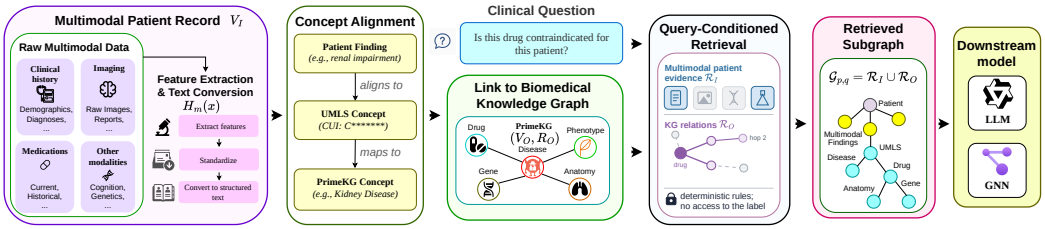


Figure 2: **Overview of MM-KG.** Harmonizers convert each record into typed, provenance-linked observations; an aligner links them to PrimeKG concepts; query-conditioned retrieval supplies a subgraph to an LLM/GNN.

3.3 Ontology Alignment

Alignment assigns at most one ALIGNS-TO edge per observation, so deleting that edge detaches the observation from its concept, and from every relation of that concept, while leaving the rest of the graph unchanged. For observation z, the aligner collects candidate concepts $C_m(z) \subseteq V_O$ through modality-specific mapping routes, such as exact identifiers, synonyms, and parent concepts, selects $c^* = \arg\max_{c \in C_m(z)} s_m(z, c)$, and adds $(z, \text{ALIGNS-TO}, c^*)$ to E only if $\Gamma_m(z, c^*) = 1$. Here s_m ranks candidates by mapping route (exact identifier, then synonym, then guarded parent, then blocklisted name match) , and Γ_m a type-and-route validity check; in our instantiation acceptance is route-based rather than thresholded. Exact identifiers are preferred over synonym or parent routes, observations without a valid mapping remain on the instance side, and all mappings are frozen before evaluation. App. C gives complete routes and coverage statistics.

3.4 Query-Conditioned Subgraph Retrieval

Serializing the full instance graph and ontology would produce an impractically large input, so the model receives only a subgraph conditioned on the question. For query q about instance p, we retrieve two typed components, $\mathcal{G}_{p,q} = \mathcal{R}_I(\mathcal{G}, p, q) \cup \mathcal{R}_O(\mathcal{G}, q)$. $\mathcal{R}_I$ returns the observations of p available at prediction time, under fixed per-modality caps, and in our experiments does not depend on q; $\mathcal{R}_O$ returns the knowledge-graph relations of the entities in the question. Both follow deterministic rules without access to the label, so the input grows with $\mathcal{G}_{p,q}$ rather than with $\mathcal{G}$.

Let $C_q \subseteq V_O$ be the entities linked from q, such as the candidate drugs a question asks about, and $R_q \subseteq R_O$ the relation types whitelisted for the question family. For one-hop retrieval, $N_{R_q}(C_q) = \{v \in V_O : \exists c \in C_q, \exists r \in R_q, (c, r, v) \in E_O\}$, and $\mathcal{R}_O$ returns the edges (c, r, v) that define this set. Multi-hop templates apply this operator recursively with a relation whitelist at each hop. We call the relations $\mathcal{R}_O$ returns for one candidate that candidate's *packet*. The two components connect when a retrieved observation aligns to a concept in a packet, $z \xrightarrow{\text{ALIGNS-TO}} v$ with $(c, r, v) \in E_O$, $c \in C_q$, and $r \in R_q$; this path from the source observation through its aligned concept to the queried relation is the patient–relation correspondence that the factorial experimental design (§5.1) tests.

For set-membership questions, each packet contains the complete list of the requested relation type rather than a truncated list; §5.5 and App. F test this choice. The subgraph is serialized as short natural-language statements grouped under source-channel headers (main experiments) or as typed triples (Table 8). App. A records the correspondence path as an evidence certificate, which documents the evidence supplied to the model but not whether the graph or the answer is correct.

3.5 Consumers and Factorial Evaluation

MM-KG supports different consumers rather than defining a single predictor. An LLM receives a serialization of $\mathcal{G}_{p,q}$ as its prompt, while a heterogeneous GNN propagates type-specific node features from frozen text and image encoders over the graph (App. J). Outcome prediction has no query entities; there, the LLM receives the patient's record with the PrimeKG relations of that patient's own drugs and conditions, and the GNN receives the patient's graph (§5.5).

We use a 2×2 experimental design in §5.1 to test whether patient evidence and external KG relations interact. All four conditions receive the question and the conventional patient record, $\mathcal{G}^{\text{base}}_{p,q}$; B adds the patient modalities retrieved by $\mathcal{R}_I$, $\mathcal{G}^{\text{mod}}_{p,q}$, C adds the packets retrieved by $\mathcal{R}_O$, $\mathcal{G}^{\text{KG}}_{p,q}$, and D adds both. Equivalently, $\mathcal{G}^{ij}_{p,q} = \mathcal{G}^{\text{base}}_{p,q} \cup i\,\mathcal{G}^{\text{mod}}_{p,q} \cup j\,\mathcal{G}^{\text{KG}}_{p,q}, \quad (A, B, C, D) = (00, 10, 01, 11)$. Here,

$0\mathcal{G} = \emptyset$ and $1\mathcal{G} = \mathcal{G}$. For metric M, the interaction is $I_M = \big[M(D) - M(C)\big] - \big[M(B) - M(A)\big]$. For metrics where higher is better, $I_M > 0$ means that one source becomes more useful when the other is present. Deleting the answer-bearing relation from a packet, swapping packets between candidates, and keeping node names without their relations test whether the interaction requires the correct correspondence rather than extra text or familiar entity names (§5.3).

4 Datasets and Implementation

We instantiate MM-KG as independently constructed MIMIC and ADNI graphs linked to PrimeKG (Chandak et al., 2023), using the same provenance-preserving construction with dataset-specific extractors (Appendices B, H, and K). The MIMIC graph covers 51,844 patients, combining MIMIC-IV EHR events (Johnson et al., 2023) with MIMIC-CXR radiographs and reports (Johnson et al., 2019a;b): diagnoses, prescriptions, procedures, and laboratory measurements become provenance-linked observations, and BiomedCLIP (Zhang et al., 2023) with CheXpert-calibrated thresholds (Irvin et al., 2019) converts radiographs into textual findings. The ADNI graph covers 1,000 participants and 4,977 visits spanning imaging, medical history, cognition, genetics, and biospecimens (Veitch et al., 2022); regional brain-MRI features from FastSurfer (Henschel et al., 2020) are compared with normative models, their deviations become imaging findings, and direction-aware anchors to HPO phenotypes connect these findings to the other clinical domains.

We evaluate outcome prediction using 28,582 MIMIC ICU stays and 619 ADNI participants, biomedical fact-checking using 1,920 balanced questions, and patient-specific joins using 3,200 MIMIC items and the ADNI paired contraindication benchmark. The ADNI benchmark contains 510 development items, a 114-item validation set and a 70-item test set; the latter provides the primary cross-cohort ADNI result. The primary MIMIC experiments use Qwen3.5-35B-A3B and a two-layer heterogeneous GraphSAGE consumer, while the primary ADNI test evaluation uses Qwen3.5-9B with thinking disabled. We report AUROC or balanced accuracy as appropriate, with paired-item bootstrap inference for MIMIC and participant-clustered inference for ADNI. Further details on entity linking, evaluation settings, model implementation, and ADNI-specific evaluation details are provided in Appendices C, D, J, and L, respectively.

5 Experiments and Results

We will evaluate MM-KG on whether patient evidence and biomedical knowledge interact (§5.1), how MM-KG compares with external methods (§5.2), whether the interaction depends on the alignment edges (§5.3), retrieval context efficiency (§5.4), and how each source performs alone (§5.5).

5.1 Aligned patient and knowledge evidence is superadditive

Because instance and ontology nodes remain separately addressable, either evidence source can be withheld or supplied independently, which permits the 2×2 design of §3.5: condition A receives neither source, B the patient modalities, C the KG relations, and D both. Fused representations do not admit this design. We call the two sources *superadditive* when the improvement from supplying both exceeds the sum of the improvements from supplying each alone, that is, when the interaction I_M of §3.5 is positive.

The MIMIC join benchmark targets questions for which such an interaction is expected: across all 28,582 MIMIC-IV ICU stays, 29.9% exhibit a laboratory-evident organ state, (e.g., creatinine >2.0, bilirubin >2.0, platelets <100, or hemoglobin <8) that is absent from the coded diagnoses while PrimeKG lists a current medication as contraindicated for that state. Items come from four question families with same-patient matched negatives and length-matched KG lists (App. E).

Table 2 reports the factorial results on the 3,200 MIMIC items and the 70-item ADNI test set. On MIMIC, supplying the patient modalities alone changes balanced accuracy by -0.002 and the KG alone by $+0.031$ ($p<0.001$), so additive effects would improve condition D by 0.029; it improves by 0.150, an interaction of $+0.122$ in balanced accuracy and $+0.209$ in AUROC ($p<0.001$). The interaction is positive and significant in every family on both metrics, including the two-hop *mechanism* questions, and remains $+0.179$ (95% CI $[+0.160, +0.197]$) on the 2,400 items outside the drug-effect family, so the pooled estimate is not driven by that family's large effect.

Table 2: Factorial results in both cohorts: balanced accuracy (chance 0.500) with AUROC in parentheses. **A**: conventional record; **B**: +patient modalities; **C**: +complete KG list; **D**: both. Interaction: I_M.

Family	A	B	C	D	Interaction
contra-report	0.496 (.59)	0.519 (.61)	0.546 (.57)	**0.751 (.82)**	+0.182 (+0.228)
contra-labs	0.507 (.61)	0.516 (.65)	0.560 (.52)	**0.636 (.68)**	+0.068 (+0.121)
mechanism	0.524 (.52)	0.510 (.55)	0.526 (.53)	**0.616 (.67)**	+0.103 (+0.118)
drug-effect	0.475 (.37)	0.449 (.31)	0.494 (.49)	**0.600 (.88)**	+0.132 (+0.463)
All (n=3,200)	0.501 (.52)	0.498 (.51)	0.532 (.53)	**0.651 (.72)**	+0.122 (+0.209)
ADNI paired contraindication, test set (n=70)					
Qwen3.5-9B	0.500 (.56)	0.571 (.57)	0.500 (.55)	**0.786 (.86)**	+0.214 (+0.299)
Qwen2.5-7B-Instruct	0.500 (.50)	0.514 (.59)	0.486 (.51)	**0.629 (.71)**	+0.129 (+0.107)
ADNI paired contraindication, validation set (n=114)					
Qwen3.5-9B	0.500 (.56)	0.535 (.55)	0.500 (.55)	**0.746 (.86)**	+0.211 (+0.319)
Qwen2.5-7B-Instruct	0.500 (.49)	0.509 (.56)	0.500 (.48)	**0.561 (.68)**	+0.053 (+0.129)

The lower blocks of Table 2 apply the same design to the independently constructed ADNI graph, using a paired contraindication benchmark that asks whether a candidate drug has a PrimeKG contraindication relation to a condition in the participant's medical history. Positive and negative items are paired across participants for the same drug and receive identical KG packets, so candidate identity alone cannot determine the label. On the 70-item test set, scored once after the validation analysis and pipeline were frozen, the AUROC interaction is $+0.299$ (95% CI $[+0.065, +0.532]$, $p = 0.0135$), close to the $+0.319$ on the 114-item validation set, and it remains positive with Qwen2.5-7B-Instruct on both. All four prespecified destructive controls fall below condition D after Holm correction, and removing the decisive contraindication edge alone reduces AUROC from 0.860 to 0.714 (App. L). Thus, while ADNI differs from MIMIC in disease domain, evidence channel, and language model, the interaction is still reproduced.

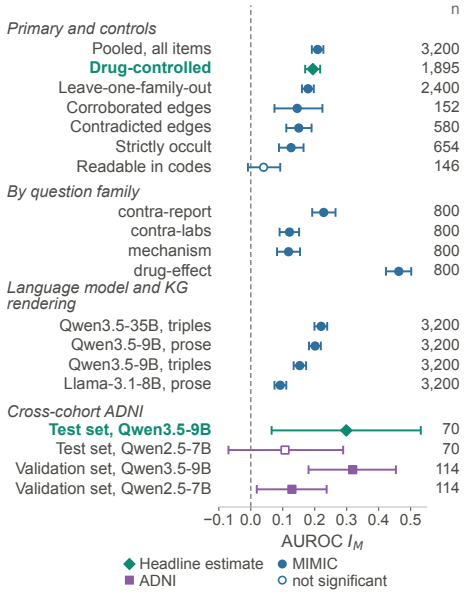


Figure 3: **Patient–KG interaction across question families, controls, language models, and cohorts:** AUROC I_M with 95% CI; marker area $\propto \sqrt{n}$ (Table 18).

This interaction persists in all conducted control experiments (Figure 3). Restricting to the 1,895 items whose candidate drug appears in both label classes brings condition A to chance (AUROC 0.485) while the interaction remains $+0.194$ ($p<0.001$); it is positive across three LLMs and under prose or typed-triple rendering (App. Table 8), and is preserved when answer-bearing edges are stratified by independent prescribing behavior (App. G). It is also specific to conditions documented outside the codes: when coded diagnoses already reveal the condition, C reaches 0.783 and the interaction disappears (App. Table 9), so the interaction requires the laboratory and report evidence that codes do not contain. Overall, neither patient evidence nor biomedical knowledge is sufficiently informative in isolation; modeling their interactions is required.

5.2 MM-KG OUTPERFORMS EXISTING METHODS ON PATIENT-SPECIFIC REASONING

We next compare MM-KG with appropriate existing methods on held-out five-candidate ranking: each item pairs one ICU stay with five administered drugs, one of which has an answer-bearing PrimeKG path, with decoys matched on prescription frequency, graph degree, packet length, and drug class (n=1,244 patients; chance Hits@1 0.200). Condition D reaches Hits@1 0.575 (95% CI $[0.548, 0.603]$) against 0.293, 0.304, and 0.278 for A, B, and C, an interaction of $+0.287$ (95% CI $[0.244, 0.331]$, $p < 0.001$).

Table 3 compares MM-KG with two external methods. MindMap (Wen et al., 2024) +PrimeKG, scored with the same Qwen3.5 protocol, trails MM-KG by 0.131 Hits@1 (95% CI $[0.100, 0.163]$,

$p < 0.001$). An adapted GraphCare-BAT model (Jiang et al., 2024) trained on 19,085 labeled pairs is statistically indistinguishable from MM-KG on the full split, but its score rises from 0.609 to 0.645 when the patient side of its input is removed ($p < 0.001$), indicating that it ranks candidates by properties of the drug rather than by their relation to the patient. On the identity-hard subset ($n = 598$), where a candidate-only prior ranks the wrong drug and a formulary lookup cannot resolve the item, MM-KG reaches 0.582 while both GraphCare variants fall to 0.375 and 0.381. On ADNI, a deterministic PrimeKG lookup reaches AUROC 0.956, so that comparison measures recovery of an explicit patient–relation match rather than differences between methods (App. Table 13); the five-candidate design could not be reproduced there because participant medication sets are too small to defeat candidate-identity shortcuts. Overall, MM-KG matches the strongest existing method where candidate identity suffices and exceeds every external method where it does not.

5.3 Patient–KG interaction requires correct alignment, not graph text

Because retrieval returns each relation as a separately addressable object, the patient–relation correspondence can be disrupted while the patient evidence and nearly all of the graph text stay in place. We apply three such disruptions to condition D on the five-candidate MIMIC ranking benchmark of §5.2: deleting the single answer-bearing relation from the answer candidate's packet, replacing the answer candidate's packet (the relations retrieved for it) with a length-matched packet from another candidate, and shuffling packets across the five candidates. All three fall below condition D ($p < 0.001$; Table 4): deleting the edge returns Hits@1 to the level of condition A, and shuffling drives it below the 0.200 chance level. Because the decoys are matched on prescription frequency, graph degree, packet length, and drug class, the drop reflects the loss of the correspondence rather than of graph text. The ADNI 70-item test set shows the same ordering, with all four prespecified disruptions below condition D (after Holm correction): removing the decisive edge lowers AUROC from 0.860 to 0.714, and keeping node names while removing their relations returns it to the level of condition A. MM-KG also cites the evidence it uses: on the ADNI validation set, all 85 attempted KG citations referred to relations in the supplied packet, and after the decisive edge was removed the model never cited it (App. L).

Table 3: External methods on held-out five-candidate ranking (MIMIC; chance Hits@1 0.2). "Identity-hard": items a candidate-only prior ranks wrongly.

	Full (n=1,244)		Id.-hard (n=598)
Method	Hits@1	MRR	Hits@1
MM-KG (condition D)	0.575	0.738	**0.582**
GraphCare-BAT adapted	**0.609**	**0.764**	0.375
(patient side removed)	*0.645*	*0.787*	*0.381*
MindMap-PrimeKG	0.444	0.646	0.448

The patient-side contribution localizes in the same way. On MIMIC, removing laboratory values or report findings reduces Hits@1 by 0.148 and 0.123 (p<0.001), whereas removing imaging findings or coded history has no measurable effect (App. N.3); on the ADNI validation set, medical history alone with the KG reaches AUROC 0.831 against 0.853 for the full record, and it is the only patient block whose enablement estimate (+0.294) exceeds the prespecified minimum detectable interaction of 0.137 (App. L). The interaction therefore depends on the alignment edges: mismatched packets return both cohorts to the no-knowledge level, and in each cohort the interaction arises from the evidence block that contains the condition the question concerns.

Table 4: **Correspondence disruption.** MIMIC: five-candidate ranking (n=1,244); ADNI: test set (n=70). [a]on MIMIC: another candidate's length-matched packet; on ADNI: degree- and relation-matched null packet.

	MIMIC (Hits@1)	ADNI (AUROC)
Correct alignment (D)	**0.575**	**0.860**
Answer-bearing edge removed	0.299	0.714
Matched control packet[a]	0.237	0.548
Shuffled packets	0.185	0.531
Node names without relations	—	0.558
No knowledge (A)	0.293	0.558
Chance	0.200	0.500

5.4 Query-conditioned retrieval recovers correspondence with less context

We next measure the cost of retrieving the correspondence on the patient-disjoint MIMIC test split (2,398 items, 1,079 patients), scored once after all policies were frozen (App. N.2). MM-KG reaches AUROC 0.731 with 1,206 input tokens, 6.8× fewer than the 8,157 used by the strongest generic policy, SapBERT vector retrieval (0.712, p=0.01); the remaining policies score 0.062–0.204 lower (Figure 4), and when capped at MM-KG's budget they recover the answer path for at most 26% of positive items and fall to 0.526–0.638. On the drug-balanced subset, whose candidate drugs appear

in both label classes, the two leading policies tie (0.695 against 0.693) while the token gap remains (6.7×), so the consistent advantage is context efficiency. MindMap-PrimeKG reaches 0.686 and recovers a complete answer path for 72.6% of positive items, against 99.7% for MM-KG.

Retrieval can fail by missing the answer path or by retrieving it without the model using it. MM-KG misses the path on only 4 of 1,199 positive items and resolves more positives than any other deployable policy: 753, against 714 for the 32K full context, which retrieves 98.2% of paths but loses 463 positives with the path present (App. Figure 6). On ADNI, varying path length from one to three hops showcases both failure modes: a static relation packet falls below closed-book answering despite usually containing the gold path, and interactive graph querying is accurate at one hop but drops to 0.670 and 0.590 balanced accuracy at two and three hops (App. M). Query-conditioned retrieval thus supplies the required correspondence at a fraction of the context, with fewer combined failures than any generic policy.

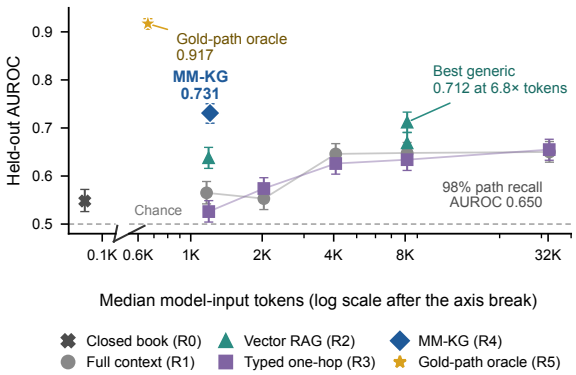


Figure 4: **Retrieval accuracy against context cost** on the held-out MIMIC split: test AUROC (95% CI) vs. median input tokens.

5.5 MM-KG'S MULTIMODAL PATIENT REPRESENTATIONS FOR OUTCOME PREDICTION

Finally, we examine each source on tasks that do not require their combination. For ICU mortality, the zero-shot LLM reaches AUROC 0.733 with the full MM-KG prompt, against 0.570 with demographics alone, and the trained GNN reaches 0.804, against 0.778 for a laboratory-only model; laboratory values contribute most for both (Table 5). The GNN also outperforms logistic regression on 30-day readmission (0.644 against 0.548; App. J), the harmonizer's structured imaging findings exceed the radiologist's verbatim prose by +0.025 AUROC (App. H), and in ADNI, adding biomarkers and imaging to clinical history improves prognosis balanced accuracy from 0.595 to 0.652 (App. L). Removing PrimeKG leaves mortality AUROC unchanged for both consumers, within run-to-run noise (App. I), consistent with §5.1: the knowledge graph contributes when a question requires the patient–relation correspondence.

Table 5: **ICU mortality ablations:** ΔAUROC when each channel is removed (ablated − full). LLM: 28,582 stays; GNN: 5,676-stay test split. $^{***}p<0.001$, $^{*}p<0.05$.

Channel	LLM (zero-shot) (Δ)AUROC	GNN (trained) (Δ)AUROC
Full model	0.733	0.804
− Laboratory	-0.049^{***}	-0.085^{***}
− Prior-visit history	-0.024^{***}	-0.005^{***}
− Medications	-0.011^{***}	−0.001 n.s.
− Radiology report	-0.011^{***}	+0.001 n.s.
− Imaging findings	−0.002 n.s.	-0.008^{*}
− PrimeKG	+0.002 n.s.	−0.002 n.s.

Conversely, when a question concerns a relation directly, complete PrimeKG edges improve balanced accuracy on 1,920 pharmacological questions in all eight relation families, by +0.063 to +0.420, whereas entity names, length-matched incorrect neighborhoods, and corrupted names do not (App. Table 7); because the model follows behaviorally corroborated and contradicted edges alike (App. G), traceability to an edge does not establish clinical correctness. Each source is therefore informative on the task suited to it, and interaction (§5.1) arises when a question requires both.

6 CONCLUSION

We introduced MM-KG, which joins multimodal patient evidence to a biomedical knowledge graph through explicit, provenance-preserving alignment edges. In two independently constructed cohorts, the two sources are individually uninformative on questions that require both but superadditive together, and replacing the aligned relations with mismatched ones returns performance to the no-knowledge level. Query-conditioned retrieval recovers these links with a fraction of the context of generic retrieval, and MM-KG outperforms external graph methods on the items that require consulting the patient. Knowledge graphs thus help clinical LLMs not as background context but as explicit links between a patient's state and the relation a question requires.

AI USE STATEMENT

Generative AI tools were used in preparing this work. They were used to assist implementation of parts of the graph-construction, retrieval, and evaluation code. They were also used to write and edit analysis and plotting code, and to assist writing with suggestions (but all final text is author-edited and verified). They were not used to generate or label data, to construct benchmark items, to propose hypotheses, to design experiments, or to assist mathematical proofs. The authors reviewed all AI-assisted text, code, and figures, and take responsibility for the final content of this work.

ETHICS STATEMENT

This work uses existing de-identified clinical and research data and collects no new data from human subjects. MIMIC-IV and MIMIC-CXR are used under the PhysioNet credentialed data use agreement, ADNI under the ADNI Data Use Agreement, and UMLS under its license. None of these data are redistributed: the released artifacts contain code, aggregate results, terminology mappings, and validated knowledge-graph partitions, and no patient-level records. The worked example in App. E reports a few values from one de-identified MIMIC-IV stay. All inference over patient records used open-weight language models served on local infrastructure. The benchmarks measure whether a model connects patient evidence to knowledge-graph relations; they are not clinical decision tools, and strong performance on them does not validate a model for decisions about patients. PrimeKG contains entries contradicted by observed practice (App. G), so a traceable path does not establish that the relation it states is clinically correct. Both cohorts come from limited sources, MIMIC-IV from a single academic medical center and ADNI from a research cohort that under-represents several demographic groups, so the results may not transfer to other populations or care settings. Clinical use would require prospective validation, calibration of alert burden, and clinician oversight.

REPRODUCIBILITY STATEMENT

Section 3 defines the graph, its alignment edges, and query-conditioned retrieval, and Section 4 specifies the cohorts, benchmarks, models, and metrics. Appendices B, H, and K describe cohort construction and feature extraction; App. C, entity linking; App. D, the evaluation settings, models, decoding, and statistical procedures; App. E, benchmark construction and leak controls; App. J, the GNN experiments; Appendices L and M, the ADNI evaluation and retrieval protocols; and App. N.2, the MIMIC retrieval policies. All evaluated language models are open-weight. PrimeKG is public, UMLS is available under license, and MIMIC-IV, MIMIC-CXR, and ADNI are available to credentialed researchers through PhysioNet and the LONI Image and Data Archive; with this access, the graphs and benchmarks can be rebuilt from source.

REFERENCES

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

Manuel Burger, Gunnar Rätsch, and Rita Kuznetsova. Multi-modal graph learning over UMLS knowledge graphs. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 52–81, 2023. URL https://proceedings.mlr.press/v225/burger23a.html.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*, 2024. doi: 10.48550/arXiv.2402.05391.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng 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*, pp. 787–795, 2017.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, Leopoldo Julian Lechuga Lopez, Krzysztof J. Geras, and Farah E. Shamout. MedMod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans. In *Proceedings of the Sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 781–803. PMLR, 2025. URL https://proceedings.mlr.press/v287/elsharief25a.html.

Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*, 2023.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503, 2022. URL https://proceedings.mlr.press/v182/hayat22a.html.

Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer – a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219: 117012, 2020. doi: 10.1016/j.neuroimage.2020.117012. URL https://doi.org/10.1016/j.neuroimage.2020.117012.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. GraphCare: Enhancing healthcare predictions with personalized knowledge graphs. In *International Conference on Learning Representations (ICLR)*, 2024.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/cb5a3b4589c1a16f14740396625cbfc8-Abstract-Conference.html.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019a.

Alistair E.W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

Alistair E.W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Ruth Johnson, Uri Gottlieb, Galit Shaham, Lihi Eisen, Jacob Waxman, Stav Devons-Sberro, Curtis R. Ginder, Peter Hong, Raheel Sayeed, Xiaorui Su, Ben Y. Reis, Ran D. Balicer, Noa Dagan, and Marinka Zitnik. Embeddings of clinical codes enable knowledge-grounded AI in medicine. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02664-9.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Haolin Li, Tianjie Dai, Zhe Chen, Siyuan Du, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. RAD: Towards trustworthy retrieval-augmented multi-modal clinical diagnosis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4109. URL https://papers.nips.cc/paper_files/paper/2025/hash/b28ca52a43881fd87196c71145f30092-Abstract-Conference.html.

Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, and Edward Choi. Graph-text multi-modal pre-training for medical representation learning. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 261–281. PMLR, 2022. URL https://proceedings.mlr.press/v174/park22a.html.

Benjamin J. Stear, Taha Mohseni Ahooyi, J. Alan Simmons, et al. Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data. *Scientific Data*, 11:1338, 2024. doi: 10.1038/s41597-024-04070-w.

Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57030–57045. PMLR, 2025. URL https://proceedings.mlr.press/v267/su25b.html.

Dallas P. Veitch, Michael W. Weiner, Paul S. Aisen, et al. Using the alzheimer's disease neuroimaging initiative to improve early detection, diagnosis, and treatment of alzheimer's disease. *Alzheimer's & Dementia*, 18(4):824–857, 2022. doi: 10.1002/alz.12422.

Yilin Wen, Zifeng Wang, and Jimeng Sun. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28443–28467, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1381. URL https://aclanthology.org/2025.acl-long.1381/.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Je5SHCKpPa.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce C. Ho, and Carl Yang. RAM-EHR: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 754–765, 2024. doi: 10.18653/v1/2024.acl-short.68. URL https://aclanthology.org/2024.acl-short.68/.

Wenfang Yao, Chen Liu, Kejing Yin, William K. Cheung, and Jing Qin. Addressing asynchronicity in clinical multimodal fusion via individualized chest x-ray generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-0913.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024b. doi: 10.1609/aaai.v38i15.29578.

Haoran Zhang, Supriya Nagesh, Milind Shyani, and Nina Mishra. DOSSIER: Fact checking in electronic health records while preserving patient privacy. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v252/zhang24a.html.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Shasha Zhou, Mingyu Huang, Jack Cole, Charles Britton, Ming Yin, Jan Wolber, and Ke Li. Assessing automated fact-checking for medical LLM responses with knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39646–39656, 2026. doi: 10.1609/aaai.v40i46.41317.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. In *Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pp. 195–204, 2025. URL https://proceedings.mlr.press/v281/zuo25a.html.

A Additional formal details

A.1 Fragment-level graph composition

Let $\mathcal{G}_A = (V_I \cup V_O, E_{\text{align}})$ denote the alignment graph. The complete graph can be written as

$$\mathcal{G} = \mathcal{G}_I \cup \mathcal{B} \cup \mathcal{G}_A, \qquad \mathcal{G}_I = \bigoplus_{p,e,m} \mathcal{G}_{p,e,m}, \tag{1}$$

where $\mathcal{G}_{p,e,m}$ is one modality fragment and $\oplus$ joins fragments through shared instance and event identifiers. This decomposition allows each component to be stored, retrieved, or removed without changing node identities in the other components. Formally, $\tau : V \to T$ assigns node types, $a(v)$ stores node attributes, and $\pi : V \cup E \to \mathcal{P}$ maps nodes and edges to provenance records. Observations with independent identities become nodes, whereas dense measurements meaningful only within an event may remain event attributes.

A.2 Construction invariants

The representation enforces three invariants. First, every derived observation has a source path:

$$\forall z \in V_I^{\text{obs}}, \quad \exists (p, e, d) : p \to e \to d \to z, \quad \pi(z) = (\text{source}, \text{id}, \text{time}). \tag{2}$$

Second, instance observations and ontology concepts have separate identities, $V_I \cap V_O = \emptyset$, even when they denote the same clinical idea. Third, harmonization, alignment, and retrieval are frozen before test labels are revealed. These invariants preserve source identity and prevent label-conditioned graph construction.

A.3 Evidence-trace certificate

A retrieved answer can include the evidence certificate

$$\Xi_{p,q} = (P_I, E_A, P_O), \tag{3}$$

where P_I traces the instance to an observation, E_A is its alignment edge, and P_O is the domain path used for the answer. The certificate is valid only if every edge exists in the frozen graph, every source predates the prediction cutoff, and every domain relation belongs to the query whitelist. It records the evidence given to the model but does not establish that the graph or answer is correct.

B COHORT CONSTRUCTION AND FLOW

MM-KG uses MIMIC-IV v3.1 and MIMIC-CXR-JPG v2.1.0. The following table traces the flow from the raw datasets to each evaluation cohort.

Stage	Patients	Units
MIMIC-IV v3.1 (hosp)	364,627	546,028 admissions
MIMIC-CXR-JPG v2.1.0	65,379	227,835 studies
Present in both datasets	61,868	—
+ ≥1 admission, ≥1 coded dx (= MM-KG)	51,844	219,720 adm.; 204,016 studies
ICU stays in MM-KG (prediction and join-task base)	18,396	28,582 stays
of which lab-occult contraindication carriers	—	8,555 stays (29.9%)
Join-task items (400 matched pairs × 4 families)	—	3,200 items
Fact-checking pair cohort (Table 7 source)	47,100	183,950 admissions

Table 6: MIMIC–PrimeKG graph composition. The two sides are connected by 13,352 typed alignment edges.

Patient-side type	Nodes	PrimeKG type	Nodes
Patient	51,844	Disease	17,080
Admission	219,720	Effect/phenotype	15,311
CXR study	204,016	Drug	7,957
Finding	28,714	Exposure	818
Procedure	9,348		
Drug / diagnosis	5,908 / 84		
Total: 560,800 nodes and 14,407,417 edges.			

"No sampling" means that every admission, study, prescription, and diagnosis is retained within the linkable population defined in row four. The evaluation cohorts below that row are selected only by task requirements, such as having an ICU stay or an occult condition. ICU stays are drawn from the MIMIC-IV `icu` module as the first stay per admission. Eligibility requires $\min(\text{deathtime}, \text{dischtime}) - \text{intime} \geq 48$ hours, so the in-hospital mortality label postdates every current-stay input. Laboratory values, medications, and radiographs enter only from the first 48 hours, and longitudinal history includes only admissions strictly before the current admission.

B.1 MIMIC MODALITY CONSTRUCTION

Each admission connects to diagnoses, prescriptions, procedures, time-bounded laboratory features, and chest-radiograph studies. Diagnoses and prescriptions are preserved as typed patient-side observations before their standardized concepts are aligned with PrimeKG. Laboratory measurements remain attached to their admission and observation window, including their values, units, timestamps, and source-table identifiers.

Each chest radiograph is processed with BiomedCLIP (Zhang et al., 2023) to produce scores for named radiographic findings. Finding-specific thresholds are calibrated against CheXpert reference labels (Irvin et al., 2019). Findings that pass their thresholds become textual `imaging-finding` nodes linked to the source radiograph and study. The ontology aligner then maps these finding nodes to compatible phenotype concepts in PrimeKG. This image-to-finding-to-concept path allows image content to participate in graph retrieval without discarding the original radiograph or its provenance.

Radiology reports are processed independently. Report-derived findings become separate observation nodes connected through `reported-finding` edges. Consequently, an image-derived finding and a report-derived finding remain distinguishable even when they align to the same biomedical concept. This separation supports channel ablations and allows retrieved evidence to be traced to either the image model or the clinical report. Entity-linking routes and coverage are described in App. C.

C ENTITY LINKING DETAILS

This section provides the terminology routes and mapping-coverage details summarized in §3.3. Diagnoses are mapped from ICD-9/10 to UMLS CUIs and then to MONDO or HPO using UMLS

2026AA. Exact source mappings cover 4,168 codes; a synonym hop through MRREL adds 3,565; a guarded `isa`/RB parent hop adds 2,238; and a blocklisted name match adds 266. Coverage reaches 80.5% of disease diagnosis rows (64.2% of all rows; administrative Z/V/Y codes are excluded from the knowledge join). The ICD-to-CUI table must be keyed by *(version, code)*, not by code alone. Twenty-one code strings have different meanings in ICD-9 and ICD-10; for example, E8341 denotes a transport accident in one version and hypermagnesemia in the other. Prescriptions map to DrugBank accessions through normalized names, UMLS synonyms, and NDC routes. We first remove 110 formulation and vehicle strings, including diluents and flushes, which account for 19.7% of raw prescription rows. The pipeline then maps 91.1% of the remaining prescription volume (47.9% of distinct strings), and 92.8% of the mapped volume has an ATC class. Coverage figures in this section use bridge v2; the frozen graph and benchmark use bridge v1 (App. N.1).

D Evaluation settings, models, and statistical procedures

Outcome prediction. The primary MIMIC cohort contains 28,582 ICU stays from 18,396 patients, including 2,306 in-hospital deaths. We retain the first ICU stay per admission and restrict current-stay inputs to the first 48 hours. Longitudinal history includes only admissions preceding the current admission. The zero-shot LLM is evaluated on the full eligible cohort. The trained GraphSAGE consumer uses a fixed 20% patient-level test split containing 5,676 stays, with no patient shared between the training and test sets. The primary metric is AUROC. A separate longitudinal ADNI prognosis experiment contains 619 participants and is evaluated across five seeds using balanced accuracy. The complete GraphSAGE architecture and training procedure are given in App. J.

Biomedical fact-checking. We construct 1,920 balanced yes/no questions across eight biomedical relation families from a source cohort of 47,100 patients and 183,950 admissions. The experimental conditions compare questions without KG evidence against complete PrimeKG relation lists, entity-only controls, shuffled neighborhoods, and corrupted-name controls. The evaluation unit is one question, and the primary metric is balanced accuracy. App. F describes the retrieval-completeness experiments and reports results for all eight families.

Patient-specific join questions. The MIMIC benchmark contains 3,200 balanced items, comprising 400 matched positive–negative pairs in each of four families. The factorial design compares a conventional patient record (A), added patient modalities (B), added PrimeKG relations (C), and both channels (D). The primary drug-controlled analysis uses 1,895 items whose candidate drug appears in both label classes. The ADNI paired contraindication benchmark contains 510 development items, 114 validation items from 61 participants, and a 70-item test set from 39 participants. The one-time test evaluation provides the primary ADNI cross-cohort result, while the validation set is retained for detailed controls and model-sensitivity analyses. Detailed benchmark construction and leak controls are provided in App. E, and the ADNI evaluation is described in App. L.

Language models and scoring. The primary MIMIC experiments use Qwen3.5-35B-A3B with 4-bit GPTQ quantization, served locally with vLLM (Kwon et al., 2023). The primary ADNI test set evaluation uses Qwen3.5-9B, and the same frozen ADNI validation set with Qwen2.5-7B-Instruct was performed as a model-sensitivity analysis. Separate ADNI retrieval studies use local Qwen3-8B and Qwen3-32B-FP8 models and are reported separately from the ADNI validation set. For yes/no questions, we normalize the first-token probabilities over `yes` and `no`, $p_{\text{yes}} = P(\texttt{yes})/[P(\texttt{yes}) + P(\texttt{no})]$, and use p_{yes} as the continuous score; balanced accuracy uses the fixed threshold $p_{\text{yes}} > 0.5$.

Uncertainty and statistical comparisons. MIMIC confidence intervals and comparisons use 2,000 paired-item bootstrap samples. The primary ADNI confidence interval treats participants as clusters, while matched-pair and candidate-drug clustering provide sensitivity analyses. Factorial interactions use the difference-in-differences definition in §3.5. Positive interaction indicates that one information source becomes more useful when the other is available.

Question: Is vancomycin contraindicated for this 83-year-old ICU patient, given her documented condition? Answer `yes` or `no`.

Shared record (A–D): no coded diagnoses; current medications include vancomycin.

Patient evidence (B,D): first-window creatinine 4.3–4.4 mg/dL and blood urea nitrogen (BUN) 91–108 mg/dL reveal severe renal impairment while the renal state is not named in the prompt.

PrimeKG evidence (C,D): vancomycin's complete contraindication list includes kidney disease.

Condition	Patient modalities	PrimeKG relation	Logged answer / score
A	no	no	`no` / 0.011
B	yes	no	`no` / 0.026
C	no	yes	`no` / 0.002
D	yes	yes	`yes` / 0.651

Figure 5: Abridged join-task example; only condition D contains both premises needed for the correct answer, `yes`. The complete prompt and its same-patient negative are provided in App. P

E CONTROLLED JOIN-BENCHMARK CONSTRUCTION AND LEAK CONTROLS

E.1 BENCHMARK ITEMS AND AN EXAMPLE

Benchmark design. The *contra-labs* and *contra-report* families connect an occult patient condition to a drug contraindication; *drug-effect* connects a report finding to an adverse effect; and *mechanism* connects a drug target to a protein implicated in an occult condition. Conditions A–D receive identical questions (Figure 5). Each positive patient–drug pair is matched to a negative drug received by the same patient. Every negative has a non-empty, length-matched KG list and no near-synonym of a visible patient concept, and all main analyses use complete relation lists (App. E).

The eligible patient pools contain 2,307 patients for *contra-labs*, 456 for *contra-report*, 2,143 for *drug-effect*, and 476 for *mechanism*. We sample 400 pairs from each family. Three shortcuts were discovered and closed during construction, in order. (1) *Presence leak*: drugs labeled safe often had no KG list, making an absent KG block a perfect predictor of "no." We therefore require every negative item to have a non-empty list. (2) *Length leak*: positive lists were approximately three times longer because drugs contraindicated in common organ states tend to have long relation lists. We match negatives on rendered character count, yielding final total-prompt positive-to-negative length ratios of 0.99–1.03. (3) *Near-synonym leak*: labels are defined by PrimeKG identifiers, but models read strings, and distinct nodes can have nearly identical names, such as "edema" and "pulmonary edema." We reject a negative if its list contains a subset-containment near-synonym of any patient concept visible in the prompt; the residual rate is 0–1% per family. Before control (3), the drug-effect family was degenerate (specificity 0.04). A pilot with permissive wording ("could *any* finding…") also saturated the decision threshold. The released question instead requires an identified match: "answer yes only if you can identify a specific finding that appears in the adverse-effect list." This wording restores a significant interaction on both metrics, although a tendency to answer yes remains at the 0.5 threshold, as reported in the main text. The stored ICU prompt used a legacy "first 24 hours" header although its values came from the prespecified first-48-hour window. This wording error is identical in all conditions and therefore does not affect their comparison. The released prompt corrects the header. Exact prompt templates and serialization rules for all conditions, together with redacted matched examples are included in App. P.

F RETRIEVAL COMPLETENESS

Across the three set-membership fact-checking families (n=720), retrieval capped at 12 edges yields balanced accuracy 0.718, with sensitivity 0.483 and specificity 0.953. This pattern is consistent with the model treating an omitted relation as absent. Complete neighborhoods raise balanced accuracy to 0.913, with sensitivity 0.911 and specificity 0.914. For the side-effect family, truncated retrieval (0.642) performs worse than no retrieval (0.725). PrimeKG lists a median of 43 side effects per drug and as many as 338, so a 12-edge list omits the queried effect for 78% of positive items. Warning the model that the evidence may be incomplete recovers $+0.142$ on one family and provides no gain on the others. The failure therefore requires more complete retrieval rather than a prompt-only correction. All main-paper KG conditions therefore use complete lists.

Table 7: Biomedical fact-checking with complete retrieval (n=1,920 balanced items). **A**: question only; **B**: entity names; **C**: retrieved KG edges; C_{shuf}: same-length neighborhood from another item; C_{corr}: true edge structure with corrupted names. Δ is C$-$A; all $p < 0.001$.

Family (hops)	A	B	C	C_{shuf}	C_{corr}	$\Delta\uparrow$
Contraindication (1)	0.537	0.608	**0.958**	0.525	0.500	+0.420
Shared drug target (2)	0.575	0.571	**0.975**	0.504	0.500	+0.400
Mechanistic overlap (2)	0.592	0.617	**0.946**	0.533	0.500	+0.355
Drug side effect (1)	0.721	0.742	**0.983**	0.533	0.529	+0.262
Drug class / ATC (1)	0.667	0.658	**0.871**	0.604	0.508	+0.203
Three-hop chain (3)	0.604	0.592	**0.787**	0.525	0.500	+0.184
Comorbidity (1)	0.808	0.767	**0.938**	0.579	0.717	+0.129
Indication (1)	0.925	0.921	**0.988**	0.696	0.500	+0.063

Table 8: Robustness of the join interaction (AUROC) across language models and KG renderings. $^{\dagger}p \geq 0.05$; all other entries $p < 0.05$.

	Qwen3.5-35B		Qwen3.5-9B		Llama-3.1-8B
Family	prose	triples	prose	triples	prose
contra-report	+0.228	+0.208	+0.138	+0.142	$+0.025^{\dagger}$
contra-labs	+0.121	+0.101	+0.063	$+0.010^{\dagger}$	+0.046
drug-effect	+0.463	+0.471	+0.513	+0.479	+0.524
mechanism	+0.118	+0.185	+0.168	+0.110	+0.084
All	+0.209	+0.220	+0.201	+0.153	+0.092

Table 9: Mechanism checks. Left: loss in balanced accuracy after dropping one patient channel with the KG on. Right: the AUROC interaction appears mainly when the state is absent from coded text. $^{*}p < 0.05$, n.s. not significant.

Leave-one-out (KG on)				Occultness split		
Family	$-$labs	$-$report	$-$image	Subset (n)	C	Interaction
contra-labs	$+0.090^{*}$	+0.004	+0.000	Strictly occult (654)	0.483	$+0.126^{*}$
contra-report	-0.017	$+0.162^{*}$	+0.005	Readable in codes (146)	0.783	+0.040 n.s.
mechanism	-0.022	$+0.087^{*}$	-0.026^{*}			
drug-effect	-0.119^{*}	$+0.087^{*}$	-0.010			

G BEHAVIORAL VALIDATION OF PRIMEKG AND LAYER REFINEMENT

For every drug–condition pair we compute the log odds ratio of prescription given the condition over 183,950 admissions. PrimeKG's indication layer is strongly corroborated: the mean log odds ratio is $+1.13$, compared with $+0.08$ for unlabeled pairs ($p<0.001$). By contrast, contraindication pairs are not avoided as a group ($+0.12$), and their mean is indistinguishable from that of unlabeled pairs. We split the cohort into disjoint patient halves and require avoidance to replicate across both halves. This procedure identifies 649 behaviorally corroborated contraindications and 1,817 relations that observed prescribing reliably contradicts, with replication rates of 97.6% and 96.3% against 50% chance. Inspection of the latter group identifies clear errors; for example, PrimeKG lists fondaparinux as contraindicated in heparin-induced thrombocytopenia even though it is a standard treatment. When we repeat the fact-checking audit on both subsets, the model follows corroborated and contradicted edges similarly ($+0.104$ versus $+0.130$; not significantly different). We additionally intersect these behavioral partitions with the answer-bearing contraindication edges of the controlled join benchmark, retaining each selected positive item together with its same-admission matched negative. The behaviorally corroborated subset contains 152 items (76 matched pairs) and retains an AUROC interaction of $+0.145$ (95% CI $[+0.074, +0.224]$, $p < 0.001$). The behaviorally contradicted subset contains 580 items (290 matched pairs) and gives a nearly identical interaction of $+0.150$ (95% CI $[+0.111, +0.190]$, $p < 0.001$). Together, these analyses show that the model can join patient evidence to a supplied relation regardless of whether observed prescribing behavior corroborates that relation; they do not establish the clinical correctness of the underlying KG edge.

H IMAGING CHANNEL AND THE INFORMATION CEILING

For a chest radiograph x, the imaging harmonizer computes finding-specific scores $s_k(x)$ and produces

$$Z_{\text{img}}(x) = \{z_k : s_k(x) \geq \eta_k\}, \tag{4}$$

where z_k is a named radiographic finding and η_k is its calibrated threshold. Each retained finding becomes an observation node linked to its source radiograph and is subsequently aligned with a compatible phenotype concept. We calibrate the vision labels against CheXpert reference labels on 203,999 studies. Findings with AUROC below 0.70 are removed, and finding-specific thresholds are selected on one calibration half. The default threshold of 0.5 predicted some findings at up to $21.8\times$ their true prevalence.

For the zero-shot LLM, image-derived findings contribute approximately zero to mortality prediction in the full context (Table 5). A separate single-channel LLM comparison explains this result. Imaging alone reaches AUROC 0.631, compared with 0.586 for demographics, so the image labels contain outcome signal. However, on this comparison cohort, the full record reaches 0.700 with or without imaging, indicating redundancy for the LLM rather than an absence of information. The trained GNN shows a different pattern: removing image features reduces AUROC by 0.008 (p=0.001).

We next use the radiologist's structured findings as an information ceiling in a representation ladder (n=7,645): finding names, severity, laterality, interval change and devices, and verbatim prose. For the image channel alone, the added graded structure improves AUROC by $+0.046$. Interval change contributes $+0.030$, severity contributes $+0.014$, and laterality contributes no gain. Once laboratory values and history are present in the LLM prompt, however, no representation rung outperforms the no-imaging condition. The structured representation also exceeds the radiologist's verbatim prose by $+0.025$. For this outcome, structured extraction is more useful to the LLM than raw clinical prose, but even the human-derived information ceiling adds no signal beyond the full record.

I RUN-TO-RUN NOISE FLOOR

For patients lacking a chest film, the full-prompt and no-imaging conditions are identical, yet scores differ across serving runs because batch composition changes floating-point reduction order. Across 19,678–21,383 identical-prompt pairs, the AUROC gap between runs is $+0.0003$–$+0.0042$, and the median per-item score drift is 2.5×10^{-3}. We therefore treat ± 0.004 AUROC as the noise floor; all interaction estimates exceed it by more than an order of magnitude. Within-run comparisons (all

join-task conditions; all ladder conditions) are scored by a single engine instance and do not incur this term.

J GNN EXPERIMENTS

The trained consumer in Table 5 is a two-layer heterogeneous GraphSAGE network. It uses `HeteroConv` with a per-edge-type `SAGEConv`, sum aggregation, hidden size 32, and dropout 0.5. Frozen pretrained features are projected to the hidden size: SapBERT name embeddings for patient-side concept nodes, PubMedBERT description embeddings for PrimeKG nodes, Bio_ClinicalBERT embeddings for matched radiology reports, and BiomedCLIP features for radiograph studies. The admission's demographics and first-48-hour laboratory features are concatenated with the pooled graph representation before a two-layer prediction head.

Training uses Adam with learning rate 3×10^{-3}, weight decay 10^{-3}, and batch size 512 for at most 100 epochs. We use validation early stopping with patience 12. The test set is a fixed 20% patient-level stratified split, so no patient appears in both training and test sets. We ensemble predictions from five training seeds. Channel ablations remove the corresponding node and edge types and their features.

On 30-day readmission, the GraphSAGE model reaches AUROC 0.644 against 0.548 for tabular logistic regression. On ICU mortality, the full model reaches AUROC 0.804 against 0.778 for the laboratory-only comparator; on the imaging-present subset, it reaches 0.854 against 0.819. In the paired ICU ablations, laboratory values have the largest contribution ($+0.085$), followed by imaging ($+0.008$) and prior-visit history ($+0.005$); all three changes are significant. Radiology reports, medications, and PrimeKG do not produce measurable gains. In particular, dropping PrimeKG changes AUROC by only -0.002 (p=0.16), matching the null result from the LLM consumer.

K ADNI GRAPH CONSTRUCTION, CONNECTIVITY, AND STATIC EVALUATIONS

Cohort and domains. The full cohort contains 1,000 participants across 4,977 visits: 284 cognitively normal participants, 39 with significant memory concern, 125 with early mild cognitive impairment (MCI), 275 with MCI, 75 with late MCI, and 202 with Alzheimer's disease (AD). The initial construction covers imaging, medical history, cognition, genetics, and biospecimens. Field-level audits add medications, neurological examinations, behavioral inventories, vital signs, and laboratory results. Every patient-local finding retains its participant, visit, source file and field, construction rule, value direction, and missingness status. The patient-level split for the static tasks is 800/100/100. Later language-model studies use the separate item banks described below, and we do not pool their results with the static tasks.

MRI harmonization. FastSurfer (Henschel et al., 2020) processes 8,990 MRI series in segmentation-only mode. Of these, 8,987 complete processing and 8,984 produce the regional-statistics file used downstream. For region k, the normative model is fitted only on cognitively normal baseline scans in the training split:

$$V_{ik} = \beta_{0k} + \beta_{1k}\,\mathrm{Age}_i + \beta_{2k}\,\mathrm{Sex}_i + \beta_{3k}\,\mathrm{SupraTentorialVol}_i + \varepsilon_{ik}, \qquad z_{ik} = \frac{V_{ik} - \widehat{V}_{ik}}{\widehat{\sigma}_k}.$$

Supratentorial volume is the head-size covariate because estimated intracranial volume is unavailable from the segmentation-only run. Each $|z_{ik}| > 2$ event becomes an abnormal imaging finding rather than an attribute merged into an ontology node.

Final graph composition. The phenotype-anchored ADNI graph contains 64,786 retained findings, 213 CUIs, 264 mapped concepts, 71,240 patient-layer nodes, and 334,764 patient-layer edges. An additional 9,264 external concept nodes and 41,246 concept-to-concept edges are stored separately. The patient-layer and concept-layer counts use different denominators and should not be summed.

Coverage did not imply connectivity. The expanded Desikan–Killiany–Tourville (DKT) to UBERON crosswalk raised imaging linkage coverage from 35.8% to 88.4% and total feature-column coverage from 44.5% to 86.6%. These values measure coverage rather than mapping correctness.

Under anatomical alignment alone, the main concept component contained 9,650 of 9,905 nodes and all nine non-imaging domains, while all 80 imaging concepts remained outside it in 46 fragments. Of the 182,298 edges touching mapped brain regions, 99.8% were tissue-expression edges from the noisy layer excluded during graph construction. No retained path therefore connected imaging concepts to disease, drug, or phenotype concepts.

The phenotype-anchored configuration retained the anatomical `located-in` edges and added direction-aware `has-imaging-phenotype` edges. Ninety-two imaging columns mapped to 22 HPO phenotypes and created 27,442 phenotype-anchoring edges. Twenty of these phenotypes entered the main component and gained two- to four-hop paths to cognitive, behavioral, genetic, and medication concepts. This was a structural repair rather than a mapping-accuracy result: none of the 96 DKT-to-HPO review rows has been adjudicated, and 60 use lower-confidence regional fallbacks. For the held-out dynamic benchmark, a separate answerable-concept set maps 112 of 126 concepts, including 104 that resolve to exact PrimeKG nodes. These values also measure coverage rather than manually verified accuracy.

Static and graph-native boundaries. On three-class diagnosis, the flat condition reaches macro-F1 0.6177 ± 0.026. The concept-mapped patient-graph condition reaches 0.6921, but a flat control model given the same added columns reaches 0.7383 ± 0.014. The additional data, rather than concept mapping, therefore explain the improvement over the original flat condition. A discontinued cognitive-slope regression is also negative: the flat model reaches Pearson correlation 0.4034 ± 0.023, whereas earlier graph-feature configurations range from 0.244 to 0.274. The cognition-dropout experiment suggests greater robustness for the graph condition. However, it lacks a non-graph concept-pooling control that would separate graph routing from feature dilution, so we do not interpret it as evidence of a graph gain.

The pre-anchor graph-native experiment also gives a null result across ten seeds. The matched no-graph network reaches macro-F1 0.6087 ± 0.0215, and the four-layer graph-attention network reaches 0.6109 ± 0.0529. Because the pre-anchor graph contains no cross-modal concept path from imaging to another domain, this experiment cannot distinguish an ineffective consumer from a graph with no useful structure to propagate. A phenotype-anchored graph-native rerun is therefore required before the phenotype repair can support a message-passing claim.

L ADNI PAIRED CONTRAINDICATION BENCHMARK: VALIDATION, TEST EVALUATION, AND EVIDENCE CONTROLS

Construction. The candidate-conditioned retriever is deterministic rule-based code. It normalizes the candidate drug, applies family-specific relation whitelists, and retrieves its complete bounded PrimeKG packet. It has no access to the answer-bearing medical-history block or the hidden label. The median, 90th-percentile, and maximum packet lengths are 623, 947, and 1,298 tokens. The paired contraindication benchmark uses cross-patient, same-candidate positive and negative pairs with identical KG packets and at most two items per participant. A positive item requires a displayed mapped condition and a contraindication edge between the candidate and that condition. A negative item means only that none of the mapped conditions displayed in the record has such an edge.

The dynamic study contains 510 development items from 265 participants, 114-item validation from 61 participants, and a 70-item test set from 39 participants, scored once with the frozen Qwen3.5-9B pipeline after the validation analysis was fixed. The validation bank contains 57 matched pairs and 53 candidate drugs; only three candidates occur in more than one pair. Qwen3.5-9B is served locally with vLLM in bfloat16, and constrained yes/no probabilities provide the continuous score. The primary confidence interval clusters by participant. Pair-clustered, candidate-clustered, and equally weighted candidate analyses serve as sensitivity checks.

Interaction and controls. With Qwen3.5-9B, validation AUROC is 0.5636 for A (neither channel), 0.5534 for B (patient only), 0.5506 for C (PrimeKG only), and 0.8590 for D (both). The interaction is $+0.3186$ with participant-clustered 95% CI $[+0.1806, +0.4539]$; pair- and candidate-clustered analyses are also positive and exclude zero. The aligned condition exceeds shuffled PrimeKG (0.5131), a degree/relation-matched null (0.5260), decisive-support removal (0.6316), and node names without relations (0.5531).

Table 10: **ADNI paired contraindication benchmark across language models.** Both models are evaluated on the same 114-item validation bank using the same item keys, arm definitions, KG packets, and non-KG inputs. Interactions are I_M with participant-clustered 95% confidence intervals.

Model	A	B	C	D	Interaction [95% CI]
Qwen2.5-7B-Instruct	0.491	0.563	0.479	0.679	$+0.1285$ $[+0.0185, +0.2372]$
Qwen3.5-9B	0.5636	0.5534	0.5506	0.8590	$+0.3186$ $[+0.1806, +0.4539]$

Table 11: **ADNI paired contraindication test-set results** ($n = 70$ items; 35 positive–negative pairs from 39 participants). AUROC intervals are participant-clustered 95% confidence intervals. The factorial interaction is $I_M = +0.2988$ (95% CI $[+0.0650, +0.5319]$; predeclared $p = 0.0135$).

Condition	AUROC	Participant-clustered 95% CI
A: base	0.5584	$[0.4050, 0.7151]$
B: patient only	0.5706	$[0.4016, 0.7305]$
C: knowledge only	0.5494	$[0.3946, 0.6988]$
D: patient + dynamic KG	**0.8604**	$[0.7433, 0.9557]$
Shuffled KG	0.5314	$[0.3949, 0.6741]$
Degree/relation null	0.5478	$[0.3979, 0.6933]$
Decisive support removed	0.7143	$[0.5676, 0.8448]$
Node names only	0.5584	$[0.3956, 0.7040]$

Modality localization. On validation, the full patient representation reaches AUROC 0.5546 without KG and 0.8529 with the correct KG packet. Medical history alone with the KG reaches 0.8313, while removing all patient information yields 0.5003 without KG and 0.5046 with KG. The medical-history enablement estimate is $+0.2936$ (95% CI $[+0.1618, +0.4237]$) and is the only tested patient block whose magnitude exceeds the prespecified 0.137 minimum detectable interaction. Independent repeats change no endpoint by more than 0.0022, below the frozen 0.010 reproducibility threshold.

Longitudinal multimodal prognosis. A separate five-seed study asks whether memory declines faster over roughly four years than the patient's baseline predicts. Among 619 participants, clinical history and examination alone reach balanced accuracy 0.595, while adding biomarkers and imaging reaches 0.652, a $+0.057$ change with the same sign in all five seeds. The graph-augmented counterparts reach 0.577 and 0.622. Both graph increments change sign across seeds and remain smaller than the variation across seeds. We therefore interpret only the multimodal increment as a consistent effect. The label is based on follow-up rather than baseline diagnosis; App. source files record the rejected label variants and their baseline and follow-up correlations.

Development-stage laboratory/vital-sign join. An earlier pooled family contains 194 matched items from 86 participants. Laboratory values or vital signs indicate a condition absent from the coded record, and the question asks whether that condition is contraindicated with a current drug. Balanced accuracy is 0.505, 0.541, 0.500, and 0.701 for A–D; the interaction is $+0.165$ with patient-clustered 95% CI $[+0.101, +0.225]$. However, the family has no protected test split, and its design choices were made while examining the pooled bank. We therefore report it as development evidence; the ADNI test set above provides the primary cross-cohort result. Here "occult" means absent from an ADNI enrollment checkbox, not a condition missed by a clinician.

ADNI explanation and provenance audit. We conducted a separate post-hoc explanation audit on the 114-item Qwen3.5 validation bank. Condition D produced valid structured output for all 114 items. Across 85 attempted KG citations, every cited relation was present in the supplied packet; 53 cited the decisive answer-bearing edge, 32 cited another valid in-packet edge, and no unsupported KG citation was observed. Relative to the 57 positive items, decisive-edge citation recall was 53/57 (0.930).

Table 12: **Destructive-control contrasts on the ADNI test set.** Differences are D minus each control; Holm adjustment covers the four prespecified destructive-control comparisons.

Contrast	ΔAUROC	95% CI	Holm p
D $-$ shuffled KG	$+0.3290$	$[+0.1484, +0.4887]$	0.0339
D $-$ degree/relation null	$+0.3127$	$[+0.1742, +0.4445]$	0.0290
D $-$ decisive support removed	$+0.1461$	$[+0.0789, +0.2192]$	0.0002
D $-$ node names only	$+0.3020$	$[+0.1637, +0.4534]$	0.0200

Table 13: Exploratory external-method comparison on the ADNI paired contraindication benchmark validation bank ($n = 114$). Methods use their method-specific evidence construction; the direct PrimeKG lookup is the primary no-model comparator. [a]MM-KG condition D is the previously frozen Qwen3.5-9B patient-plus-KG result on the same validation bank; its complete candidate KG packet differs from the evidence representations used by MindMap and GraphCare, so the row is shown for within-bank context rather than as a controlled architecture-only comparison.

Method	AUROC
MM-KG (condition D)[a]	0.8590
MindMap v2	0.9674
Direct PrimeKG lookup	0.9561
GraphCare GIN adapted	0.9437
GraphCare BAT adapted	0.8122
Candidate-only prior	0.5000

Removing the decisive supporting edge did not induce fabrication of the missing relation. All 114 support-removed outputs remained structurally valid; among the 57 positive items, the model cited another available in-packet relation in 48 cases and abstained in 9, and it never cited the removed identifier. The positive-item citation-attempt rate decreased from 55/57 to 48/57 ($\Delta = -0.123$, 95% CI $[-0.204, -0.039]$). Removing only provenance headers leaves predictive performance essentially unchanged: AUROC is 0.8587 with the original D prompt and 0.8480 without provenance headers, for a paired difference of -0.0108 (95% CI $[-0.0579, +0.0322]$). Pair- and candidate-clustered intervals likewise include zero. Thus, provenance labels improve traceability without being responsible for the observed predictive gain.

M ADNI RETRIEVAL-REGIME STRESS TEST

Design. The hop ladder contains five balanced families with 100 items each and manipulates only shortest-path length. The three one-hop families are the individual legs of the longer drug–protein–disease–disease chain. The closed-book condition supplies no graph. Static context serializes a relation-guided packet, while the query loop lets the model issue one `neighbors`, `find`, or `answer` action per round. Hop exclusivity is checked during item construction, and positives and negatives are matched on evidence length. The experiment uses full PrimeKG because the cohort-specific phenotype-anchored graph is materialized only one hop from cohort seeds. It therefore evaluates graph access rather than the sufficiency of the ADNI cohort-specific graph.

Table 14: Balanced accuracy for three graph-access regimes in the ADNI-derived hop ladder. Gold-path presence is measured in the static packet.

Relation family	Hops	Closed book	Static context	Query loop	Gold path
Drug–protein	1	0.760	0.370	0.920	100%
Protein–disease	1	0.570	0.400	0.970	100%
Disease–disease	1	0.820	0.290	0.980	100%
Drug–protein–disease	2	0.450	0.430	0.670	100%
Drug–protein–disease–disease	3	0.470	0.450	0.590	90%

The regimes fail differently. Closed-book performance falls to chance at two and three hops, showing that the model does not reliably supply the missing relations from its parameters. Static context falls below closed book in every family even though the gold path is usually present. At two hops, it answers yes on 82% of items and has specificity 0.080. The query loop is strong at one hop but reaches only 0.670 and 0.590 at two and three hops. At two hops, its sensitivity is 0.340 because failed retrieval is often interpreted as evidence that a relation is absent. Complete retrieval is therefore necessary but not sufficient, and the two graph-enabled regimes fail in opposite directions.

Qualitative traces. On a one-hop enzyme question, the larger Qwen3 model independently queries the enzyme neighbors of celecoxib and warfarin, intersects the two lists, and correctly identifies CYP2C8, CYP2C9, and CYP3A4. On a whole-regimen question, the smaller model repeats the same truncated disease-neighborhood query for all 15 rounds, never queries from the drug side, and returns the wrong answer. A paired two-hop example shows a complementary static-context failure. The packet contains the positive path from lutein through ALB to gastric neoplasm, but the

model also answers yes to a negative naproxen–plasma-cell-myeloma item with no such path. These examples show why a retrieved answer should be checked against an explicit path rather than treated as grounded simply because graph text was present.

Result boundaries. The hop ladder is a knowledge-access stress test, not a clinical outcome task, and it uses full PrimeKG rather than the materialized ADNI graph. The reported language-model runs also require an archived repeat-run noise estimate. In addition, the 32-token constrained-decoding configuration produced whitespace-only truncations before the answer field in some earlier studies. These issues must be resolved before interpreting small differences between retrieval policies.

N Additional frozen evaluation results

N.1 Audit boundary for the retrospective MIMIC split

The original 3,200-item join benchmark was constructed without a development–test split, and shortcut controls were tuned while inspecting the complete item bank. We subsequently created a patient-disjoint retrospective split (development: 802 items from 360 patients; test: 2,398 items from 1,079 patients), selected policies and thresholds on development, froze all configurations, and scored the test split once. The held-out results therefore test model and retrieval-policy generalization within a previously designed benchmark; they are not an independent validation of benchmark construction. The patient-side block used by these policies is admission-scoped rather than question-specific, while the PrimeKG block is candidate- and relation-conditioned. The frozen graph and benchmark use the bridge-v1 alignment, whereas separately quoted MIMIC mapping-coverage audits use bridge v2; coverage percentages from bridge v2 should not be treated as properties of the frozen benchmark prompts. Finally, in 4 of 1,600 positive items, the answer-bearing report finding falls beyond the frozen prompt's 20-name report cap. These prompts were not changed after freezing, so those cases remain in the evaluation as unavailable patient evidence.

N.2 Held-out MIMIC retrieval policies

Table 15: **Retrieval-policy results on the retrospective MIMIC test split** (n=2,398 items; 1,079 patients). R0–R5 are the frozen retrieval policies; MindMap-PrimeKG is included as an external comparator evaluated on the same test items. Input is the median number of model-input tokens. Path recall is the fraction of the 1,199 positive items for which the gold patient evidence is visible under the frozen patient-side caps and every edge of at least one gold answer path is retrieved. Δ and Holm-adjusted p values compare each policy with R4. The final column gives retrieval/use failures among the 1,199 positive items.

Policy	Configuration	Input	Path recall	AUROC [95% CI]	Δ	Holm p	Failures (retrieval / use)
R5	Gold-path oracle	660	0.997	0.917 [0.906, 0.927]	+0.186	< 0.001	4 / 251
R4	MM-KG	1,206	0.997	**0.731** [0.710, 0.750]	ref.	—	4 / 442
R2	SapBERT, KG-only, 8K	8,157	0.746	0.712 [0.691, 0.733]	−0.019	0.010	305 / 263
MindMap	evidence + yes-token	4,448	0.726	0.686 [0.665, 0.708]	−0.044	< 0.001	329 / 401
R2	SapBERT, multi-query, 8K	8,157	0.649	0.669 [0.647, 0.691]	−0.062	< 0.001	421 / 216
R3	Typed one-hop, 32K	32,353	0.747	0.655 [0.633, 0.677]	−0.076	< 0.001	304 / 373
R1	Full admissible, 32K	32,420	0.982	0.650 [0.629, 0.672]	−0.081	< 0.001	22 / 463
R1	Full admissible, 8K	8,156	0.736	0.648 [0.626, 0.670]	−0.082	< 0.001	316 / 312
R1	Full admissible, 4K	4,080	0.549	0.646 [0.623, 0.667]	−0.085	< 0.001	541 / 280
R2	SapBERT, KG-only, matched	1,189	0.257	0.638 [0.616, 0.660]	−0.092	< 0.001	891 / 69
R3	Typed one-hop, 8K	8,154	0.735	0.634 [0.612, 0.656]	−0.097	< 0.001	318 / 343
MindMap	graph-of-thought generation	4,409	0.726	0.631 [0.612, 0.651]	−0.100	< 0.001	329 / 288
R3	Typed one-hop, 4K	4,080	0.629	0.626 [0.604, 0.648]	−0.104	< 0.001	445 / 280
R3	Typed one-hop, 2K	2,032	0.194	0.574 [0.551, 0.597]	−0.157	< 0.001	966 / 38
R1	Full admissible, matched	1,166	0.001	0.565 [0.542, 0.589]	−0.166	< 0.001	1,198 / 0
R1	Full admissible, 2K	2,032	0.155	0.553 [0.530, 0.577]	−0.177	< 0.001	1,013 / 103
R0	Closed book	74	0.000	0.548 [0.526, 0.572]	−0.182	< 0.001	1,199 / 0
R3	Typed one-hop, matched	1,191	0.018	0.526 [0.504, 0.549]	−0.204	< 0.001	1,177 / 4

On the drug-balanced sensitivity subset (n=1,418), R4 reaches AUROC 0.695 (95% CI $[0.666, 0.722]$) with 1,217 tokens, while the strongest R2 policy reaches 0.693 ($[0.665, 0.721]$) with 8,161 tokens ($\Delta = -0.002$, Holm-adjusted p=0.886). The oracle reaches 0.920 ($[0.904, 0.934]$).

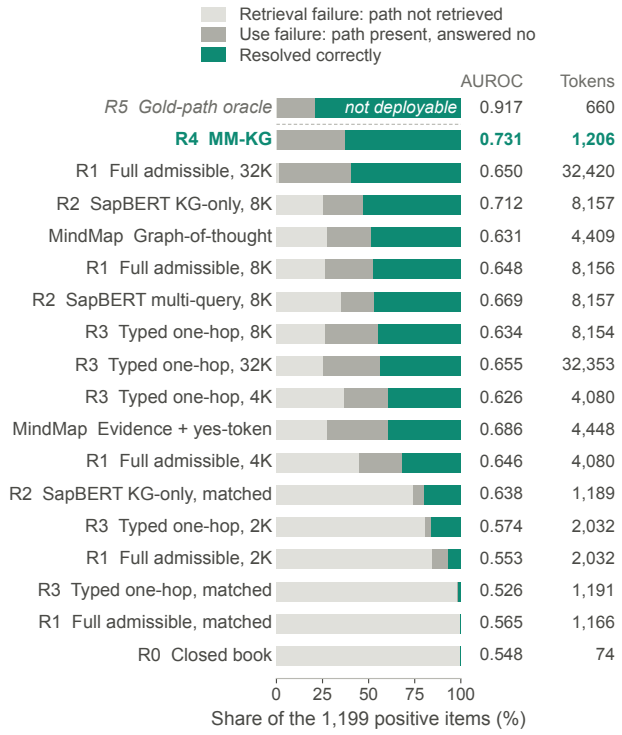


Figure 6: **All sixteen frozen retrieval policies and two MindMap comparator configurations, ordered by the proportion of positive items resolved.** Each bar decomposes the 1,199 positive items of the frozen MIMIC test split into the answer path never being retrieved, the path being present while the model still answers negatively, and the remainder, resolved correctly, ordering rows by the share of the 1,199 positive items each policy resolves correctly rather than by AUROC. The two orderings are not equivalent: the 32K full-admissible policy resolves more positive items than the strongest vector policy despite a lower AUROC, because this decomposition counts only positive items at the decision threshold whereas AUROC is threshold-free and also reflects performance on the negatives.

N.3 HELD-OUT CANDIDATE RANKING, CONTROLS, AND EXPLANATION DIAGNOSTICS

An earlier five-packet rendering provides a separate audit of explanation traceability. With provenance present, 98.2% of outputs are valid JSON and 79.5% of cited edges exist in PrimeKG, but only 16.5% reproduce the exact answer path; evidence precision and recall are 0.312 and 0.447. Removing provenance headers unexpectedly raises Hits@1 to 0.492 ($\Delta = +0.030$, Holm-adjusted p=0.016), while evidence precision and recall become zero because source tags are unavailable. Deleting the gold edge drives condition-identification accuracy to 0.007 and exact-path recovery to 0.001. Provenance therefore makes supplied evidence inspectable, whereas correct semantic alignment is what supports the decision; neither property establishes that the underlying KG relation is clinically valid.

Table 16: Frozen candidate-ranking results on MIMIC (n=1,244 independent patients; chance Hits@1 0.200). MM-KG uses the per-candidate query-conditioned R4q formulation. Marginal Hits@1 intervals and paired Δ intervals use 2,000 bootstrap resamples of the 1,244 independent patients; Holm-adjusted p values compare each row with condition D. The identity-hard subset (n=598) contains items for which a candidate-only prior based on training-bank drug frequencies ranks the wrong drug. Structured-explanation diagnostics were evaluated separately under the earlier five-packet rendering and are not mixed with the R4q results here.

Group	Condition / policy	Hits@1 [95% CI]	Δ vs. D [95% CI]	Holm p	MRR	Identity-hard
Factorial	A: neither channel	0.293 [0.269, 0.319]	−0.281 [−0.318, −0.245]	< 0.001	0.539	0.268
Factorial	B: patient only	0.304 [0.277, 0.330]	−0.271 [−0.307, −0.236]	< 0.001	0.555	0.276
Factorial	C: PrimeKG only	0.278 [0.254, 0.304]	−0.297 [−0.328, −0.265]	< 0.001	0.523	0.296
Factorial	D: patient + PrimeKG (R4q)	**0.575** [0.548, 0.603]	ref.	—	0.738	**0.582**
Disruption	Gold edge removed	0.299 [0.273, 0.323]	−0.276 [−0.305, −0.250]	< 0.001	0.540	0.279
Disruption	Matched wrong packet	0.237 [0.213, 0.261]	−0.338 [−0.371, −0.303]	< 0.001	0.484	0.253
Disruption	Shuffled packets	0.185 [0.162, 0.206]	−0.390 [−0.425, −0.355]	< 0.001	0.451	0.176
Patient channel	No laboratories	0.427 [0.400, 0.455]	−0.148 [−0.177, −0.120]	< 0.001	0.627	0.415
Patient channel	No report	0.452 [0.424, 0.479]	−0.123 [−0.150, −0.097]	< 0.001	0.652	0.448
Patient channel	No imaging	0.590 [0.562, 0.616]	+0.015 [−0.003, +0.034]	0.21	0.747	0.599
Patient channel	No codes	0.596 [0.569, 0.623]	+0.021 [−0.001, +0.042]	0.18	0.749	0.600
Patient channel	No medications	0.611 [0.583, 0.637]	+0.036 [+0.011, +0.060]	0.015	0.761	0.625
Audit channel	No provenance headers	0.602 [0.575, 0.628]	+0.027 [+0.006, +0.048]	0.044	0.754	0.602
External	MindMap-PrimeKG (score)	0.444 [0.416, 0.472]	−0.131 [−0.163, −0.100]	< 0.001	0.646	0.448
External	GraphCare-BAT adapted	0.609 [0.582, 0.636]	+0.034 [−0.006, +0.072]	0.21	0.764	0.375
External control	GraphCare-BAT, patient removed	0.645 [0.618, 0.671]	+0.070 [+0.033, +0.107]	< 0.001	0.787	0.381

For transparency, the earlier five-packet rendering of condition D reached Hits@1 0.461 (95% CI [0.433, 0.488]; MRR 0.665). The per-candidate R4q formulation was selected on development before its test score was examined and increased test Hits@1 by +0.114 (95% CI [+0.084, +0.146], $p < 0.001$).

Table 17: Strict-matching d8 sensitivity (n=380; chance Hits@1 0.200). The ordering is preserved, but the D–A comparison is not conclusive at this sample size (p=0.060).

Condition	Hits@1 [95% CI]	Hits@3	MRR	Candidate AUROC
A: neither channel	0.342 [0.297, 0.390]	0.726	0.567	0.581
B: patient only	0.250 [0.208, 0.292]	0.766	0.520	0.606
C: PrimeKG only	0.274 [0.226, 0.318]	0.747	0.532	0.564
D: patient + PrimeKG	**0.408** [0.361, 0.458]	0.818	0.626	0.672

N.4 ADNI VALIDATION TABLE AND TARGETED PREDICTION BOUNDARY

O ADNI DATA USE AND ACKNOWLEDGEMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators is available through the official ADNI acknowledgement list.

ADNI acknowledgement Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and

Table 18: The join interaction across every subset, model, rendering and cohort tested. All entries are AUROC interactions I_M as defined in §3.5; positive means the two evidence channels are superadditive. $^{***}p < 0.001$; $^{*}p < 0.05$; n.s. not significant. [a] §5.1 reports the interaction as positive and significant for every family on both metrics without printing per-family p values. [b] Per the Table 8 caption, every entry without a dagger satisfies $p < 0.05$. [c] Excluding the drug-effect family. [d] Behavioral subsets retain a positive item when its answer-bearing contraindication edge belongs to the independently defined prescribing-behavior partition and retain that item's same-admission matched negative. The corroborated subset contains 76 positive–negative pairs; the contradicted subset contains 290 pairs.

Analysis	n	Interaction	95% CI	Source
Primary				
Pooled, all items	3,200	$+0.209^{***}$	$[+0.191, +0.227]$	§5.1
Drug-controlled	1,895	$+\mathbf{0.194}^{***}$	$[+0.170, +0.217]$	§5.1
Leave-one-family-out[c]	2,400	$+0.179^{***}$	$[+0.160, +0.197]$	§5.1
Behaviorally corroborated contraindications[d]	152	$+0.145^{***}$	$[+0.074, +0.224]$	App. G
Behaviorally contradicted contraindications[d]	580	$+0.150^{***}$	$[+0.111, +0.190]$	App. G
contra-report	800	$+0.228^{***}$	$[+0.192, +0.265]$	§5.1
contra-labs	800	$+0.121^{***}$	$[+0.090, +0.151]$	§5.1
mechanism	800	$+0.118^{***}$	$[+0.082, +0.154]$	§5.1
drug-effect	800	$+0.463^{***}$	$[+0.423, +0.502]$	§5.1
Strictly occult	654	$+0.126^{***}$	$[+0.088, +0.165]$	Table 9
Readable in codes	146	$+0.040$ n.s.	$[-0.009, +0.092]$	Table 9
Qwen3.5-35B, triples	3,200	$+0.220^{***}$	$[+0.199, +0.239]$	Table 8
Qwen3.5-9B, prose	3,200	$+0.201^{***}$	$[+0.182, +0.219]$	Table 8
Qwen3.5-9B, triples	3,200	$+0.153^{***}$	$[+0.134, +0.173]$	Table 8
Llama-3.1-8B, prose	3,200	$+0.092^{***}$	$[+0.074, +0.111]$	Table 8
Cross-cohort				
ADNI test set (Qwen3.5-9B)	70	$+\mathbf{0.2988}^{*}$	$[+0.0650, +0.5319]$	Table 11
ADNI test set (Qwen2.5-7B)	70	$+0.107$ n.s.	$[-0.070, +0.289]$	Table 2
ADNI validation set (Qwen3.5-9B)	114	$+\mathbf{0.3186}^{***}$	$[+0.1806, +0.4539]$	Table 19
ADNI validation set (Qwen2.5-7B)	114	$+0.1285^{*}$	$[+0.0185, +0.2372]$	Table 10

Table 19: ADNI validation set (n=114 items; 57 matched pairs from 61 participants). A–D form the factorial design; remaining rows disrupt the evidence. Rows are point estimates; the participant-clustered interaction is $+0.3186$ (95% CI $[+0.1806, +0.4539]$).

Condition	AUROC↑	Evidence available or disrupted
A: base	0.5636	Neither patient nor KG
B: patient only	0.5534	Patient record, no KG relation
C: PrimeKG only	0.5506	KG relation, no patient condition
D: patient + PrimeKG	**0.8590**	Both aligned channels
Shuffled PrimeKG	0.5131	Wrong candidate packet
Degree/relation-matched null	0.5260	Nuisance structure without true content
Decisive support removed	0.6316	Correct packet without answer-bearing edge
Node names only	0.5531	Entities without relations

Table 20: Targeted APOE-relation augmentation does not improve ADNI biomarker prediction. Intervals are 95% confidence intervals for the AUROC change.

Endpoint	Patient representation	+ targeted KG	ΔAUROC [95% CI]
Amyloid status	0.777	0.773	-0.004 $[-0.027, +0.018]$
Tau status	0.708	0.687	-0.021 $[-0.058, +0.017]$

Table 21: Prompt blocks included in factorial conditions A–D.

Block	Content (exact header)	A	B	C	D
`demo`	"A {age}-year-old {sex} patient in intensive care."	✓	✓	✓	✓
`codes`	"Coded diagnoses on record: …" (prior admissions; ≤ 25 names), or "No coded diagnoses on record."	✓	✓	✓	✓
`meds`	"Current medications ({n}): …" (first 48 h; ≤ 40 names)	✓	✓	✓	✓
`labs`	"Laboratory values, first 24 hours: creatinine {min–max}, BUN …" (up to 17 analytes; legacy header, see below)		✓		✓
`report`	"Radiology report for the chest film describes: …" (≤ 20 UMLS finding names)		✓		✓
`imaging`	"Chest radiograph, automated image analysis reports: …" (BiomedCLIP findings)		✓		✓
`kg`	Candidate packet, Supplementary Section P.3			✓	✓

Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research &Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

P Benchmark prompts and worked examples

P.1 Common prompt template

Every item of the controlled join benchmark is scored with one chat prompt. The system message is fixed across items and conditions:

```
You are a clinical pharmacology expert reviewing an intensive-care
medication list.
Answer the question with a single word: yes or no.
```

The user message is the concatenation of the blocks admitted by the condition (App. P.2), separated by blank lines, followed by the question. The chat template is applied with the generation prompt appended and, for Qwen3.5, thinking disabled. Decoding is deterministic (temperature 0, one output token, top-20 log-probabilities); the item score is $P(\texttt{yes})/(P(\texttt{yes}) + P(\texttt{no}))$ over the first output token, and the logged answer is `yes` when that score exceeds 0.5.

P.2 Condition A–D block definitions

Each item stores seven text blocks. The four conditions differ only in which blocks are included; the question and every included block are byte-identical across conditions.

For exact reproduction, the frozen prompts retain the legacy "first 24 hours" laboratory header, although all laboratory values were computed from the prespecified first-48-hour feature window; this header was identical across experimental conditions.

The `labs`, `report`, and `imaging` blocks are omitted when empty (no chest film in the window: 380 items lack a report block and 1,458 an imaging block), whereas an empty `codes` block is rendered as the fixed sentence above. Because a positive and its matched negative share the admission, patient-side block presence is identical within every pair. The `kg` block is never empty: every negative has a non-empty packet (the presence-leak control of App. E).

P.3 KNOWLEDGE-GRAPH PACKET SERIALIZATION

The main experiments use the prose-list serialization. A packet is a header sentence naming the candidate drug (its PrimeKG name) and the relation family, followed by the complete list of related PrimeKG concept names, one per line, sorted alphabetically:

```
Facts from a curated biomedical knowledge graph. {Drug} {relation phrase}
(complete list, {k} entries):
  {concept 1}
  {concept 2}
  ...
```

with relation phrases "is contraindicated in the following conditions" (contra-labs, contra-report), "has the following documented adverse effects" (drug-effect), and "acts on proteins implicated in the following conditions" (mechanism). For mechanism, the retriever applies the two-hop drug→protein→disease template of §3; the packet lists the disease endpoints of that path and does not name the intermediate protein, so the model joins the endpoint to the patient's finding. The typed-triple rendering compared in Table 8 carries the same entries as

```
Knowledge-graph retrieval results (complete; {k} triples):
(:Drug {name: "{Drug}"})-[:CONTRAINDICATED_IN]->(:Condition {name:
"{concept 1}"})
...
```

with relation names `CONTRAINDICATED_IN`, `HAS_ADVERSE_EFFECT`, and `TARGETS_PROTEIN_IMPLICATED_IN`, exactly as written by the v2 preparation step into the frozen item file.

P.4 QUESTION WORDING BY FAMILY

The candidate drug is inserted as the MIMIC prescription string (for example "doxycycline hyclate" or "0.9% sodium chloride").

```
contra-labs, contra-report:
  Question: Is {drug} contraindicated for this patient, given their
  documented
  condition? Answer yes or no.
drug-effect:
  Question: Is at least one of this patient's documented findings a
  listed adverse
  effect of {drug}? Answer yes only if you can identify a specific
  finding that
  appears in the adverse-effect list; otherwise answer no. Answer yes or
  no.
mechanism:
  Question: Does {drug} act on a protein implicated in a condition this
  patient
  has? Answer yes or no.
```

P.5 WORKED EXAMPLES

Each family is illustrated by one positive item and its same-admission matched negative (same record, different candidate drug). Only the blocks that carry the answer are shown in full; patient-level values that are not already reported in the main text are replaced by bracketed placeholders, and omitted list entries are counted. The positive's answer-bearing entries are marked with ◀. Item numbers index the frozen 3,200-item benchmark file.

Contra-labs (items 658 / 1720; Figure 5). Shared record (A–D):

```
A 83-year-old female patient in intensive care.

No coded diagnoses on record.
```

```
Current medications (33): [31 entries omitted; the list includes
doxycycline
  hyclate and vancomycin].
```

Patient modalities (B, D):

```
Laboratory values, first 24 hours: creatinine 4.3-4.4, BUN 91-108,
[14 further analytes omitted].                                    <|
Radiology report for the chest film describes: [15 finding names
omitted].
Chest radiograph, automated image analysis reports: Pneumonia, Pulmonary
Edema.
```

Packet (C, D), positive candidate:

```
Facts from a curated biomedical knowledge graph. Vancomycin is
contraindicated in
the following conditions (complete list, 6 entries):
  agranulocytosis
  auditory system disease
  hearing disorder
  kidney disease                                                   <|
  neutropenia
  pyoureter
```

Packet (C, D), matched negative:

```
Facts from a curated biomedical knowledge graph. Doxycycline is
contraindicated in
the following conditions (complete list, 3 entries):
  Clostridium difficile colitis
  dyskinesia of esophagus
  perinatal necrotizing enterocolitis
```

Questions: "Is vancomycin contraindicated ..." (gold `yes`) and "Is doxycycline hyclate contraindicated ..." (gold `no`). The creatinine value satisfies the renal-impairment rule (> 2.0), the coded history is empty, and only the vancomycin packet lists a renal condition; the state is never named in the prompt. Logged scores (Qwen3.5-35B): positive A 0.011, B 0.026, C 0.002, D 0.651; negative A 0.033, B 0.039, C 0.005, D 0.025.

Contra-report (items 3101 / 2975).

```
A [age]-year-old [sex] patient in intensive care.

No coded diagnoses on record.

Current medications (40): [40 entries omitted; propofol is listed,
omeprazole falls
  beyond the 40-name cap and is named only in the question].

Laboratory values, first 24 hours: [15 analytes omitted].
Radiology report for the chest film describes: [3 findings omitted],
  Congestive heart failure, [2 findings omitted].
  <|
Chest radiograph, automated image analysis reports: Multiple pulmonary
nodules.
```

Positive packet (omeprazole, 8 entries): bone fracture, chronic atrophic gastritis, congestive heart failure ◀, gallbladder disease, liver disease, liver failure, osteoporosis, postmenopausal osteoporosis. Negative packet (propofol, 6 entries): epilepsy, familial hyperlipidemia, hypotensive disorder, inherited lipid metabolism disorder, intracranial arteriosclerosis, respiratory failure. The report finding aligns to PrimeKG *congestive heart failure*, which appears only in the omeprazole packet; no entry of either packet is a near-synonym of any other visible finding.

Drug-effect (items 1662 / 1285).

```
A [age]-year-old [sex] patient in intensive care.

No coded diagnoses on record.

Current medications (33): [31 entries omitted; the list includes 0.9%
sodium
  chloride and chlorhexidine gluconate 0.12% oral rinse].

Laboratory values, first 24 hours: [15 analytes omitted].
Radiology report for the chest film describes: Hypervolemia (finding),
  Left pleural effusion.                                              <|
```

(No imaging block: the study had no finding above threshold, so the block is omitted.) Positive packet: "Sodium chloride has the following documented adverse effects (complete list, 22 entries)": Agitation, Coma, Confusion, Congestive heart failure, Dehydration, Drooling, Edema, Epiphora, Excessive daytime somnolence, Excessive salivation, Fever, Headache, Hypernatremia, Hypervolemia ◀, Loss of consciousness, Pain, Respiratory failure, Seizure, Tachycardia, Thrombophlebitis, Venous thrombosis, Vertigo. Negative packet: "Chlorhexidine has the following documented adverse effects (complete list, 25 entries)": Abnormal joint morphology, Anaphylactic shock, Arthralgia, Arthropathy, Back pain, Bronchitis, Cough, Dysmenorrhea, Erythema, Gingival bleeding, Gingival overgrowth, Glossitis, Headache, Myalgia, Oral ulcer, Pain, Paresthesia, Recurrent aphthous stomatitis, Recurrent fractures, Rhinitis, Scaling skin, Sensory impairment, Sinusitis, Stomatitis, Xerostomia. This pair shows the length matching: the negative list is longer than the positive one (rendered-character ratio 1.10, within the 0.85–1.18 tolerance), so list length cannot stand in for the answer.

Mechanism (items 101 / 2202).

```
A [age]-year-old [sex] patient in intensive care.

No coded diagnoses on record.

Current medications (24): [22 entries omitted; the list includes
alteplase and
  chlorhexidine gluconate 0.12% oral rinse].

Laboratory values, first 24 hours: [15 analytes omitted].
Radiology report for the chest film describes: [5 findings omitted],
  Pulmonary Embolism, [2 findings omitted].                           <|
```

Positive packet: "Alteplase acts on proteins implicated in the following conditions (complete list, 62 entries)", which include *pulmonary embolism (disease)* ◀ (reached through alteplase's target fibrinogen alpha chain, FGA, which is not shown in the prompt); the other 61 entries are omitted here. Negative packet: "Chlorhexidine acts on proteins implicated in the following conditions (complete list, 71 entries)"; none of the 71 entries aligns to, or is a near-synonym of, a finding visible in the record. The protein hop is composed by the retriever. Typed one-hop retrieval (R3) returns drug→protein edges but never the protein→disease hop, so its mechanism path recall is 0 at every budget in the frozen test analysis (App. N.2 reports the pooled recall).

P.6 Five-candidate ranking prompts

The ranking benchmark of §5.2 reuses the patient-side blocks of App. P.2 unchanged (same renderer and caps, including the 40-name medication cap; test items list 13–40 medications, median 37). For the per-candidate scoring, the record is followed by a candidate line and one question per candidate:

```
Candidate drugs received by this patient: 1. {drug 1}; 2. {drug 2}; ...;
5. {drug 5}.

Question: Does this patient have a documented condition in which {drug i}
is
```

```
contraindicated, or a documented finding that is a listed adverse effect
of {drug i}?
Answer yes or no.
```

Condition D in the R4q formulation includes only the queried candidate's contraindication and adverse-effect packets (both relation families, same prose serialization as App. P.3); the earlier five-packet rendering included all five candidates' packets in every question. The five candidates are ranked by their yes-probabilities.