MediScribes at MultiClinSum-2: A Dutch-English Clinical Summarization System

TNO at CLEF 2026

Maaike H. T. de Boer[1,*], Quirine T. S. Smit[1,2] and Roos M. Bakker[1,3]

[1]Dutch Organisation for Applied Scientific Research (TNO), Department Data Science, Anna van Buerenplein 1, 2595 DA, The Hague, The Netherlands

[2]Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

[3]Leiden University Centre for Linguistics, Reuvensplaats 3 - 4, 2311 BE Leiden, The Netherlands

Abstract

Medical documentation is currently burdening the medical treatment process. One way to ease the process is to automate the documentation process and summarization. The automation should, however, be factually correct. In this paper, we show the results of our participation in the MultiClinSum-2 challenge. We focused on Dutch and English. To account for factuality, we researched how we can use knowledge graphs to enhance clinical summarization. We compare methods that use a small version of the FHIR ontology and automatic knowledge graph creation using GraphRAG. The results show that, on the test set of this challenge, including knowledge graphs for the generation of summaries does not improve performance for the selected evaluation metrics. Our method using GraphRAG slightly outperforms the LLM-only method on faithfulness and consistency, but the other methods currently fall short, especially for the ROUGE metrics. Future work involves further experimentation with our methods on conversational summarization of medical consults.

Keywords

Summarization, MultiClinSum, Knowledge Graphs, GraphRAG, LLM

1. Introduction

In the medical domain, administrative tasks can consume up to 30% of clinicians' working hours, diverting attention from patient care and contributing to physician burnout [1, 2]. Medical documentation is, however, important within the medical treatment process. In this work, we aim to lessen the workload for medical specialists by automatically creating summaries from medical documentation. The European MediSpeech project aims to reduce administrative waste in healthcare by creating an open digital healthcare ecosystem for automated medical reporting[1]. One of the research topics is to automatically summarize conversations using Automatic Speech Recognition (ASR) techniques and Large Language Models (LLMs). The summarized conversations should follow the same structure as the documentation requires, and they should be factually correct. In this part of the project, we are responsible for the Dutch (and English) summarization based on the transcribed conversations.

To further the research for this project, we participate in English and Dutch runs of the MultiClinSum-2 task by BioASQ [3, 4]. We explore different ways of creating knowledge graphs to research how knowledge graphs can be used to enhance the factuality of clinical summarization.

In the next section, related work on summarization, the first edition of the challenge and knowledge sources from the medical domain is provided. The experimental setup with the dataset, methods and evaluation metrics is explained in Section 3. Section 4 contains the results, whereas the discussion is provided in Section 5. The final section concludes the paper and shows an overview of future work.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

maaike.deboer@tno.nl (M. H. T. d. Boer); quirine.smit@tno.nl (Q. T. S. Smit); roos.bakker@tno.nl (R. M. Bakker)

0000-0002-2775-8351 (M. H. T. d. Boer); 0009-0008-8890-4261 (Q. T. S. Smit); 0000-0002-1760-2740 (R. M. Bakker)

[1]https://itea4.org/project/medispeech.html

2. Related Work

Summarization is a very common task in the field of Natural Language Processing (NLP), dating back to the 1950s [5]. It has gone through the phases of using feature-based systems, deep learning frameworks and now it is dominated by Large Language Models. Summaries can be either extractive (directly choosing different sentences from the original text), or abstractive (novel sentences are created). Many datasets have been created over the years, such as the summarization of the Document Understanding Conferences (DUC[2]), Text Analysis Conference (TAC[3]), Gigaword [6] or CNN Daily Mail dataset [7], and many evaluation metrics were proposed [5], such as comparison to a golden standard summary, lexical overlap metrics (e.g. ROUGE), semantic overlap metrics (e.g. BERTScore) or LLM-as-a-judge metrics (fluency, coherence) [8].

In the MultiClinSum challenge of 2025 [8], the top-performing teams used abstractive approaches with decoder-only architectures fine-tuned on biomedical corpora. A few teams also combined LLMs with knowledge-based approaches, such as the ExtraSum team which compared graph-based, concept-based, topic-based and cluster-based approaches [9]. Team MedCOD extracts medical keywords and incorporates them into a structured prompt to provide contextual input [10]. This approach, together with finetuning, leads to higher performance for non-English languages. Team JohannaUE uses an agentic multilingual framework and combines abstractive and extractive approaches [11]. They use a LangGraph-based architecture which has modules such as a NER-guided entity preserver, knowledge graphs and prompted language models. The recommendations of last year involved examination of the impact of clinical entities for automatic summarization approaches [8], which we have incorporated in one of our methods.

In the medical domain, standardization and the use of ontologies is very common [12]. They provide structured, machine-readable frameworks that contain biomedical knowledge, concepts, and the relationships between them. Large sources of knowledge include SNOMED [13], UMLS [14] and MeSH [15]. The Fast Healthcare Interoperability Resources (FHIR) standard is a framework used to represent clinical information and is often used in Electronic Health Records (EHRs) [16]. To the best of our knowledge, there are no papers yet that use FHIR to enhance LLMs for summarization.

3. Experimental Setup

3.1. MultiClinSum-2 Data and Evaluation

The MultiClinSum-2 challenge focuses on the automatic summarization of lengthy clinical case reports [3]. There are 15 languages available, including Dutch and English. The training dataset contains about 26000 full case-summary pairs in all languages and the sample dataset contains about 50 full case-summary pairs. The submitted test dataset contains 1000 cases. We have only participated in the English and Dutch cases.

The created summaries are evaluated on the test set using different metrics [3]. For the lexical overlap several ROUGE versions are used: ROUGE1 measures the unigram overlap between the ground truth summary and the generated summary, ROUGE2 does the same with bigrams, ROUGE-Lsum takes the longest common subsequence overlap (sentence-by-sentence). For the semantic similarity BERTScore with multilingual DeBERTa embeddings is used. LLM-as-a-judge scores on faithfulness (only information present in the original case), completeness (covers key clinical information following CARE guidelines), fluency (grammatical correctness and readability in the language) and consistency (factual equivalence across all language versions) to measure other dimensions [3].

[2] https://duc.nist.gov/
[3] https://tac.nist.gov/

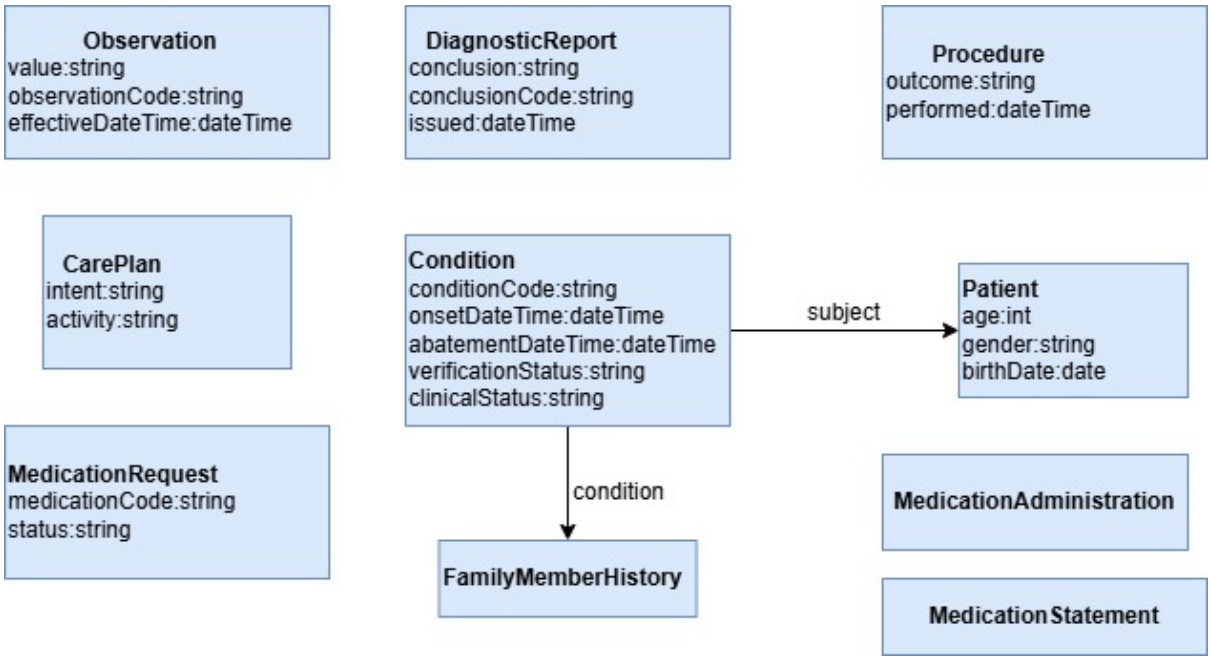

Figure 1: Overview of the FHIR-light ontology

3.1.1. MultiClinSum-2 Data Analysis

We have examined the sample dataset and found that both in English and Dutch, there are on average slightly more than 5 sentences and around 96 words in the summaries. The compression ratio from the full text to the summaries is about 0.2. Each summary contains 1) patient information, at least age and gender, sometimes race, origin, impression and history, 2) sometimes an initial symptom, 3) always a finding such as clinical measurement (e.g blood pressure), diagnostic procedure (e.g. MRI) or diagnosis, 4) sometimes as treatment or outcome such as disease progression, treatment, affected systems and functional impact.

3.2. KG Construction: FHIR-light

To provide semantic grounding for the extracted knowledge graphs, we incorporated a lightweight ontology based on the FHIR standard. FHIR was selected after the analysis in Section 3.1.1, as it provides a comprehensive representation of clinical information, including patient characteristics, diagnoses, observations, treatments, and care plans. These categories closely matched the information types that frequently occurred in the summaries and source documents.

Because the complete FHIR ontology contains a large number of resources and relations that were unnecessary for our summarization task, we constructed a reduced subset tailored specifically to the extraction pipeline (see Figure 1). The subset ontology contains ten core classes: Patient, FamilyMemberHistory, Condition, Observation, MedicationRequest, MedicationAdministration, MedicationStatement, Procedure, CarePlan, and DiagnosticReport. These classes capture the central entities required for representing patient state, diagnoses, measurements, medication usage, procedures, and treatment planning.

In addition to the classes, a selection of data and object properties was included to model relations and attributes relevant to the extraction task. Patient-related properties include demographic information such as gender, birth date, and age. Clinical conditions are represented through properties describing clinical status, verification status, onset and abatement dates, and diagnostic codes. Observations contain measurement codes, values, and timestamps, while medication-related entities describe prescribed or administered medications and their statuses. Procedures include information about execution time and outcome, and diagnostic reports contain conclusions, conclusion codes, and issue dates. Finally, the

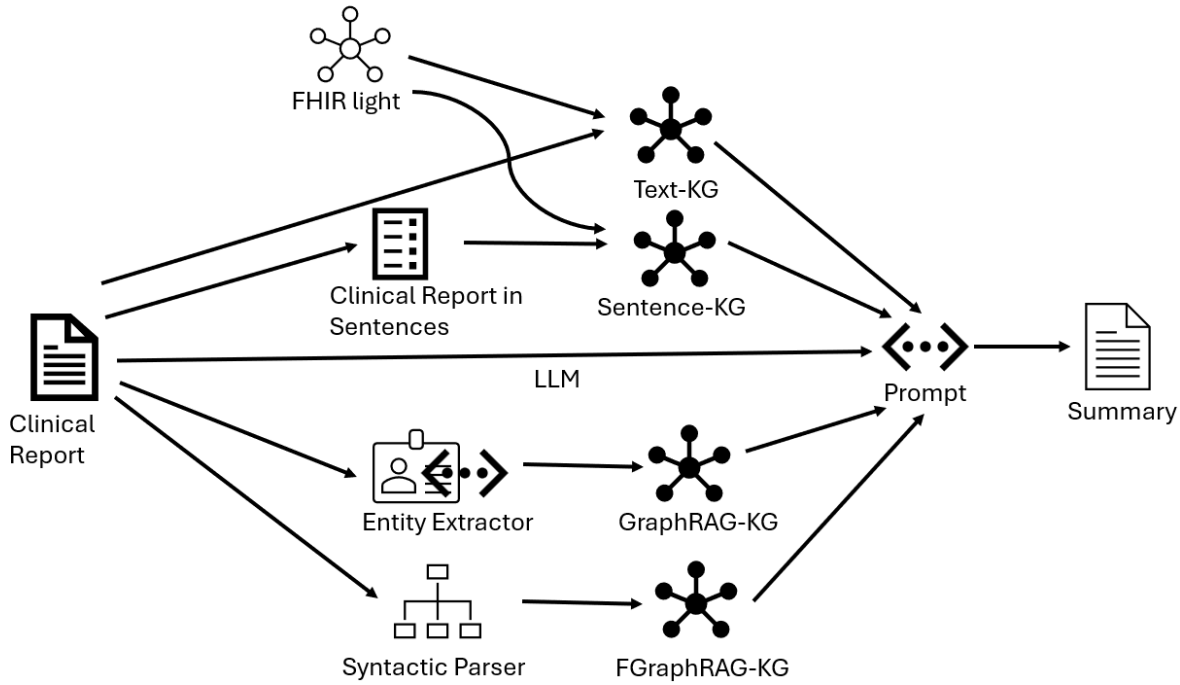

Figure 2: Overview of our methods

care plans include information on intended treatments and planned activities.

The ontology was implemented in OWL/Turtle format using standard RDF, RDFS, and OWL constructs. Domain and range constraints were defined for all properties to ensure semantic consistency during extraction and graph generation. By constraining the extraction process to this reduced ontology, the generated knowledge graphs remained focused, interpretable, and suitable for downstream summarization tasks while avoiding unnecessary complexity from the full FHIR specification.

3.3. Methods

To automatically create summaries from clinical case reports, we compare 5 different methods as visualized in Figure 2: **LLM-only** As a baseline, we provide the text and a prompt to the LLM with the instruction to extract a summary. Then, to give the LLM a more factual foundation, we tested several methods for going from text, to a knowledge graph (KG), and then to a summary based on that knowledge graph. These methods include:

- **Text-KG** A text is first converted to triples, then to a KG and then together with a prompt provided to an LLM
- **Sentence-KG** A text is first split in sentences and then follows the same steps as Text-KG
- **GraphRAG-KG** uses Microsoft GraphRAG to create a KG
- **FGraphRAG-KG** uses Microsoft fastGraphRAG to create a KG

We use a modular setup which is a priori model-agnostic. This means that a method is easily added and an LLM model is easily changed (although new prompts might be required). In all runs submitted to the challenge, the OpenAI GPT-5.1 model is used, as this is one of the newer models and it is focused on adaptive reasoning, consistency and tone [17, 18]. The Azure platform is used to perform our experiments, with an API version of 2024-12-01-preview. If there was a connection error or no response, 3 attempts were done. Otherwise an empty summary was provided.

Our prompt is based on the data analysis; a shortened version is shown below.

The full prompts for English can be found in Appendix A - C. To create our prompts, we have followed prompt conventions for this GPT version[4], such as defining the role of the LLM, the task, the input information, clear short instructions and output rules. We have explored different namings of the system, the mentioning of the patient's medical record and clinical case report, the constraints (including number of words instead of sentences and without constraints), and the wording in the order of the summary. We have calculated the Rouge, Bleu, and BERTScore (using the Evaluate package on Huggingface[5]) on the sample dataset for several prompt versions, and this was the most optimal prompt based on our exploration.

```
You are a medical summarization system.

Your task is to create a short summary of a patient medical record.

You are given:
- A clinical case report

INSTRUCTIONS:
- Read the whole text
- Select the most relevant information to summarize
- Constraints: maximum 5 sentences. Focus only on key takeaways.
- The summary should have the following order: 1. patient information, 2. symptoms,
3. clinical findings, 4. treatment, 5. outcome

OUTPUT RULES:
- Return ONLY the summary
```

3.3.1. Text-KG

In the Text-KG method the text is first converted to triples and then to a knowledge graph, based on the FHIR-light ontology as explained in section 3.2. Triples are extracted using the prompt as displayed in Appendix A. The triples are converted using the rdflib package [19]. The knowledge graph is then, together with the prompt from Appendix B, used to create the summary.

3.3.2. Sentence-KG

The sentence-KG method is similar to Text-KG, but instead of extracting triples in one go, we first use the sentence tokenizer from NLTK [20] to split the text into sentences. Then the same prompts are used, but now per sentence. All triples are also converted to a graph in the same manner as with Text-KG.

3.3.3. LLM-only

This is the baseline method in which we use the text as input and the prompt from Appendix C to generate the summary. We used 2-shot prompting, so we added 2 examples to the end of the prompt. For the Dutch run, we used the Dutch texts and summaries as examples.

3.3.4. GraphRAG-KG

For the GraphRAG-KG method, we use the graph extracted based on the Microsoft GraphRAG implementation [21]. This implementation first splits the documents into text chunks, then extracts entities and relations within a chunk. An LLM generates short descriptions for the entities and relations, and in the final step, a knowledge graph is created in which entities and relations are aggregated to nodes and

[4]https://developers.openai.com/cookbook/examples/gpt-5/gpt-5-1_prompting_guide
[5]https://github.com/huggingface/evaluate/tree/main

Method	ROUGE1	ROUGE2	rLsum	BERTScore	f	compl	fluency	cons
Text-KG	0.349	0.129	0.229	0.864	0.766	0.919	0.700	0.690
Sentence-KG	0.321	0.115	0.217	0.859	0.719	0.776	0.698	0.661
LLM-only	**0.373**	**0.142**	**0.249**	**0.866**	0.782	**0.964**	**0.702**	0.697
GraphRAG-KG	0.300	0.107	0.199	0.862	**0.804**	0.943	0.701	**0.706**
FGraphRAG-KG	0.221	0.062	0.159	0.846	0.733	0.710	0.700	0.702

Table 1
Test Results for English; rLsum is ROUGELsum, f is faithfulness, compl is completeness and cons is consistency

Method	ROUGE1	ROUGE2	rLsum	BERTScore	f	compl	fluency	cons
Text-KG	0.135	0.037	0.093	0.845	0.736	0.939	0.701	0.666
Sentence-KG	0.130	0.034	0.093	0.841	0.695	0.740	0.697	0.589
LLM-only	**0.336**	**0.109**	**0.223**	**0.862**	0.747	**0.983**	**0.708**	0.669
GraphRAG-KG	0.227	0.073	0.164	0.858	**0.755**	0.938	0.700	**0.699**
FGraphRAG-KG	0.233	0.055	0.164	0.843	0.700	0.680	0.699	0.667

Table 2
Test Results for Dutch; rLsum is ROUGELsum, f is faithfulness, compl is completeness and cons is consistency

edges. These element summaries are then used to create graph communities and community summaries. These communities are an aggregation of multiple elements (documents), after which the answers can be created from the communities instead of the separate documents. As our method aims to create a KG, and we need the KG per document, the steps to create graph communities and community summaries are not used in our method.

In our implementation, we create an extended version of the entities, which are organization, person, geo, event, patient, condition, observation, procedure, medication, and outcome. These entities generally follow the FHIR-light. We use the default setting of the implementation.

3.3.5. FGraphRAG-KG

For the FGraphRAG-KG, we use the Microsoft fastGraphRAG implementation, in which NLP tools are used [22] instead of an LLM to extract entities. We use the syntactic parser option to extract entities, with the default model for English and the "nl_core_news_lg" model for Dutch.

4. Results

The results on the test set are shown in Table 1 and 2. The results show that the LLM-only model performs best on the ROUGE and BERTScore metrics, as well as completeness and fluency, both for Dutch and English. The GraphRAG-KG method is best in terms of faithfulness and consistency. If we further compare performance of the other methods (not including LLM), Text-KG has better ROUGE and BERTScores, followed by GraphRAG-KG, Sentence-KG and FGraphRAG-KG is worst. The LLM-as-a-judge metrics follow a slightly different trend: GraphRAG-KG, Text-KG / FGraphRAG-KG, Sentence-KG.

Overall, differences in performance between Dutch and English are small. We observe a drop in performance, and especially completeness, when comparing Text-KG and Sentence-KG and when comparing GraphRAG and fastGraphRAG-KG.

A statistical analysis on the test set is shown in Table 3, with summaries and knowledge graphs for each method included for one test sample in Appendices D and E. The FGraphRAG-KG method is the only approach that generated summaries for all test instances in both English and Dutch, while the LLM baseline showed the highest number of missing summaries, particularly for English. The Sentence-KG method produces the shortest summaries on average, whereas the GraphRAG-KG and FGraphRAG-KG methods generate substantially longer summaries. In terms of graph size, Sentence-KG produces the

Method	Missing Summaries	Average Words	Average KG Size (triples)
Text-KG	1 | 1	117.06 | 108.14	75.85 | 75.27
Sentence-KG	1 | 0	95.81 | 85.00	64.19 | 51.24
LLM	4 | 1	119.03 | 111.08	- | -
GraphRAG-KG	2 | 0	303.13 | 437.03	93.84 | 87.92
FGraphRAG-KG	0 | 0	339.74 | 206.90	465.17 | 246.55

Table 3
Statistical analysis on the test set per method; English values are shown on the left and Dutch values on the right.

smallest graphs, while FGraphRAG-KG produces the largest graphs, especially for English. Overall, the English graphs are slightly larger than the Dutch graphs for most methods.

5. Discussion

The results of the experiments show that only using the LLM has higher performance than using a knowledge graph, at least for all metrics except for faithfulness and consistency. This could be due to the few shot prompting used in the LLMs and/or because of the accuracy and completeness of the KG methods.

An example of resulting KGs are shown in Appendix E. For Sentence-KG and Text-KG we included the FHIR-light ontology as a structure for the LLM to create the KG. However, the LLM did not follow the ontology structure. In the Sentence-KG the LLM hallucinated more than in the Text-KG. The LLM hallucinated new concepts which are not in the ontology, such as the concept "Activity_limitation_scale" and the predicate "status" in "Procedure". There was inconsistency with capitalization which created two separate objects "Observation_1" and "observation_1". It also did not follow the typesetting, age is defined as having an integer value but was written as a string. In Text-KG the connections between objects Patient and Condition was not followed according to the ontology.

The KGs for GraphRAG-KG and FGraphRAG-KG show that there is no consistent format for these generated KGs. These graphs were much larger and therefore we show a snippet in Appendix E. Both graphs shown do not have directional relations, nor defined edges. The graph by GraphRAG-KG has specific types such as CONDITION and OBSERVATION whereas the other graph has a general NOUN PHRASE type for all objects. Moreover the KG by GraphRAG-KG has very long texts with information, whereas the FGraphRAG-KG graph has none. A substantial part of the information from the clinical note is repeated over the nodes in the KG by GraphRAG-KG. The concepts and relations between the nodes in the KG by FGraphRAG-KG are always logical to the human reader, for instance "49".

Since we only performed one run on the test set, there no information about robustness or replicability. Considering the variety in the produced KGs, especially in the GraphRAG-KG and FGraphRAG-KG methods, we expect that rerunning the same prompts could produce significantly different KGs with potentially different summaries.

In general, if we combine this with the results of Table 3, we see that the summaries of Sentence-KG are shorter, which could have influenced the completeness score. The GraphRAG-KG method has very long Dutch summaries, but this seems not clearly visible in the evaluation metric scores.

Another observation is that, although we did not know in advance, our prompt contained the key clinical information that was measured in the completeness metric, which is named as patient demographics, diagnosis, intervention, outcome and follow-up. As our naming is slightly different (patient information, symptoms, clinical findings, treatment, outcome), the other naming could have improved performance a bit for all methods.

Additionally, running LLMs comes at a cost. For most KG-based methods we had to perform two LLM calls per summary: one for the KG creation and one for the summarization. This was repeated for both Dutch and English. The costs are higher for the output tokens than for the input tokens. This means that especially KG creation from the full texts was expensive, as some of the resulting KGs are

up to 13,162 tokens or 41185 characters (42KB). By working in batches the costs were reduced. This means a large set of calls was sent to the LLM in one batch, and the LLM would respond to all calls within 24 hours, when there was space on the server, for example at night. The syntactic parser or other non-LLM-based methods to create KGs will help circumvent this problem. We recognize the environmental impact that these amounts of LLM calls have.

A major limitation of our work is that the LLM uses few-shot prompting (adding 2 examples), and the KG-based methods did not. In our runs on the sample set, the examples did not have a much higher performance, whereas it did add to the costs. This choice could, however, have influenced the results on the test set, as the summaries that are created with few-shot might be closer to the ground truth summaries, for example in style. For the KG-based methods, we view that the methods all produce quite different knowledge graphs. As all methods use an LLM to generate a graph, this has caused errors to appear in the KG generation. Another limitation is that we have not used the train set for these runs or used a specific medically trained LLM. Additionally, summarization is a very hard task as it is much dependent on the ground truth summaries and the chosen metrics.

6. Conclusion

In this work, we describe our modular setup to explore different ways of creating knowledge graphs to support the automatic generation of summaries. The results on the Dutch and English cases of the MultiClinSum-2 task by BioASQ show that using few-shot prompting with the OpenAI gpt-5.1 model is more powerful than zero-shot with a knowledge graph, at least for the summaries evaluated in this task. In future work, we plan to further evaluate and optimize the created knowledge graphs and integrate our setup in a pipeline to automatically summarize conversations instead of clinical reports.

7. Acknowledgments

This paper was created with funding from the ITEA project 22032 MediSpeech. We would like to thank prof. Stephan Raaijmakers for his review on the paper, Lucia Tealdi for leading the project within TNO and Gabriel Hoogerwerf for his contribution to an early version of our code pipeline.

Declaration on Generative AI

The authors have not employed any Generative AI tools for writing this paper.

References

[1] N. Clynch, J. Kellett, Medical documentation: part of the solution, or part of the problem? A narrative review of the literature on the time spent on and value of medical documentation, International journal of medical informatics 84 (2015) 221–228.

[2] E. Asgari, J. Kaur, G. Nuredini, J. Balloch, A. M. Taylor, N. Sebire, R. Robinson, C. Peters, S. Sridharan, D. Pimenta, Impact of electronic health record use on cognitive load and burnout among clinicians: narrative review, JMIR Medical Informatics 12 (2024) e55499.

[3] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, A. Mash, M. Krallinger, Overview of MultiClinSum-2 at CLEF 2026: Data, Evaluation, and Results for Multilingual Clinical Case Report Summarization Across 15 Languages, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2025 Working Notes, 2026.

[4] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli,

G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[5] H. Shakil, A. Farooq, J. Kalita, Abstractive text summarization: State of the art, challenges, and improvements, Neurocomputing 603 (2024) 128255.

[6] C. Napoles, M. R. Gormley, B. Van Durme, Annotated gigaword, in: Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX), 2012, pp. 95–100.

[7] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1073–1083. URL: https://www.aclweb.org/anthology/P17-1099. doi:10.18653/v1/P17-1099.

[8] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Giménez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of multiclinsum task at bioasq 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results, in: CLEF, 2025.

[9] S. Rhazzafe, S. Colreavy-Donnelly, N. S. Nikolov, Extrasum @ multiclinsum: Extractive summarization of english, spanish, french and portuguese clinical case reports, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, 2025.

[10] M. S. Salim, L. Fu, A. A. Ramakrishnan, S. Kwon, Z. Yao, H. Yu, Enhancing multilingual medical summarization via contextual keyword augmentation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, 2025.

[11] J. Angulo, V. Yeste, Agentic mcs: A multilingual clinical summarization framework, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, 2025.

[12] M. Ivanović, Z. Budimac, An overview of ontologies and data resources in medical domains, Expert Systems with Applications 41 (2014) 5158–5166.

[13] R. Cornet, N. de Keizer, Forty years of snomed: a literature review, BMC medical informatics and decision making 8 (2008) S2.

[14] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[15] C. E. Lipscomb, Medical subject headings (mesh), Bulletin of the Medical Library Association 88 (2000) 265.

[16] C. N. Vorisek, M. Lehne, S. A. I. Klopfenstein, P. J. Mayer, A. Bartschke, T. Haese, S. Thun, Fast healthcare interoperability resources (fhir) for interoperability in health research: systematic review, JMIR medical informatics 10 (2022) e35724.

[17] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, et al., Openai gpt-5 system card, arXiv preprint arXiv:2601.03267 (2025).

[18] OpenAI, Introducing gpt-5.1 for developers, https://openai.com/index/gpt-5-1/, 2025. Accessed: 2026-05-27.

[19] D. Krech, G. AAstrand Grimnes, G. Higgins, J. Hees, I. Aucamp, N. Lindström, N. Arndt, A. Sommer, E. Chuc, I. Herman, et al., Rdflib, Zenodo (2002).

[20] E. Loper, S. Bird, Nltk: The natural language toolkit, in: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, 2002, pp. 63–70.

[21] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024).

[22] Microsoft, Extract graph nlp, https://microsoft.github.io/graphrag/config/yaml/#extract_graph_nlp, Accessed: 2026-05-27. Part of the GraphRAG project.

Appendices

A. Text-to-KG Prompt

You are an ontology developer who extracts RDF triples from text. Your task is to extract RDF triples from text using a given ontology.

You are given:

- ontology
- input_text

Instructions:

- Use only concepts from the ontology
- Create instances when needed and assign `rdf:type`
- Do not infer information not present in the text
- Use prefixed names (e.g., `fhir:patient_1`, `fhir:condition_1`)
- Use one triple per individual (e.g., condition_1, condition_2 for each condition)
- No explanations
- If no triples are found, return `[]`

Content guidelines: If present in the text, always include:

- Patient information: age, gender, race, family history
- Observations or conditions
- Treatment information
- Outcome information

Output: Return ONLY valid JSON in the form:

```
[
  {"subject": "...", "predicate": "...", "object": "..."}
]
```

B. KG-to-Summary Prompt

You are a medical summarization system.

Your task is to create a short summary of a knowledge graph of a patient medical record.

You are given:

- A knowledge graph of a clinical case report

Instructions:

- Read the entire knowledge graph
- Select the most relevant information
- Maximum 5 sentences
- Follow this order: patient information → symptoms → clinical findings → treatment → outcome

Output: Return ONLY the summary

C. Text-to-Summary Prompt

You are a medical summarization system.

Your task is to create a short summary of a patient medical record.

You are given:

- A clinical case report

Instructions:

- Read the entire text
- Select the most relevant information
- Maximum 5 sentences
- Follow this order: patient information → symptoms → clinical findings → treatment → outcome

Output: Return ONLY the summary

Example 1

Input: A 63-year-old man had been subjected to a percutaneous coronary intervention via his right femoral artery. At the end of the uncomplicated procedure, the puncture site was closed using an Angio-Seal™. The patient had no post-interventional complaints and no signs of ischemia. Doppler ultrasound showed a mobile structure in the femoral artery. The device was surgically removed. The patient was discharged without complications.

Output: A patient developed a femoral artery murmur after PCI caused by a dislocated Angio-Seal™ device. The condition was diagnosed via Doppler ultrasound and treated surgically, with full recovery.

Example 2

Input: A 30-year-old woman with an 8-year history of infertility and vaginal bleeding. She previously underwent laparoscopic surgery for endometriosis and multiple fertility treatments. Imaging revealed abnormal vascular structures consistent with UAVM. Despite repeated embolization, symptoms persisted, but she later achieved successful pregnancies via IVF.

Output: A 30-year-old woman with long-term infertility and persistent vaginal bleeding. Imaging suggested uterine arteriovenous malformation (UAVM). Despite repeated embolization, the condition persisted, but she later achieved successful pregnancies through IVF.

D. Example of a Summary per method (ex. 7 from Test Set; English)

Original Text: A 25-year-old female presented to our patient clinic with complaints of pain and swelling over the dorsum of her right foot for a period of 1 year. Swelling was insidious in onset, gradually progressive and reached the present size. Pain was insidious in onset, intermittent, moderate intensity, dull aching, no radiation, aggravated by walking, and relieved with medications. There is no history of fever, trauma, loss of weight, and loss of appetite. The patient had no other medical comorbidities. On examination, there was a localized ovoid-shaped swelling of 2 by 2 cm over the dorsum of right foot, 7–8 cm in front of medial malleolus, and 5 cm behind the base of 2nd toe. Surface appeared to be smooth. Edge of the swelling was clearly defined. There was no hyperpigmentation of skin. On palpation, there was no warmth and tenderness was present. Swelling had well-defined margins, non-mobile, and hard in consistency. Skin over the swelling was pinchable. There was no enlargement

of any regional lymph node. X-ray of right foot AP and oblique () was done which revealed well-defined osteolytic lesion in the center of the intermediate cuneiform, geographic pattern of destruction, narrow zone of transition, and not breaching the cortex. MRI of right foot („) was done which showed an expansile osteolytic lesion with multiple internal septations in intermediate cuneiform with thinning of the cortex. Patient was planned for surgery, the incision was made over the dorsum of the foot, the bone was approached between extensor hallucis longus and extensor digitorum brevis. The intermediate cuneiform was identified and excised completely and was sent for histopathological examination. There was no extension of the lesion to the surrounding soft tissues. Postop xray () showed the removal of intermediate cuneiform. Microscopic examination showed the presence of focal giant cell-rich lesions with background stromal cells, and areas of hemorrhage were also noted (and). The finding in the microscopic picture helped us in narrowing our diagnosis to GCT evolving into secondary ABC. The patient was made to partially weight bear for 1 month and following that full weight bearing was started. The patient was not on any sort of immobilizing splint to the ankle or foot. The patient was followed periodically and by 1-year patient was able to weight bear without pain and there was no recurrence of lesion in the foot. Foot Function Index [] preoperatively was 49% (84/170) and got improved to 5% (9/170) after 1 year postoperatively. Activity limitation scale improved from 13% to 3%.

D.1. Text-KG (120 words)

A 25-year-old female patient was diagnosed with an active giant cell tumor evolving into a secondary aneurysmal bone cyst of the intermediate cuneiform, with onset about 1 year before presentation. Symptoms are implied by impaired foot function and activity, with a preoperative Foot Function Index of 49% and activity limitation scale of 13%. Clinical findings centered on the intermediate cuneiform lesion. Treatment consisted of surgical excision of the intermediate cuneiform lesion and a care plan including partial weight bearing for 1 month followed by full weight bearing. At 1-year follow-up, the Foot Function Index improved to 5% and the activity limitation scale to 3%, and the patient was able to bear weight without pain with no recurrence of the lesion.

D.2. Sentence-KG (93 words)

A 25-year-old female patient presented with pain and swelling over the dorsum of her right foot for one year. X-ray imaging of the right foot showed a well-defined osteolytic lesion in the center of the intermediate cuneiform with a geographic pattern of destruction, narrow zone of transition, and intact cortex. A planned surgical procedure involved an incision over the dorsum of the foot with bone approached between the extensor hallucis longus and extensor digitorum brevis. Her activity limitation score improved from 13% pre-treatment to 3% post-treatment. Overall clinical outcome was documented as improved.

D.3. LLM (118 words)

A 25-year-old woman presented with a 1-year history of pain and gradually progressive swelling over the dorsum of her right foot. Symptoms included intermittent, moderate, dull aching pain aggravated by walking, without systemic features or history of trauma. Examination and imaging (X-ray and MRI) revealed a well-defined, expansile osteolytic lesion in the intermediate cuneiform with cortical thinning but no soft tissue extension. She underwent complete excision of the intermediate cuneiform, and histopathology confirmed a giant cell tumor (GCT) evolving into secondary aneurysmal bone cyst (ABC). Postoperatively, she progressed from partial to full weight bearing, and at 1 year had no recurrence with marked functional improvement (Foot Function Index from 49% to 5%, activity limitation from 13% to 3%).

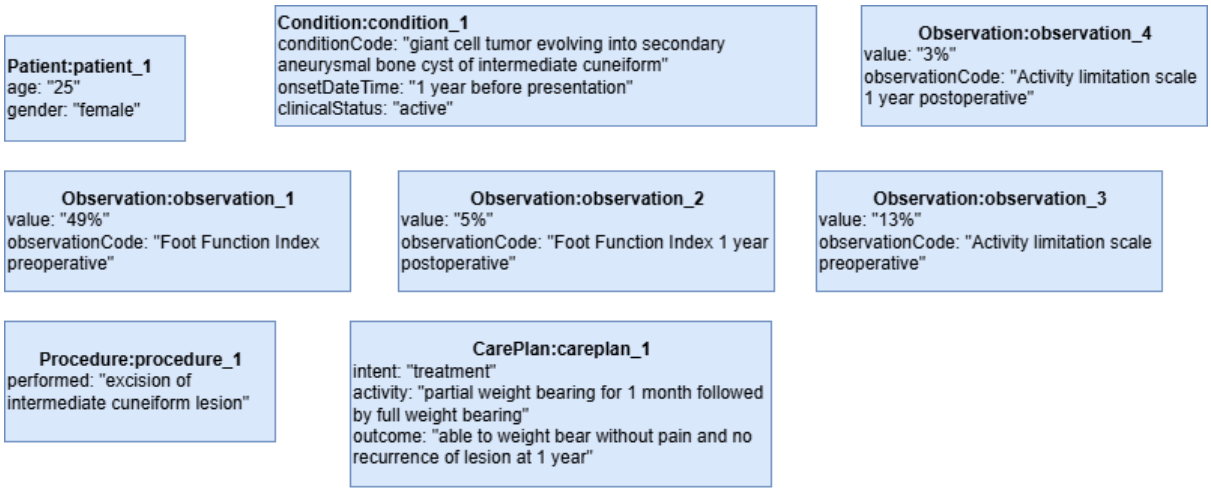

Figure 3: Schematic representation of the KG of clinical note 7 by Text-KG

D.4. GraphRAG-KG (163 words)

A 25-year-old woman with no medical comorbidities presented with a 1-year history of insidious, intermittent, moderate dull aching pain and gradually progressive swelling over the dorsum of the right foot, without history of trauma or systemic symptoms. Examination showed a tender, hard, non-mobile, well-defined 2×2 cm swelling with pinchable, normal-overlying skin and no regional lymphadenopathy. Imaging (X-ray and MRI) revealed a well-defined expansile osteolytic lesion confined to the intermediate cuneiform with a geographic pattern of destruction, narrow zone of transition, cortical thinning, and no soft tissue extension; histopathology of the excised bone confirmed a giant cell tumor evolving into a secondary aneurysmal bone cyst. She underwent complete surgical excision of the intermediate cuneiform, followed by partial weight bearing for 1 month and then full weight bearing, without use of an immobilizing splint. At 1-year follow-up there was no recurrence, she could bear weight without pain, and her Foot Function Index improved from 49% to 5%, with activity limitation decreasing from 13% to 3%.

D.5. FGraphRAG-KG (77 words)

A 25-year-old woman presented with a 1-year history of dull, aching pain in the right foot. Examination and imaging revealed an expansile osteolytic lesion measuring approximately 2–5 cm with well-defined margins, involving the region of the extensor hallucis longus and surrounding soft tissues, but without regional lymph node involvement or other medical comorbidities. After about 1 month she was allowed full weight-bearing. Microscopic examination confirmed the nature of the lesion (details not specified in the graph).

E. Example of a Summary per method (ex. 7 from Test Set; English)

E.1. Text-KG

See figure 3.

E.2. Sentence-KG

See figure 4.

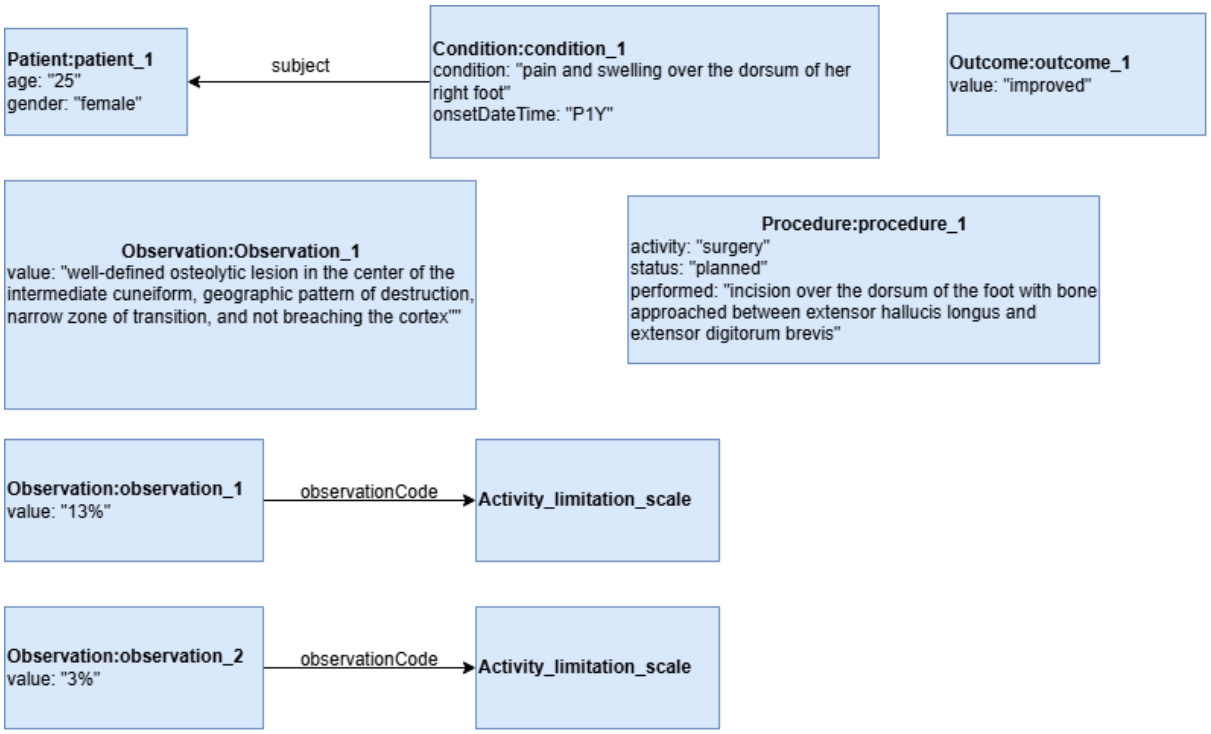

Figure 4: Schematic representation of the KG of clinical note 7 by Sentence-KG

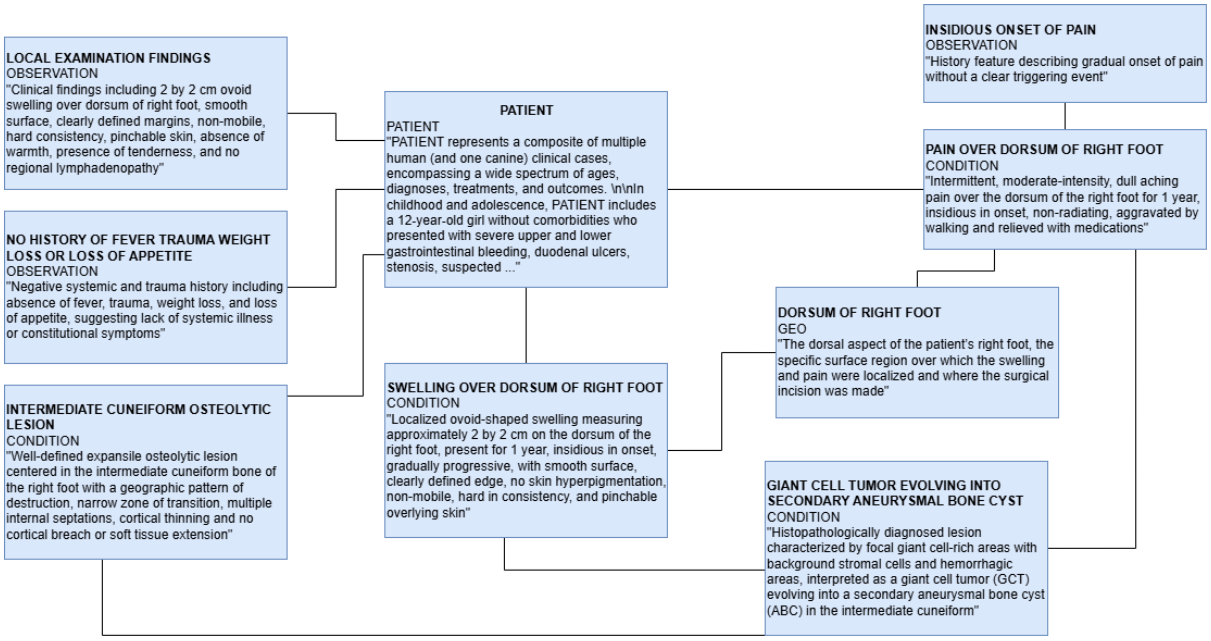

Figure 5: Schematic representation of the KG of clinical note 7 by GraphRAG-KG

E.3. GraphRAG-KG

See figure 5.

E.4. FGraphRAG-KG

See figure 6.

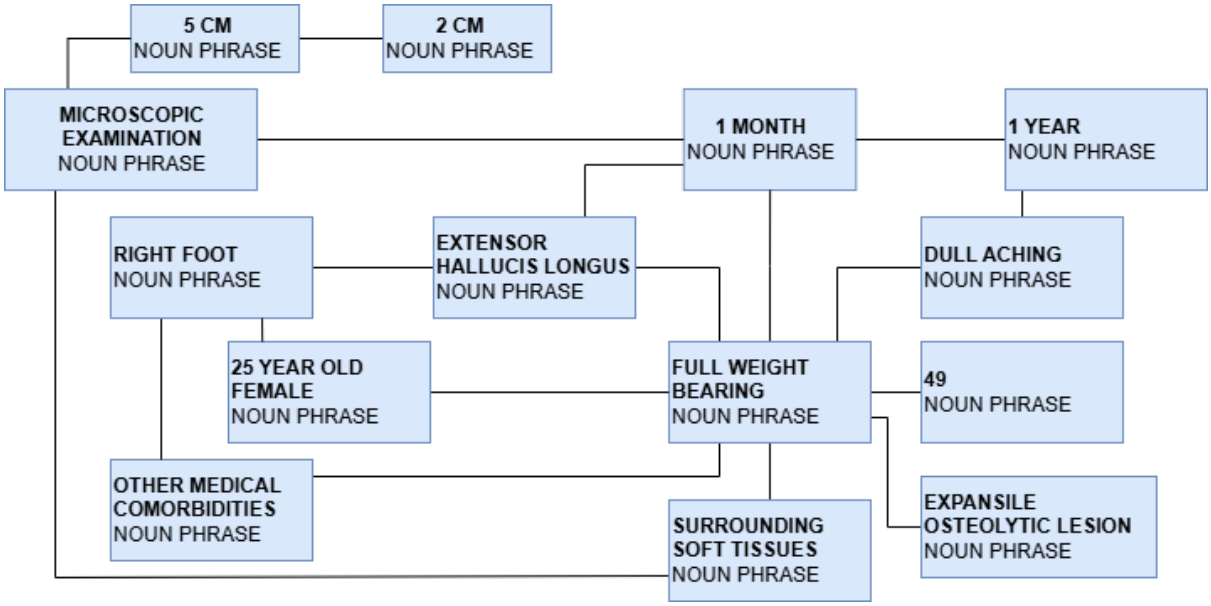

Figure 6: Schematic representation of the KG of clinical note 7 by FGraphRAG-KG