Monopath DAGs: Structured Patient Trajectories from Clinical Case Reports

Anson Zhou[1,*], Aaron Fanous[1,*], Vasiliki Bikia[1,2], Yaowei Deng[1], Shlok Natarajan[1], Sonnet Xu[6], Ank A. Agarwal[1], Akshat Garg[1], Noah Fanous[3], Lyndon Huang[3], Preston Tolbert[5], Jonathan Luu[4], Emily Alsentzer[1,6,†], Roxana Daneshjou[1,7,†]

[1]*Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*
[2]*Institute for Human-Centered AI & Computer Science, Stanford University, Stanford, CA, USA*
[3]*Department of Medicine, University of Texas Dell Medical School, Austin, TX, USA*
[4]*Department of Psychiatry and Behavioral Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA*
[5]*Department of Otolaryngology–Head and Neck Surgery, LSU Health Shreveport, Shreveport, LA, USA*
[6]*Department of Computer Science, Stanford University, Stanford, CA, USA*
[7]*Department of Dermatology, Stanford University, Stanford, CA, USA*

[*]*Equal contribution.* [†]*Co-senior authors.*
E-mail: aron7628@stanford.edu

Clinical machine learning has limited access to rare diseases, atypical treatment responses, and uncommon care pathways, many of which are documented in clinical case reports but remain difficult to model from free text. We introduce Monopath Directed Acyclic Graphs (DAGs), a structured representation that converts case reports into ordered patient states linked by explicit clinical transitions. We construct Monopath DAGs from 485 lung cancer case reports and evaluate their fidelity to the source narratives. Graph-based reconstruction achieves a ClinicalBERT BERTScore of $F_1 = 0.795 \pm 0.041$; human review finds 94% node semantic validity, 87% node temporal validity, 94% edge temporal validity, and 98% edge semantic validity. We next evaluate the representation in embedding and generation tasks. Node-text embeddings derived from Monopath DAGs achieve higher Calinski–Harabasz scores than embeddings of the original reports (136 vs. 78; difference 95% CI 34–79), although this analysis does not encode graph order or edges. In a blinded physician evaluation, graph-conditioned synthetic cases were rated approximately 0.45 points higher on a 5-point scale than controls generated from the same source report without the graph for timeline validity, clinical actionability, and perceived safe actionability ($p = 0.005$, $p = 0.005$, and $p = 0.001$, respectively). Clinical validity did not differ significantly between conditions. These results show that structuring case reports as explicit patient states and transitions preserves source-trajectory information and produces synthetic narratives that physicians judge more temporally coherent and actionable than those generated from a flat textual representation.

Keywords: clinical trajectories; directed acyclic graphs; synthetic clinical data; case reports; clinical language models; patient trajectory modeling

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Healthcare remains a challenging domain for AI model development. Although clinical AI models have increasingly achieved strong performance on retrospective benchmarks, these gains have failed to consistently translate to prospective clinical evaluation and routine deployment.[1–3] Limited and inconsistent clinical data contribute to this translation gap, particularly for rare diseases and uncommon clinical presentations. As a result, synthetic data generation has become a promising technique for ML practitioners to expand coverage of these underrepresented cases.

Although synthetic data can expand the coverage of rare clinical cases, additional samples do not guarantee the preservation of informative qualities. A generated case may remain coherent and clinically plausible while omitting findings that alter diagnostic reasoning or subsequent actions. Clinical summarization studies have demonstrated this failure: compressed representations can preserve a high-level account of a patient's medical history while excluding information that clinicians need for downstream decisions.[4,5] Synthetic generation therefore depends on preserving information that constrains subsequent clinical decisions, not only representing information sufficient to maintain clinical validity.

Longitudinal EHRs introduce additional loss of representational fidelity across the clinical data pipeline. Clinical encounters encompass temporally-distributed observations, interventions, treatment responses, and outcomes, which clinicians selectively document, and downstream systems subsequently normalize, aggregate, or summarize for analysis. Each transformation can alter the temporal and contextual relationships between events, including treatment-response sequences, unresolved findings, and dependencies between prior observations and subsequent actions. Recent longitudinal EHR studies demonstrate that event-ordering itself carries information associated with later clinical outcomes.[6] Consequently, a representation may retain sufficient information to characterize the current clinical state while failing to preserve the relationships that constrain subsequent clinical decisions.

Current evaluations of synthetic clinical data primarily assess distributional fidelity, task-specific utility, privacy risk, and broad measures of clinical plausibility.[7] These criteria determine whether generated data reproduce statistical properties of observed cohorts, retain performance on predefined prediction tasks, or satisfy general clinical-validity constraints. They do not directly assess whether a synthetic case preserves the sequence of findings, interventions, and responses needed to discriminate between subsequent clinical actions. As a result, two cases may achieve comparable validity or utility scores while materially differing in temporal coherence, actionability, and the safety of the decisions they support. Clinical representations can therefore be measured by clinicians' perceived ability to act on them.

Narrative case reports provide a useful setting in which to study this gap. Clinicians use case reports to document unusual presentations, diagnostic workups, interventions, treatment responses, and longitudinal disease progression. These reports often retain the sequence of findings, decisions, and responses that condensed EHR representations omit, while disproportionately representing rare or diagnostically challenging cases. However, case reports encode these relationships in free text. Temporal dependencies, decision points, and transitions between clinically-meaningful states often remain implicit within the narrative, limiting

systematic extraction and large-scale modeling.

Although case reports may describe competing diagnoses, concurrent findings, and alternative management considerations, they ultimately document a single observed clinical course. Therefore, we can represent each case as a directed path through clinically-meaningful patient states and transitions. We retain this path within a directed acyclic graph (DAG) structure so that individual states and transitions remain explicitly addressable and the representation can support graph operations such as sub-trajectory selection. This study uses the complete observed trajectory rather than selected subgraphs, and tests whether conditioning generation on it changes how clinicians judge the resulting cases.

We present Monopath DAGs as structured intermediates for synthetic case generation. The framework extracts clinically-relevant states and transitions from case-report narratives, organizes them into temporally-ordered trajectories, and conditions generation on the resulting graph structure. This design allows us to compare synthetic cases generated with and without explicit trajectory structure while holding the underlying source material constant.

We evaluated this framework on 485 curated lung cancer case reports and 200 rare-disease cases across four disease groups. For each report, we generated two synthetic case summaries: one conditioned on the report's graph and one generated from the report without the graph (the control). Physicians rated each summary on clinical validity, timeline validity, clinical actionability, clinically safe action, and language appropriateness. This evaluation directly tests whether cases with similar clinical validity can still differ in the information they preserve for subsequent clinical decision-making.

This work makes four key contributions:

- **A decision-relevant evaluation framework** for synthetic clinical data that distinguishes clinical validity from timeline validity and clinical actionability and clinically safe action.
- **Evidence that clinical validity ratings can understate differences in clinical actionability**: graph conditioning raised clinical actionability by about 0.45 points on a 5-point scale, whereas the difference in clinical validity (+0.21, 95% CI −0.05 to 0.48) was smaller and not significant in the primary model.
- **A structured trajectory representation** that extracts ordered, ontology-grounded Monopath DAGs from free-text case reports, representing each case as a sequence of patient states and transitions.
- **A synthetic data generation pipeline and evaluation dataset** spanning 485 lung cancer case reports and 200 rare disease cases across four disease groups.

2. Related Work

Clinical Narrative Information Extraction. Information extraction from clinical text has progressed from identifying entities to recovering the relations among them. Rule- and vocabulary-based systems such as cTAKES and MetaMap normalize entities to controlled terminologies. Statistical and transformer-based methods extend this to temporal events and their ordering, including semi-supervised CRFs for temporal event detection in EMRs,[8] ClinicalBERT[9] and BioGPT,[10] and GraphTrex on i2b2 TempEval.[11] Other methods extract

cause–effect relations through dependency patterns.[12] Other work targets case reports directly, through LLM-based extraction,[13] graph-based representations of cardiovascular case reports in CREATe,[14] and the 167,000 case summaries of PMC-Patients for retrieval.[15] These approaches recover entities, relations, temporal events, or condensed summaries, but do not represent a case as the sequence of patient states it describes and the transitions between them. In contrast, we represent each report as an ordered sequence of patient states, with the transition between each pair of consecutive states stated explicitly.

Patient Trajectory Modeling. Trajectory models represent a patient's course as a sequence of structured events. Sequence models such as RETAIN,[16] BEHRT,[17] and Med-BERT[18] treat diagnosis codes or visits as tokens in an ordered sequence, and graph-based methods such as GraphCare[19] add knowledge-graph structure over codes and visits for prediction. In these models a trajectory is a sequence of coded events drawn from structured records. They need coded inputs that case reports lack, and they omit the states and decisions a narrative describes. Few studies recover such trajectories from case reports.

Synthetic Patient Data Generation. Researchers generate synthetic patient data to offset data scarcity and privacy limits. Rule-based simulators such as Synthea[20] encode disease progression by hand, while learned generators synthesize longitudinal records with GANs and encoder–decoder models (EHR-M-GAN,[21] EHR-Safe[22]), hierarchical Transformers over code sequences (HALO[23]), or language models for discharge summaries.[24] These generators are typically evaluated on distributional fidelity, downstream utility, and privacy,[7] criteria that do not test whether a generated case preserves the information needed for the next clinical decision. Generators learned from cohorts also see few examples of the rare and atypical trajectories that case reports document.

Here, we recover state-level trajectories from case narratives, condition generation on them, and ask physicians whether the generated cases support the next clinical decision.

3. Methods

3.1. *Modular Framework for Patient Trajectory Graph Extraction*

We present a modular framework for converting free-text clinical case reports into structured patient trajectories represented as *Monopath Directed Acyclic Graphs (Monopath DAGs)*.

3.2. *Task Definition*

We represent each clinical case report $d_i \in \mathcal{D}$ as a directed acyclic graph $G_i = (V_i, E_i)$. Each node corresponds to a patient state, defined as a snapshot of the patient as perceived by the clinician at a given point in the trajectory. States reflect the clinical information available at that point rather than an assumed complete representation of the patient's underlying condition. Nodes are annotated with free-text descriptions, structured clinical attributes, and, when available, explicit timestamps. Directed edges represent transitions between states and are labeled by transition type and the clinical entities associated with each transition. See

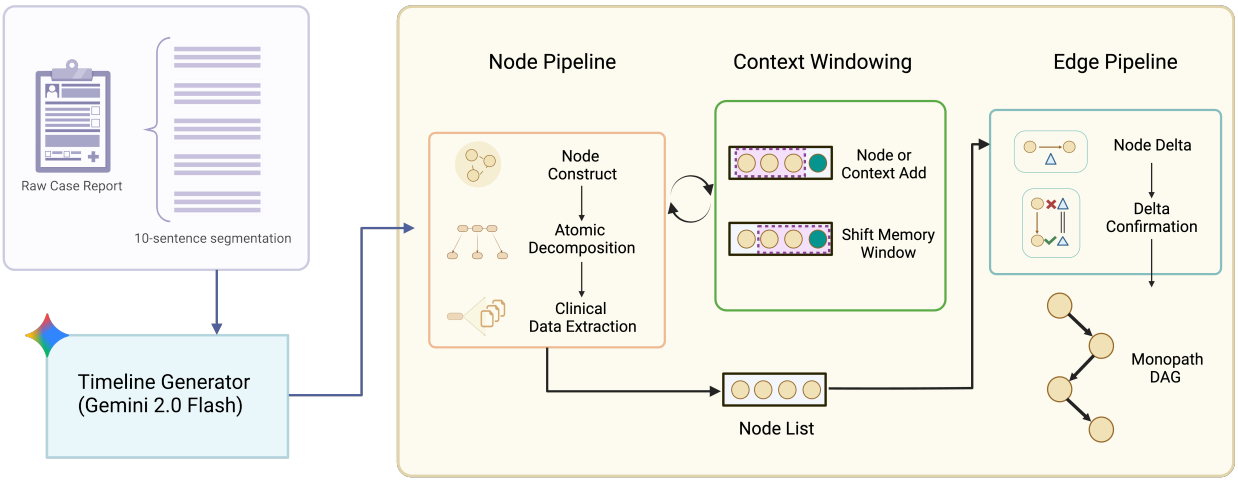


Fig. 1: DAG generation pipeline for patient trajectory modeling.

Supplement S1 for the complete formal definition of patient states and graph structure.[a]

3.3. *Patient Trajectory Generation*

Trajectory construction follows three structural constraints: (i) the graph is acyclic, reflecting the forward progression of clinical time; (ii) each node has at most one incoming and one outgoing edge, and the graph is connected; and (iii) nodes are temporally ordered using explicit timestamps when available and inferred narrative order otherwise. These constraints define the Monopath representation and reduce ambiguity during trajectory construction. Topology validation flags graphs that violate these constraints (Section 5).

Preprocessing & Timeline Extraction. Each case report is retrieved from PubMed Central as UTF-8 HTML, with non-narrative boilerplate removed. We segment the remaining narrative into 10-sentence blocks and pass them to the timeline generator $T_i = \texttt{PatientTimeline}(p_1, \ldots, p_k)$. In pilot runs, we judged qualitatively that Gemini 2.0 Flash extracted timelines more consistently than Llama 3.1 8B and 70B at acceptable cost. All language-model steps use Gemini 2.0 Flash at $T = 0.2$ with an approximately 128k-token context window.

Node Construction. We divide each patient timeline T_i into four-sentence blocks c_j and process them sequentially to construct patient states. For each block, we generate a node

$$v_j = \{\texttt{node_id}, \texttt{step_index}, s_j, a_j, t_j\},$$

where s_j is the free-text state description, a_j contains structured clinical attributes, and t_j is an optional timestamp. To preserve local temporal context, each node v_j is generated using a compressed representation of the three preceding states.

[a]Supplementary material, prompts, code, and data:
`https://github.com/DaneshjouLab/MonopathDAGs/tree/main/supplement`

We recursively decompose each state description s_j into atomic clinical facts, defined as statements containing a single clinical assertion about the patient state. We then map these facts to structured ontology terms. Facts that cannot be mapped are retained as free text to preserve information. We merge nodes with matching identifiers (appending new content and combining clinical data), apply a final model pass to remove duplicate states, and reindex the sequence (Supplement S3). We do not model negation or uncertainty explicitly; temporality is carried by node order, with timestamps only when stated.

Edge Construction. Given the final node sequence $\{v_1, \ldots, v_n\}$, we construct an edge between each adjacent pair (v_i, v_{i+1}) by identifying the clinical change between states. Each edge is represented as

$$e_{i \to i+1} = \{\texttt{edge_id}, \tau_{ij}, \phi_{ij}\},$$

where τ_{ij} is a free-text transition description and ϕ_{ij} is a structured event containing the trigger type, clinical domain, and associated ontology terms. We use three few-shot demonstrations (Supplement S5) to keep transition labels consistent.

Graph Serialization & Metadata Tracking. Each patient trajectory is serialized as a JSON directed graph following a consistent schema with ontology-grounded labels and structured metadata (Supplement S3). We enforce schema compliance during construction while deferring full structural validation for downstream evaluation.

3.4. *Lung Cancer Case Report Dataset*

We curated a dataset of full-text lung cancer case reports from PubMed Central (PMC). Using the Entrez API, we searched for open-access case reports published from 2019 onward using lung cancer-related terms, including lung cancer, lung carcinoma, non–small cell lung cancer (NSCLC), small cell lung cancer (SCLC), mesothelioma, pulmonary carcinoma, and bronchogenic carcinoma, matched against title and abstract fields. This search yielded 550 open-access case reports.

4. Evaluation and Utility

4.1. *Graph Construction and Fidelity*

We assess graph fidelity with human review, BERTScore, and structural checks.

Human Graph Review. For each edge and node of a random sampling of 30 graphs, two graduate students compare each generated graph with the source case report.Reviewers assess whether each node and edge preserves the patient states represented in the source case, the clinically relevant information contained within those states, and the temporal ordering of the trajectory (Supplement S4). The graduate students made 568 annotations of 445 elements (243 nodes, 202 edges) from 30 graphs; on the 123 elements annotated twice, agreement on G1 (correct order) and G2 (accurate content) was 94% and 92%. Cohen's κ is 0.73 for G1; for G2, nearly all ratings were positive, which makes κ uninformative.

LLM-Based Narrative Reconstruction and BERTScore Evaluation. We reconstruct the input case narrative from the generated graph using Gemini 2.0 Flash. The model receives a graph's nodes and edges and is prompted to generate a coherent textual summary that

aims to mirror the structure and content of the original case. The reconstructed narrative $\hat{R}$ is compared to the original case report R using BERTScore,[25] which measures semantic similarity via contextualized embeddings from ClinicalBERT. Reconstruction-based BERTScore captures presumed semantic coverage between the graph and the source narrative through the distance, it does not explicitly detect hallucinated content or subtle clinical inaccuracies, motivating complementary human evaluation. We calculate BERTScore with Bio _ClinicalBERT[9] (layer 12, without idf or rescaling), calculated on the full texts: each text is embedded in consecutive 510-token windows, as in Section 4.2, and all tokens enter the greedy matching. Scores may partly reflect self-consistency, since one model constructs and reconstructs the graphs, or the model's prior exposure to published reports.

Topology Validation. We sanity checked the structural integrity of the patient graphs through topological audits. We checked acyclicity by applying depth-first search, count connected components, and flag malformed graphs.

4.2. *Clustering of Patient Trajectories*

We embedded graphs and case reports with ClinicalBERT. A graph was represented by the mean of its node-text embeddings (node order and edges not encoded); a report, by the mean embedding of consecutive 510-token windows, so that no text was truncated (case presentations average 1,137 ClinicalBERT tokens, median 967, and 73% exceed 512). To test whether edges add information, we also embedded each edge description with the same model and represented each graph by the mean over all node and edge embeddings. We analyzed embeddings in 768 dimensions and after PCA or UMAP reduction to 10, and assessed cluster structure with K-means ($k \in [2, 10]$, selected by silhouette[26]) and the Calinski–Harabasz (CH) index.[27] For the 484 patients with metastasis status, annotated manually for each report, we computed CH with patients partitioned by status, and the cross-validated AUC of logistic regression predicting status from each embedding. 95% CIs are from patient-level bootstrap resamples; settings are in Supplement S9.

4.3. *Synthetic Trajectories and Clinical Expert Review*

Synthetic Generation. We prompted a language model to generate 77 narrative vignette pairs from conditioning on the source or from the graph. The control extracts a list of clinical events from the full report in one pass and writes a narrative from it; the graph-conditioned narrative is generated from the ordered nodes and edges of G_i. Both use nearly identical prompting, and the same case per narrative pair(Supplement S5).

Clinician-centered evaluation criteria

Q1. (Clinical validity): Baseline check that case actions are medically sound.
Q2. (Timeline validity): Assesses whether the case progression follows a clinically appropriate order, capturing temporal reasoning.
Q3. (Clinical actionability): Measures whether the physician judges the case specific and granular enough to guide the next clinical decision.
Q4. (Clinically safe action): Measures whether the physician judges the information suffi-

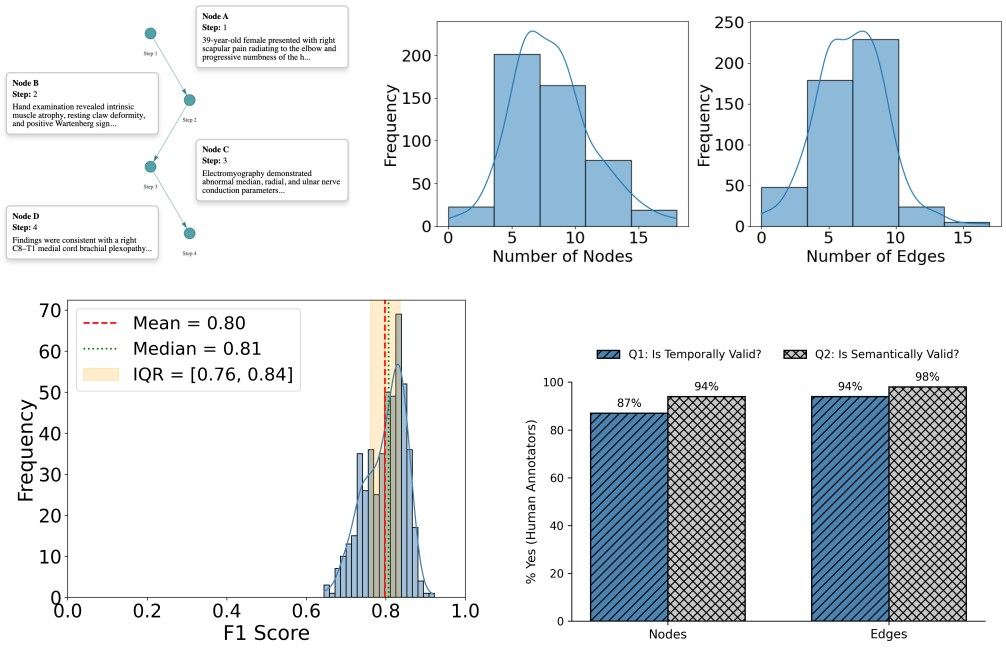


Fig. 2: **Monopath DAG trajectory representations and evaluation.** (a) Schematic of a Monopath DAG representing a patient case. (b) Distribution of node and edge counts across constructed graphs. (c) BERTScore F1 similarity between reconstructed narratives and original texts. (d) Human evaluation comparing graph fidelity and ordering accuracy.

cient to make a safe decision with confidence, reflecting the physician's decision certainty.
Q5. (Appropriate language): Evaluates whether the narrative uses medical language appropriately, avoiding generic or ambiguous phrasing.

Clinical Expert Review. Resident physicians (PGY-1 to PGY-3), blinded to narrative generation conditions, rated each narrative on 5-point Likert scales. They receive cases in the same order, to ensure overlap coverage but generation conditioning order is randomized. Three of four physicians contributed 227 ratings (105 graph-conditioned, 122 control) across 76 cases: in 98 physician–case combinations they rated both versions (97 with complete answers), and in 31 they rated one version only (24 control, 7 graph-conditioned). The fourth rated one narrative and is excluded. We fit a linear mixed-effects model to each question, with crossed random intercepts for physician and case to absorb harsh raters and hard cases, and applied Holm[28] correction across the five questions. The graph-conditioned arm passes through more processing steps than the control, so the two arms do not receive equal input. The five questions (Q1–Q5), reviewed by medical experts, cover clinical validity, timeline validity, clinical actionability and clinically safe action, and language. Q1 assesses medical correctness in isolation, whereas Q4 evaluates whether the information provided supports confident and safe decision-making under clinical uncertainty (Supplement S4).

5. Results and Discussion

From the initial set of 550 case reports, we remove non-English entries and case series, resulting in a curated dataset of 485 individual narratives. Semantic fidelity of the graph-derived reconstructions is evaluated using BERTScore, yielding a mean F1 of 0.795 (95% CI 0.791–0.798), median 0.802, interquartile range [0.770, 0.824], range 0.643–0.879, and standard deviation 0.041. Reconstructions remain semantically close to their sources, although BERTScore does not detect omissions or clinical errors. Precision and recall components of BERTScore are 0.807 and 0.786, respectively (Supplement S2). On average, graphs contained 8 ± 3 nodes and 7 ± 3 edges. Node temporal validity is 87% and node semantic validity 94%; edge temporal validity is 94% and edge semantic validity 98%, indicating that most reviewed elements (30 graphs) were correctly ordered and accurate to the source (Figure 2d). Most reviewed nodes and edges were therefore accurate to the source and correctly ordered; node ordering was the least accurate of the four measures.

Graph embeddings yielded higher Calinski–Harabasz indices than case-report embeddings in every space: 136 (95% CI 119–156) vs. 78 (69–90) in the original space, 224 (198–258) vs. 147 (127–171) after PCA, and 738 (649–841) vs. 358 (318–408) after UMAP, with the difference excluding zero in each case. Both representations separated metastatic from non-metastatic patients to a similar degree (CH 10.6 vs. 7.0; difference 3.6, 95% CI −1.6 to 10.7), and the AUC for predicting metastasis status was slightly lower for graphs, not significantly different (0.73 vs. 0.77; difference −0.04, 95% CI −0.09 to 0.01). Graphs reduce each report (mean 1,137 tokens) to about eight states. When we embedded each edge description with the same model and added the edges to the nodes, AUC for metastasis status rose by 0.01–0.03 in every cross-validation split and setting we tried, but the patient-level 95% CI for the difference included zero (−0.01 to 0.06). With edges, AUC approached that of the full text (0.755 vs. 0.769), and edges alone scored slightly below nodes. Adding edges lowered cluster compactness (CH 101 vs. 136) (Supplement S9). Graph embeddings therefore form more compact clusters than the report text, while metastasis status is predicted with similar accuracy from both, although each graph contains about eight states. Adding edges raised the AUC slightly but lowered compactness.

In a blinded evaluation, we analyzed the three of the four participating resident physicians (PGY-1 to PGY-3) who rated both versions of cases, comparing control-generated and graph-conditioned synthetic cases across the five clinical criteria (Figure 3, Table 1). PGY-level physicians were selected because they routinely perform frontline clinical reasoning and decision-making, making them well positioned to assess timeline coherence, actionability, and safety in narrative cases. We did not detect a difference in clinical validity (Q1; mixed-model estimate +0.21, 95% CI −0.05 to 0.48, $p = 0.12$), whereas graph-conditioned cases were rated higher on timeline validity (Q2, +0.46, 0.18 to 0.73, $p = 0.005$), clinical actionability (Q3, +0.45, 0.17 to 0.72, $p = 0.005$), and clinically safe action (Q4, +0.45, 0.21 to 0.68, $p = 0.001$; means 2.74 vs. 2.31, both below the scale midpoint, with 19% vs. 13% of ratings 4–5). Excluding physician A or B removes the significance of timeline validity and both actionability measures after Holm correction; excluding physician C does not, and without C, clinical validity also differs (+0.36, $p = 0.008$; Supplement S8). Agreement between physicians on individual ratings

was low (Spearman ρ −0.10 to 0.37). Language appropriateness (Q5) did not differ significantly after correction ($p = 0.058$). Graph conditioning therefore raised ratings of timeline validity, clinical actionability, and clinically safe action by about 0.45 points on a 5-point scale, while the difference in clinical validity was smaller and not significant. Ratings of clinically safe action stayed below the scale midpoint in both conditions, the effects depended on physicians A and B, and agreement between physicians was low. Although the binary evaluation schema (G1: order, G2: accuracy) provides only a coarse-grained view of graph fidelity, it is designed to complement automated semantic metrics such as BERTScore with a minimal human check that could be performed within the constraints of clinician time. This design emphasized inter-reviewer agreement and feasibility, but it does not capture node informativeness or the clinical validity of edges on its own. We consider this a key limitation of the current evaluation and plan to extend our framework with more fine-grained protocols, including explicit node-level assessments and validation of transition plausibility. As noted in Section 4.3, clinician evaluation compared synthetic narratives with control narratives on textual criteria; physicians rated narratives only; they did not view the graphs or compare formats. In this study, graphs are used as structured intermediates to support generation, rather than as interfaces intended for direct clinical decision-making.

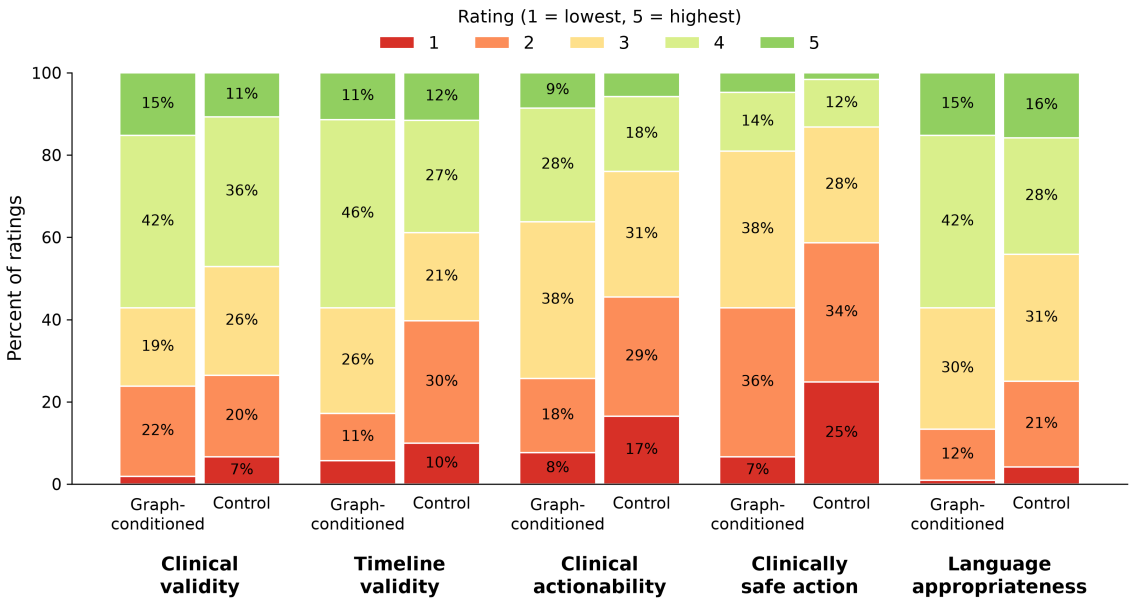


Fig. 3: Physician ratings of graph-conditioned and control narratives.

6. Exploratory Application to Rare Diseases

While our primary experiments focus on lung cancer—chosen for its prevalence in case report literature, detailed longitudinal trajectories, and heterogeneous treatment paths—the pipeline is agnostic to disease.

Table 1: Linear mixed-effects estimates (graph-conditioned − control) with crossed random intercepts for physician and case, across five Likert-scale questions. Four physicians took part; the 227 ratings analyzed, across 76 cases, came from the three physicians who contributed paired ratings. CIs are unadjusted; p values are Holm-adjusted across questions ($^*p < 0.05$).

Question	**Estimate**	**95% CI**	p
Clinical validity	+0.21	[−0.05, 0.48]	0.117
Timeline validity	+0.46	[0.18, 0.73]	0.005*
Clinical actionability	+0.45	[0.17, 0.72]	0.005*
Clinically safe action	+0.45	[0.21, 0.68]	0.001*
Appropriate clinical language	+0.28	[0.03, 0.53]	0.058

6.1. *Rare Diseases.*

We conducted additional experiments on four rare diseases across body systems (Gaucher, Wilson's, Charcot–Marie–Tooth, and Retinitis Pigmentosa), using 50 case reports per disease (200 total, see Supplement S2). These conditions were selected because they span distinct organ systems, disease progression patterns, and diagnostic pathways. We report this as an exploratory analysis of clustering only (Supplement S2), without human or downstream evaluation; it does not establish generalization or clinical utility.

7. Conclusion and Future Work

Monopath DAGs provide a structured intermediate for converting case reports into synthetic clinical narratives. The graphs are faithful to the source on most reviewed elements, yield more compact embeddings than the original text, and support synthetic narratives that physicians align more with timeline validity and clinical actionability and perceived clinically safe action. Clinical validity did not differ significantly. These findings suggest that evaluating synthetic cases on clinical plausibility alone can miss differences in meaningful components that lead to clinical decision making.

Limitations. Our evaluation is limited to recent, English-language lung cancer case reports and may not generalize across specialties, languages, or narrative conventions. Monopath DAGs enforce a single dominant clinical thread, collapsing concurrent events, branching trajectories, and cyclic complications. The extraction pipeline also depends on general-purpose LLMs and retains errors in negation, temporal ordering, and dosage extraction that are not fully captured by BERTScore or our limited human evaluation. Synthetic trajectories were evaluated through small-scale expert review, without model-leakage audits or downstream task benchmarking. We also do not explore privacy leakage from the case reports themselves. The single-path representation trades expressivity for a simple temporal structure that can be checked against the source. We did not quantify the information lost to the single-path assumption. The generation comparison does not separate segmentation, ordering, and composition, and the two arms differ in processing steps and prompt length. Case reports are curated, single-author narratives of unusual courses, whereas EHR trajectories are fragmented across

many notes and authors; we have had some success testing this on synthea data, but not at a larger scale and note that case reports are quite different from EHR knowledge.

Future Work. We plan to extend Monopath DAGs to *Multipath DAGs* with controlled branching to represent clinical uncertainty, reversible events, and alternative pathways. We will also investigate biomedical language models and graph-imputation methods to improve extraction of under-specified clinical states. These extensions could support trajectory-informed outcome prediction, counterfactual synthetic generation, cross-disease benchmarking, and trajectory-aware case matching for clinical decision support.

References

1. Why high scores do not mean application readiness for health ai, *Nature Medicine* (2026).
2. R. M. W. W. Tseng, L. C. Ong, J. H. L. Goh, Y. Chen, T. Chen, E. Lum and Y.-C. Tham, Prospective real-world implementation of deep learning systems in healthcare: a systematic review guided by implementation science, *npj Digital Medicine* **9**, p. 172 (2026).
3. Show us the evidence for the value of medical ai, *Nature Medicine* **32**, 1163 (2026).
4. J. D. Oliveira, H. D. P. Santos, A. H. D. P. S. Ulbrich, J. C. Couto, M. Arocha, J. Santos, M. M. Costa, D. Faccio, F. O. Tabalipa and R. F. Nogueira, Development and evaluation of a clinical note summarization system using large language models, *Communications Medicine* **5**, p. 376 (2025).
5. A. X. You, N. M. Kahl, A. Patel, J. Cao, E. M. Castillo, P. S. Bedmutha, T. Chan, K. Singh and C. A. Longhurst, Artificial intelligence note summarization in the emergency department, *JAMA Network Open* **9**, e2618021 (06 2026).
6. B. Edelman, H. Kim and J. Skolnick, Scalable discovery and validation of order-specific electronic health record event trajectories for interpretable adverse-outcome risk estimation, *Journal of Biomedical Informatics* **177**, p. 105008 (2026).
7. Y. Kim, J. Lee, S. B. Koh and K. H. Lee, An integrated evaluation framework for synthetic clinical data in severely imbalanced settings: fidelity, privacy-risk profiling, and diagnostic utility, *BMC Medical Informatics and Decision Making* (2026).
8. G. Moharasar and T. B. Ho, A semi-supervised approach for temporal information extraction from clinical text, in *2016 IEEE RIVF international conference on computing & communication technologies, research, innovation, and vision for the future (RIVF)*, 2016.
9. E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann and M. B. A. McDermott, Publicly available clinical bert embeddings (2019).
10. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon and T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* **23**, p. bbac409 (09 2022).
11. R. Chaturvedi, P. Baghershahi, S. Medya and B. Di Eugenio, Temporal relation extraction in clinical texts: A span-based graph transformer approach, *arXiv preprint arXiv:2503.18085* (2025).
12. M. A. H. Kabir, A. Almulhim, X. Luo and M. A. Hasan, Informative causality extraction from medical literature via dependency-tree-based patterns, *Journal of Healthcare Informatics Research* **6**, 295 (may 2022).
13. V. Sciannameo, D. J. Pagliari, S. Urru, P. Grimaldi, H. Ocagli, S. Ahsani-Nasab, R. I. Comoretto, D. Gregori and P. Berchialla, Information extraction from medical case reports using openai instructgpt, *Computer Methods and Programs in Biomedicine* **255**, p. 108326 (2024).
14. Y. Zhou, W.-T. Chen, B. Zhang, D. Lee, J. H. Caufield, K.-W. Chang, Y. Sun, P. Ping and W. Wang, Create: Clinical report extraction and annotation technology (feb 2021).

15. Z. Zhao, Q. Jin, F. Chen, T. Peng and S. Yu, A large-scale dataset of patient summaries for retrieval-based clinical decision support systems, *Scientific Data* **10** (2022).
16. E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart and J. Sun, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism (2017).
17. Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi and G. Salimi-Khorshidi, Behrt: transformer for electronic health records, *Scientific reports* **10**, p. 7155 (2020).
18. L. Rasmy, Y. Xiang, Z. Xie and C. S. and Jianying Hu, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, *npj Digital Medicine* **4**, p. 86 (jun 2021).
19. P. Jiang, C. Xiao, A. Cross and J. Sun, Graphcare: Enhancing healthcare predictions with personalized knowledge graphs, *arXiv preprint arXiv:2305.12788* (2023).
20. J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher and S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, *Journal of the American Medical Informatics Association* **25**, 230 (mar 2018).
21. J. Li, B. J. Cairns, J. Li, A. D. Gilden, M.-I. Z. Li and J. H. Saltz, Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications, *npj Digital Medicine* **6**, p. 98 (jun 2023).
22. J. Yoon, M. Mizrahi, N. F. Ghalaty, M. S. Oberst and P. Szolovits, Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records, *npj Digital Medicine* **6**, p. 141 (sep 2023).
23. B. Theodorou, C. Xiao and J. Sun, Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model, *Nature Communications* **14**, p. 5305 (sep 2023).
24. M. Falis, A. P. Gema, H. Dong, L. Daines, S. Basetti, M. Holder, R. S. Penfold, A. Birch and B. Alex, Can gpt-3.5 generate and code discharge summaries?, *Journal of the American Medical Informatics Association* **31**, p. 2284–2293 (oct 2024).
25. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
26. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* **20**, 53 (1987).
27. T. Caliński and J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods* **3**, 1 (1974).
28. S. Holm, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65 (1979).