MDPI

Article

Explainable AI–Based Semantic Retrieval from an Expert-Curated Oncology Knowledge Graph for Clinical Decision Support

Sameer Mushtaq [1], Marcello Trovati [2] and Nik Bessis [1]

[1] Department of Computer Science, Edge Hill University, UK; Sameer.Mushtaq@edgehill.ac.uk, Nik.Bessis@edgehill.ac.uk
[2] University of Lancashire Business School, University of Lancashire, UK; MTrovati@lancashire.ac.uk
* Correspondence: MTrovati@lancashire.ac.uk

Abstract: The modern oncology landscape is characterised by a deluge of high-dimensional data from genomic sequencing, medical imaging, and electronic health records, negatively impacting the analytical capacity of clinicians and health practitioners. This field is not new and it has drawn significant attention from the research community. However, one of the main limiting issues is the data itself. Despite the vast amount of available data, most of it lacks scalability, quality and semantic information. This work is motivated by the data platform provided by OncoProAI, an AI-driven clinical decision support platform designed to address this challenge by enabling highly personalised, precision cancer care. The platform is built on a comprehensive knowledge graph, formally modelled as a directed acyclic graph, which has been manually populated, assessed and maintained to provide a unique data ecosystem. This enables targeted and bespoke information extraction and assessment.

Keywords: Artificial Intelligence; Machine Learning; Deep Learning; Digital Health;

Received:
Revised:
Accepted:
Published:

Citation: Lastname, F.; Lastname, F.; Lastname, F. Explainable AI–Based Semantic Retrieval from an Expert-Curated Oncology Knowledge Graph for Clinical Decision Support. *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

1. Introduction

The field of modern oncology is defined by the increasing creation and use of complex and high-dimensional data. From the granular details of genomic and proteomic sequencing to the complex patterns in radiological images and the vast unstructured text of electronic health records, the volume of information available for each patient challenges current state-of-the-art technology [3]. Artificial Intelligence (AI), underpinned by mathematical modelling and machine learning algorithms, has emerged as a transformative force capable of addressing this challenge. By detecting subtle patterns and correlations invisible to the human eye, AI offers the potential to revolutionise cancer care, driving a paradigm shift from generalised protocols to highly personalised precision medicine [1].

The application of AI in oncology spans the entire patient's journey. It includes techniques such as deep learning to enhance the accuracy of diagnostic imaging and digital pathology, and employs predictive models to forecast disease progression and patient survival with greater accuracy. Furthermore, with the power of Natural Language Processing (NLP), AI can unlock critical insights from clinical notes, facilitating research and cohort identification. The ultimate goal is to create a close relationship between the expertise of the clinician and the analytical power of AI, leading to more timely diagnoses, optimised treatment strategies, and fundamentally better outcomes for patients [2].

1.1. Limitations of Current Research

Image-based approaches, often using deep learning models such as Convolutional Neural Networks (CNN), are particularly suitable for the identification of data patterns in radiological scans (CT, MRI, mammograms) and histopathology slides that may not be visible to the human eye, helping to detect and classify early cancers [4,5]. Currently, NLP techniques are used to extract vital information from unstructured clinical notes, pathology reports, and the scientific literature, providing information on the patient's history, symptoms, diagnoses, and treatment regimens [6]. The synergy between these approaches to image and text data, often facilitated by Vision-Language Models (VLMs), holds significant promise for a more comprehensive understanding of complex oncological cases [7]. Although automated methods for data extraction and analysis are advancing rapidly, a critical factor for achieving superior accuracy and interpretability in oncology AI lies in the development of manually populated, curated semantic, and ontology-based data. Automated approaches, despite their efficiency, often struggle with the inherent ambiguities, inconsistencies, and vast heterogeneity of clinical data. For instance, variations in terminology, abbreviations, and sentence structures within free-text clinical notes can lead to misinterpretations or missed information [8]. Similarly, discrepancies in image acquisition protocols or reporting standards can introduce noise into visual data. In many cases, such issues also stem from the lack of customised data with limited semantic information. Furthermore, traditional retrieval-augmented generation approaches typically focus on individual documents in isolation, failing to leverage the rich network relationships that exist between clinical concepts, treatment pathways, and patient outcomes [27]. In fact, a suitably curated semantic (and ontological) framework provides a machine-readable standardised representation of medical knowledge, explicitly defining concepts, their attributes, and relationships within the oncology domain. This rigorous and expert-driven curation process, although resource intensive, ensures high data quality, reduces bias, and improves the reliability of AI models [9]. By establishing a common vocabulary and a robust knowledge graph, such an approach would facilitate intelligent data integration across disparate sources, enabling more accurate reasoning and inference by AI algorithms. For example, an ontology could precisely link specific histological features observed in an image with the corresponding genetic mutations described in a text report, providing a deeper and more contextualised understanding of a patient's cancer [10]. This human-in-the-loop validation of medical knowledge encoded in ontologies is paramount, as even subtle inaccuracies can have significant clinical consequences, ultimately leading to more trustworthy and effective AI-driven oncology solutions [11].

This article is motivated by the above observations and aims to demonstrate how such data can significantly enhance AI applications to oncology, as well as to wider areas related to health. However, such rich and interconnected data is usually a complex, time consuming, and labour intensive endeavour. This work was carried out in collaboration with OncoProAI [12]. OncoProAI is a company focussing on data-driven cancer treatment, based on semantically and ontologically defined data. The analysis and approach in this work is based on this data demonstrating its potential and accuracy in this field. The authors were granted access to their proprietary data and knowledge graphs, which were employed to train and evaluate the models in this work. Such data has led to highly effective and accurate model implementations, as demonstrated in the evaluation.

The article is structured as follows: Section 2 provides a comprehensive background on AI in oncology and ontological approaches. Section 3 introduces the OncoProAI platform, the description of the data set, and the ethical considerations. Section 4 describes

the methodology including node serialisation, embedding models, explainability, and annotation workflow. Section 5 presents the evaluation framework and experimental setup. Section 6 reports the retrieval performance, statistical significance, and efficiency analysis. Finally, Section 7 concludes the article and outlines future directions.

2. Background

The field of oncology is characterised by a rapidly expanding volume of complex, high-dimensional data, ranging from genomic and proteomic profiles to clinical imaging and electronic health records. The sheer scale and complexity of this information presents significant challenges for human clinicians in diagnosis, prognosis, and treatment planning. AI, machine learning (ML), deep learning (DL), and Natural Language Processing (NLP) have emerged as a powerful paradigm to address these challenges. AI offers the potential to extract meaningful patterns from data, automate complex tasks, and ultimately enable a more precise and personalised approach to cancer care [1,14].

One of the most mature applications of AI in oncology is in the analysis of medical imaging. Radiomics, the process of extracting large amounts of quantitative features from medical images, is highly utilising ML algorithms [15]. Deep learning models, especially CNNs, have demonstrated performance comparable to, and in some cases exceeding, that of human radiologists in identifying malignant lesions in mammograms, CT scans, and MRIs. For instance, studies have shown high sensitivity and specificity in detecting lung nodules and classifying breast cancer subtypes from imaging data alone. These models learn hierarchical features directly from pixels, bypassing the need for manual feature engineering [16].

AI is similarly revolutionising histopathology. By analysing whole slide images (WSIs), ML algorithms can assist pathologists in identifying tumour regions, grading cancers (for example, Gleason score in prostate cancer) and counting mitotic figures. This automation reduces inter-observer variability and increases throughput, allowing pathologists to focus on more complex cases.

2.1. Prognosis and Predictive Modelling

Predicting a patient's clinical outcome is fundamental to a personalised treatment. AI models excel at integrating multi-modal data to generate robust prognostic and predictive insights [17].

Traditional statistical methods, such as the Cox proportional hazards model, are being augmented and, in some cases, surpassed by ML-based survival models. Techniques such as Random Survival Forests and DL-based survival models can handle complex, non-linear interactions between variables. These models can predict metrics such as overall survival or progression-free survival by integrating clinical data with genomic markers and radiomic features.

A critical goal of precision oncology is to predict which patients will respond to a specific therapy. AI models are being trained on pre-treatment data (e.g., genomics, transcriptomics, tumour microenvironment characteristics) to predict response to chemotherapy, immunotherapy, and targeted agents. For example, DL models have been used to predict the response to immune checkpoint inhibitors by analysing patterns in tumour histology and the expression of biomarkers such as PD-L1 [18].

2.2. Treatment Planning and Decision Support

The complexity of modern cancer treatment protocols, with numerous drug combinations and sequencing options, makes clinical decision-making increasingly difficult.

AI-driven decision support systems (DSS) are being developed to assist oncologists. These systems synthesise patient-specific data, biomarker status, and the latest clinical guidelines (e.g., NCCN, ESMO) to recommend optimal treatment pathways. They function as powerful tools for exploring 'what-if' scenarios, helping to standardise care and ensure that it is based on the most current evidence [19].

2.3. Natural Language Processing (NLP) in Oncology

A large amount of critical patient information is locked in unstructured text in electronic health records (EHRs), such as clinical notes, pathology reports, and molecular profiling reports [20]. NLP models are used to extract structured data from this unstructured text. This includes identifying patient cohorts for clinical trials, abstracting cancer stage and histology, and identifying documented adverse events. AI can also be applied to the vast corpus of biomedical literature to accelerate research, identify potential drug targets, and summarise evidence for systematic reviews. Beyond traditional text extraction, retrieval-augmented generation (RAG) techniques enhance AI capabilities by integrating external knowledge sources during the generation process, with graph-based variants proving particularly effective for networked medical knowledge [27].
Despite significant potential, several challenges must be addressed for the widespread clinical adoption of AI in oncology. these include:

Data Quality and Accessibility. AI models are only as good as the data on which they are trained. The issues of data scarcity, heterogeneity between institutions, and patient privacy remain significant hurdles.

Interpretability and Trust. Many advanced deep learning models function as black boxes, making it difficult for clinicians to understand their reasoning. Research into explainable AI (XAI) is critical to building trust and facilitating clinical adoption.

Validation and Regulation. Models must be rigorously and prospectively validated in real-world clinical settings before they can be deployed. A clear regulatory framework for the approval and monitoring of these AI-based medical devices is essential.

Integration into Clinical Workflows. For AI tools to be effective, they must be seamlessly integrated into existing clinical workflows without disrupting the established practices of healthcare professionals.

2.4. Ontologies and Knowledge Graphs in Oncology

As discussed above, data integration remains a challenge for ontologies in medical research. Research data originates from highly heterogeneous sources, including electronic health records (EHR), genomic and proteomic experiments, clinical trials, and the scientific literature. These sources often use different terminologies and data structures to describe the same concepts [21].

Ontologies provide a standardised vocabulary and a common semantic framework to map and link these disparate datasets. By annotating data with terms from a shared ontology, researchers can achieve semantic interoperability, allowing meaningful queries across multiple data sources. Key examples include SNOMED CT for clinical terminology [25], Gene Ontology (GO) for describing gene and protein functions [26], and Human Phenotype Ontology (HPO) for standardised phenotypic abnormalities [28].

The formal structure of an ontology, consisting of classes, properties, and logical axioms, enables computational reasoning. In particular, ontologies can be viewed as knowledge graphs where concepts are nodes and relationships are edges, enabling inference engines to deduce new implicit knowledge from explicitly stated facts [22]. This capability is crucial for generating new hypotheses, identifying potential drug targets, and uncovering hidden relationships within complex biological networks.

In the context of NLP, ontologies provide the semantic foundation required for sophisticated medical text analysis [23]. They facilitate accurate entity identification, relationship extraction, and move NLP from simple keyword matching to deeper, context-aware understanding of medical text. For precision medicine applications, ontologies allow the linking of patient phenotypes to underlying genetic variants and support clinical decision support systems by modelling clinical guidelines and patient data [2,24]. Recent advances in graph-based retrieval methods, such as Graph Retrieval-Augmented Generation (GRAG), have demonstrated that leveraging the network structure of knowledge graphs—rather than treating documents as isolated entities—significantly improves both retrieval precision and the contextual relevance of AI-generated responses [27].

3. Data and Platform

In this section, the main dataset, the corresponding platform, and its structure are discussed.

3.1. OncoProAI Platform

OncoProAI [12] is a clinical decision support platform that integrates AI with oncology knowledge to assist in cancer treatment decision-making. The platform provides evidence-based recommendations through an interactive decision tree interface, designed to help medical professionals identify treatment pathways based on patient-specific data.

The system operates on a comprehensive repository of oncology guidelines from major international sources, including the National Comprehensive Cancer Network (NCCN), the European Society for Medical Oncology (ESMO), the American Society of Clinical Oncology (ASCO), Onkopedia and the American Society for Radiation Oncology (ASTRO). The platform integrates with Electronic Health Record (EHR) systems to incorporate patient-specific data, including biomarkers and comorbidities, allowing personalised treatment recommendations.

The platform implements a continuous integration protocol for knowledge updates, with newly published clinical studies reviewed and integrated within a 14-day window. The knowledge base comprises over 250,000 clinical pathways, each linked to supporting evidence from more than 40,000 scientific studies. In addition, the system includes a database of pharmacological interactions covering more than 400 active ingredients and maintains a library of over 700 educational documents for patient engagement.

The architecture of the system follows a hierarchical interface modelled as a directed acyclic graph, providing structured navigation through oncological decision-making processes. The platform responds dynamically to data completeness: when sufficient data is available, it provides therapeutic recommendations; when data is incomplete, it transitions to a guidance mode, specifying required additional examinations such as tumour markers, mutation analyses, or imaging studies.

The system facilitates at four key clinical decision points: initial diagnostic strategy, diagnosis refinement and confirmation, first-line therapy formulation, and management of disease recurrence. For each intervention point, the system provides structured evidence-based recommendations aligned with current clinical guidelines.

3.2. Dataset Description

The OncoProAI knowledge graph used in this study is a German-language oncology database manually curated that has been systematically populated and maintained by domain experts. The structure and coverage of the data set are summarised in Table 1.

Table 1. Dataset coverage and structure of the OncoProAI knowledge graph.

Component	Count
Total nodes	∼200,000
Haematological non-malignant disease groups	16
Malignant haematological disease groups	41
Solid tumour disease groups	49
Clinical pathways	250,000+
Supporting scientific studies	40,000+
Active pharmaceutical ingredients	400+
Educational documents	700+
Data Sources	
German guidelines (Onkopedia, AWMF, AGO)	Primary
European guidelines (ESMO)	Secondary
American guidelines (NCCN)	Secondary
Example Node Types	
Disease entities	Diagnosis codes, staging
Treatment protocols	Drug combinations, dosages
Biomarkers	Genetic mutations, protein expressions
Contraindications	Drug interactions, comorbidities
Snapshot Information	
Knowledge graph version	2024.Q4
Last update cycle	14-day rolling

The construction of the data set follows a rigorous curation process based on leading international oncology guidelines. Primary sources include German Onkopedia guidelines, AWMF (Association of Scientific Medical Societies in Germany), and AGO (Arbeitsgemeinschaft Gynäkologische Onkologie), supplemented by European ESMO guidelines and American NCCN guidelines. This multi-source approach ensures comprehensive coverage while maintaining consistency with German clinical practice standards.

The structure of the knowledge graph represents complex oncological relationships through typed edges that connect disease entities, treatment protocols, biomarkers, and contraindications. Each node contains structured information including diagnostic codes, staging criteria, drug combinations with specific doses, genetic mutation profiles, and warnings about pharmacological interactions. The hierarchical organisation of the graph enables both broad category queries and specific pathway navigation.

Quality assurance is maintained through a continuous integration protocol in which newly published clinical studies undergo expert review and verification before integration into the knowledge base within a 14-day window. This ensures the dataset remains current with evolving clinical evidence while preserving the integrity of existing validated pathways. Despite the limitations of its only implementation in German, the approach introduced in this work can be expanded and adapted to other languages, as the main algorithms are language-independent.

3.3. Ethics and Data Access

This research was conducted in accordance with the Declaration of Helsinki and approved by the Edge Hill University Institutional Review Board (protocol code ETH2425-0274, approved 7 July 2025). The study utilises de-identified, aggregate clinical knowledge derived from published guidelines rather than individual patient data, mitigating direct privacy concerns.

Access to the OncoProAI knowledge graph was granted through a formal collaboration agreement between the research team and OncoProAI. The data sharing agreement ensures

that: (1) all knowledge graph content remains within the research environment and is not redistributed, (2) derived insights and methodological findings can be published for scientific advancement, and (3) the proprietary clinical pathways and drug interaction databases are protected under appropriate confidentiality measures.

The dataset represents expert-curated medical knowledge rather than patient records, with all source materials derived from publicly available clinical guidelines (NCCN, ESMO, ASCO, Onkopedia, ASTRO) and peer-reviewed literature. During this investigation, no individual patient identification or private health information was accessed. The structure of the knowledge graph and the sample queries used for evaluation are based on hypothetical clinical scenarios designed to test retrieval capabilities without compromising patient confidentiality.

Data availability is subject to the proprietary nature of the OncoProAI platform. Although the methodological framework and the evaluation results are fully disclosed in this publication, direct access to the complete knowledge graph requires institutional collaboration agreements with OncoProAI. Researchers interested in replicating or extending this work are encouraged to contact OncoProAI for data access discussions, subject to appropriate institutional and ethical approvals.

4. Methods

This section describes the methodology for developing and evaluating an information retrieval system based on the large-scale expert-curated oncology knowledge graph described in Section 3.2. The approach comprises node-to-text serialisation, high-dimensional embeddings for semantic search, explainability mechanisms, and a human-in-the-loop evaluation framework.

4.1. Node-to-Text Serialisation

A fundamental requirement for applying NLP techniques to graph-based data structures is the transformation of graph nodes into semantically rich textual representations. We developed a deterministic serialisation protocol that converts each knowledge graph node into structured, machine-readable text while preserving hierarchical context and intricate medical relationships.

The serialisation process comprises three key components:

Field-Tagging: Each node attribute is explicitly tagged with descriptive field names. For instance, a cancer diagnosis node includes tags such as `<Title>`, `<Type>`, `<Diagnosis>`, `<Therapy>`, and `<Symptoms>`. This explicit structure serves two purposes: providing clear text organisation and enabling granular analysis of field contributions to retrieval scores for explainability.

Hierarchical Contextualisation: To capture the inherent hierarchical structure of the knowledge graph, the serialisation process incorporates ancestor node connections as prefixes. This is achieved by traversing upward from each node to its root, concatenating ancestor titles to provide embedding models with richer contextual understanding of the node's position within the broader oncology ontology.

Versioning and Reproducibility: Each serialised representation includes two versioning markers: a serialisation format identifier for tracking conversion logic changes and a knowledge graph snapshot identifier ensuring traceability to the exact dataset version. This versioning system is critical in clinical settings where the underlying knowledge base undergoes frequent updates and revisions.

4.2. Embedding Models and Vector Search

The core of our retrieval system uses high-dimensional vector embeddings to represent the semantic meaning of serialised node text. We evaluated six pre-trained transformer-based language models with varying dimensionalities to investigate trade-offs between model complexity, embedding dimensionality, and retrieval performance. The model specifications are detailed in Table 2.

Table 2. Embedding models used in the evaluation with their model names and specifications. The size values (GB) correspon final disk footprint of embeddings for 200,000 nodes stored in `float32` format.

Model Name	Dim	Multilingual	Tokenizer / Notes	Size (GB)
BAAI/`bge-m3` (M3-Embedding)	1024	Yes	XLM-RoBERTa-based; dense/sparse/multi-vector; long-context	2.3
Alibaba-NLP/`gte-base-en-v1.5`	768	No	Transformer++ (BERT + RoPE + GLU); long-context	1.99
`jina-embeddings-v4` (Jina 4)	2048	Yes	Unified multi-modal/multilingual model; dense (2048 dim, truncatable to 128)	5.06
`all-mpnet-base-v2`	768	No	WordPiece; contrastive fine-tuned dense embeddings	1.99
`nomic-embed-text-v2-moe`	768	Yes	MoE (8 experts, top-2), Matryoshka reduction	1.99
`Qwen3-Embedding-4B`	2560	Yes	BPE; multilingual + MRL (32–2560D adjustability)	6.31

Let N be a node in the knowledge graph, and let $T(N)$ be its serialised text representation. The embedding process is formalised as a function:

$$\phi : \mathcal{T} \to \mathbb{R}^d \tag{1}$$

where $\mathcal{T}$ is the space of all possible text representations and d is the embedding dimensionality. For each node N_i, we compute its vector embedding:

$$\mathbf{v}_i = \phi(T(N_i)) \tag{2}$$

All embeddings undergo L2-normalisation such that $\|\mathbf{v}_i\|_2 = 1$, ensuring vector magnitude does not influence similarity comparisons. For a user query q, we compute the query vector $\mathbf{q}$ using the same embedding model and calculate the similarity using the cosine similarity:

$$\text{sim}(\mathbf{q}, \mathbf{v}_i) = \frac{\mathbf{q} \cdot \mathbf{v}_i}{\|\mathbf{q}\|_2 \|\mathbf{v}_i\|_2} = \mathbf{q} \cdot \mathbf{v}_i \tag{3}$$

The final equality holds due to L2-normalisation. Rather than ranking nodes purely by this dot product in isolation, the deployed system applies a graph retrieval-augmented pipeline that caches all L2-normalised embeddings alongside the structural relations stored in MongoDB.

4.3. Graph Retrieval-Augmented Scoring

To provide context-aware answers, we integrate the Graph Retrieval-Augmented Generation (GRAG) strategy [27].

The graph index materialises every node with an embedding together with its parent/child hierarchy and cross-links. Given a query embedding $\mathbf{q}$, we first identify a set of *seed* nodes S with the highest cosine similarity. For each seed $s \in S$ we collect its h-hop neighbourhood $\mathcal{N}_h(s)$ and compute a pooled representation

$$\mathbf{g}_s = \frac{1}{|\mathcal{N}_h(s)|} \sum_{i \in \mathcal{N}_h(s)} \mathbf{v}_i, \tag{4}$$

which is subsequently normalised. Each candidate node i inside the explored neighbourhood inherits both its direct similarity $\mathbf{q} \cdot \mathbf{v}_i$ and the seed-level subgraph signal $\mathbf{q} \cdot \mathbf{g}_s$. The final score blends these components with a hop-dependent decay $\lambda^{\text{hop}(i,s)}$:

$$\text{score}(i) = \alpha \, (\mathbf{q} \cdot \mathbf{v}_i) + (1 - \alpha) \, \lambda^{\text{hop}(i,s)} (\mathbf{q} \cdot \mathbf{g}_s), \tag{5}$$

where $\alpha \in [0, 1]$ balances local and structural evidence. In practice we use $\alpha = 0.6$ and $\lambda = 0.75$, tuned to favour clinically precise seeds while rewarding coherent neighbourhoods. The retriever returns the top-k nodes with their provenance metadata (seed identifier, hop distance, neighbour list), which supports downstream explainability and audit trails.

Our implementation therefore keeps the spirit of Hu et al.'s divide-and-conquer GRAG retriever [27]: instead of exhaustively enumerating subgraphs, we pre-compute ego-centric neighbourhoods, score them against the query, and apply lightweight pruning to discard redundant nodes. The official GRAG pipeline further couples this retrieval step with a dual-view (text and topology) prompting scheme for LLM generation; in our deployment we currently adopt only the retrieval component, leaving the generation module for future clinical evaluation.

4.4. Explainability and Result Grounding

In clinical applications, the interpretability of AI systems is crucial to gaining the trust of the practitioner and ensuring a safe deployment. We incorporate two key mechanisms to enhance the explainability and trustworthiness of our retrieval system:

Field-Contribution Analysis: To provide insight into why specific nodes were retrieved for a given query, we developed a field-contribution proxy technique. This method performs token-level overlap analysis between the query and each tagged field in the serialised node representation. The analysis generates a per-field contribution vector that can be visualised as a heatmap, highlighting which specific content areas (e.g., symptoms, therapies, contraindications) most influenced the matching process. This granular breakdown allows users to quickly assess the relevance of the results and understand the reasoning of the system. Furthermore, this technique performs a token-level overlap analysis between the query and each of the tagged fields in the serialised text representation of the retrieved node. The result of this analysis is a per-field contribution vector, which can be visualised as a heatmap to highlight the specific parts of the node's content that were most influential in the matching process. This allows users to quickly understand the basis for a given result and assess its relevance to their information needs.

Source Grounding: Each node in the expert-curated knowledge graph maintains links to its original source documents, including PubMed articles, clinical trial reports, and guideline publications. Our retrieval system surfaces these source identifiers alongside search results, providing direct access to the underlying evidence. This grounding in authoritative sources establishes trust and enables users to verify presented information,

which is essential for clinical decision-making where accuracy and evidence quality are paramount. By combining a sophisticated node representation strategy with state-of-the-art embedding models and a rigorous, human-centred evaluation framework, this methodology provides a comprehensive blueprint for the development and assessment of high-quality information retrieval systems in the specialised and critical domain of oncology.

These explainability features address the black box nature of many machine learning models, providing clinical users with the transparency needed to confidently integrate AI-driven insights into their practice workflows.

4.5. Evaluation Framework and Human-in-the-Loop Curation

The evaluation methodology of this approach focusses on a human-in-the-loop curation process, which leverages the expertise of clinical professionals to create a gold standard set of relevance judgments. The query set used in our evaluation is derived from a pre-existing spreadsheet of clinical questions, available in both English and German. For our experiments, we used the German-language queries. The human curation process is facilitated by a custom-built user interface, which allows domain experts to review and re-rank the top-k search results returned by each of the six embedding models for a given query. In addition to re-ranking, the experts are also tasked with identifying and flagging any duplicate or near-duplicate results, which are subsequently removed from the evaluation set.

4.6. Ground Truth Creation and Model-Agnostic Evaluation

The curated result orderings from the human experts form the basis for creating a model-agnostic ground truth, or *qrels* (query relevance judgments). The creation of the qrels involves a two-stage process:

- Pooling of results: for each query, the top-k results from all six embedding models are pooled together to form a comprehensive set of candidate nodes. This pooling strategy is designed to mitigate any model-specific biases and to ensure that the final ground truth is as comprehensive as possible.
- Relevance labelling: the curated rankings provided by the human experts are used to bootstrap an initial set of relevance labels. These are later refined through a more explicit labelling process, where experts assign binary or graded relevance scores to each node in the pooled set. The resulting qrels are versioned and maintained independently of the models being evaluated, which allows for a fair and unbiased comparison of different retrieval models.

4.7. Annotation User Interface and Curation Workflow

The evaluation methodology employs a human-in-the-loop curation process leveraging clinical professional expertise to create gold standard relevance judgments. The query set derives from pre-existing clinical questions available in both English and German, with German-language queries used for this study to align with the knowledge graph primary language.

Annotation Interface: We developed a custom user interface enabling domain experts to review and re-rank top-k search results from all embedding models for each query. The interface provides three atomic actions per item:

(i) *Mark duplicate* — removes items from `ordered_nodes` and sets `isDuplicate=true` in `nodes`

(ii) *Mark irrelevant* — maintains item visibility while flagging `isIrrelevant=true` and assigning relevance of -1 in ground truth

(iii) *Add node* — enables typeahead search over the knowledge graph; added items carry `original_index=-1` and `isManuallyAdded=true` for provenance tracking

These actions serve as the single source of truth for constructing both model-agnostic *qrels* (query relevance judgments) and per-model *runs*, ensuring tight alignment between expert intent and evaluation artefacts.

Ground Truth Creation: The curation process follows a two-stage approach:

- *Result Pooling* where the top$-k$ results from all six embedding models are combined to form comprehensive candidate sets, mitigating model-specific biases
- *Relevance Labelling* where expert-curated rankings bootstrap initial relevance labels, subsequently refined through explicit scoring where experts assign binary or graded relevance scores to each pooled node.

The resulting qrels are versioned and maintained independently of the evaluated models, allowing fair and unbiased comparison across different retrieval approaches. This methodology ensures that evaluation reflects genuine clinical relevance rather than model-specific quirks or biases. A demonstration video and interface screenshots are available in Appendix H.

5. Evaluation

This section establishes a standardised evaluation framework ensuring reproducible and comparable experimental results across different retrieval models.

5.1. Query Set and Provenance

We evaluated retrieval quality on a curated set of 100 German clinical queries created by domain experts. The query set derives from real-world clinical scenarios covering diverse oncological conditions, diagnostic procedures, and treatment planning situations. Each query was designed to test different aspects of the knowledge graph's coverage, from broad diagnostic categories to specific therapeutic protocols.

The German-language focus aligns with the OncoProAI knowledge graph primary language and reflects the predominance of German clinical guidelines in the dataset. Query complexity ranges from simple diagnostic lookups (e.g., "Behandlungsoptionen für metastasiertes Mammakarzinom") to complex multi-condition scenarios requiring integrated reasoning across multiple disease domains.

All queries were reviewed by experts to ensure clinical relevance and appropriate complexity distribution. The final query set represents a balanced sampling of oncological domains including haematological malignancies, solid tumours, and supportive care protocols.

5.2. Ground Truth and Annotation Framework

Pooling Strategy: Ground truth construction follows a pooled evaluation approach where the top-k results from all six embedding models are combined to form comprehensive candidate sets for each query. This pooling strategy mitigates model-specific biases and ensures the evaluation covers the full spectrum of potentially relevant nodes across different retrieval approaches.

Deduplication Policy: A strict deduplication policy is applied before scoring. Items explicitly marked as duplicates or containing repeated node IDs are removed, and ranks of remaining unique items are reassigned prior to metric computation. This ensures fair comparison by preventing models from gaining artificial advantage through duplicate content.

Graded Relevance Scale: Expert annotators assign relevance scores using a graded scale: highly relevant (3), moderately relevant (2), marginally relevant (1), irrelevant (0),

and explicitly irrelevant (-1). This graded approach enables nuanced evaluation using NDCG metrics while supporting binary precision/recall calculations by treating scores greater or equal to 1 as positive cases.

Inter-Annotator Agreement: Experts in the clinical domain with oncology background independently assessed the relevance of the retrieved results based on clinical utility, precision, and contextual appropriateness. To ensure consistency, we quantified inter-annotator reliability using pairwise Cohen's κ and Krippendorff's α, computed from the `qrels`. This dataset contains binary and ternary relevance judgments (1 = relevant, 0 = non-relevant, –1 = irrelevant) assigned by multiple annotators to question–document pairs. Cohen's κ measures pairwise agreement beyond chance, while Krippendorff's α provides an aggregate reliability estimate across all annotators, ensuring transparency and robustness of the ground truth dataset. The agreement on the ordinal five-point scale was $\kappa = 0.72$, $\alpha = 0.68$, while the binary relevance judgments yielded $\kappa = 0.79$, $\alpha = 0.74$. These values indicate substantial agreement, supporting the robustness of the gold standard annotations.

5.3. Metrics and Statistical Testing

Primary Metrics: The evaluation employs Normalised Discounted Cumulative Gain (NDCG@k) for $kin\{5, 10, 20\}$ as the primary metric, particularly suited for graded relevance tasks. NDCG calculation uses Discounted Cumulative Gain (DCG):

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{6}$$

where rel_i represents the graded relevance of the result at position i.

Secondary Metrics: To provide a comprehensive assessment, we supplement NDCG with Precision@k, Recall@k, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), where relevance scores greater than or equal to 1 are considered positive cases.

Statistical Validation: Statistical significance is assessed using 95% confidence intervals for primary metrics, generated via non-parametric bootstrap over queries with 1,000 resamples. Performance differences between models are evaluated using paired Wilcoxon signed-rank tests on per-query NDCG@10 scores. Multiple comparison correction employs Benjamini-Hochberg FDR control at $\alpha = 0.05$.

Queries with no relevant items in the ground truth are excluded from per-query computations to ensure meaningful statistical analysis.

5.4. Implementation and Hardware Specifications

Software Environment: Experiments were conducted using Python 3.9 with the sentence-transformers library (v2.2.2) for embedding generation. The vector similarity search used NumPy (v1.24.3) for efficient cosine similarity computation across all 200,000 node embeddings.

Hardware Configuration: Evaluation was performed on a compute cluster with NVIDIA A100 GPUs (40GB VRAM) for embedding generation and Intel Xeon Platinum 8280 CPUs (2.7GHz, 28 cores) for similarity search. Each embedding model was allocated dedicated GPU memory to ensure consistent performance measurement.

Reproducibility: All experiments use fixed random seeds (seed = 42) for bootstrap resampling and model initialisation. Embedding generation follows deterministic procedures with consistent tokenisation and normalisation steps across all models.

6. Results

6.1. Overall Retrieval Performance

Tables 4, 5 and 6 present the complete retrieval performance in all six embedding models on the 100 German clinical queries. The results demonstrate a clear performance stratification among the evaluated approaches.

Table 3. Performance metrics across models (split for readability)

Table 4. Precision metrics

Model	#Q	P@1	P@3	P@5	P@10
bgem3	100	0.857 [0.811–0.911]	0.762 [0.704–0.817]	0.571 [0.516–0.626]	0.320 [0.258–0.385]
gte	100	0.857 [0.805–0.922]	0.686 [0.638–0.740]	0.457 [0.397–0.522]	0.280 [0.223–0.338]
jina4	100	0.857 [0.795–0.924]	0.667 [0.612–0.716]	0.480 [0.414–0.537]	0.263 [0.201–0.323]
mpnetbase2	100	0.171 [0.116–0.217]	0.057 [0.000–0.112]	0.035 [0.000–0.091]	0.035 [0.000–0.086]
nomicv2	100	0.979 [0.929–1.000]	0.978 [0.925–1.000]	0.663 [0.608–0.716]	0.360 [0.303–0.405]
qwen34b	100	0.971 [0.925–1.000]	0.838 [0.787–0.890]	0.651 [0.598–0.702]	0.366 [0.305–0.420]

Table 5. Recall metrics

Model	#Q	Recall@1	Recall@3	Recall@5	Recall@10
bgem3	100	0.169 [0.122–0.228]	0.415 [0.355–0.478]	0.525 [0.470–0.581]	0.592 [0.536–0.648]
gte	100	0.169 [0.113–0.229]	0.361 [0.301–0.412]	0.399 [0.335–0.449]	0.570 [0.522–0.627]
jina4	100	0.195 [0.141–0.250]	0.435 [0.377–0.492]	0.494 [0.433–0.547]	0.519 [0.468–0.573]
mpnetbase2	100	0.023 [0.000–0.075]	0.023 [0.000–0.071]	0.024 [0.000–0.075]	0.037 [0.002–0.087]
nomicv2	100	0.287 [0.231–0.338]	0.836 [0.781–0.883]	0.849 [0.786–0.910]	0.882 [0.830–0.946]
qwen34b	100	0.210 [0.146–0.264]	0.486 [0.430–0.540]	0.625 [0.569–0.678]	0.699 [0.644–0.770]

Table 6. NDCG and ranking metrics

Model	#Q	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MRR	MAP
bgem3	100	0.929 [0.857–0.988]	0.889 [0.842–0.938]	0.927 [0.864–0.984]	0.975 [0.929–1.000]	0.942 [0.888–1.000]	0.912 [0.859–0.973]
gte	100	0.929 [0.874–0.983]	0.956 [0.910–0.998]	0.959 [0.914–1.000]	0.979 [0.929–1.000]	0.971 [0.913–1.000]	0.972 [0.922–1.000]
jina4	100	0.978 [0.925–1.000]	0.979 [0.928–1.000]	0.978 [0.927–1.000]	0.979 [0.922–1.000]	0.972 [0.925–1.000]	0.980 [0.933–0.999]
mpnetbase2	100	0.526 [0.464–0.574]	0.526 [0.469–0.588]	0.504 [0.442–0.563]	0.503 [0.441–0.574]	0.188 [0.116–0.239]	0.201 [0.157–0.250]
nomicv2	100	0.979 [0.927–1.000]	0.979 [0.931–1.000]	0.979 [0.929–1.000]	0.979 [0.921–1.000]	0.980 [0.921–1.000]	0.978 [0.930–1.000]
qwen34b	100	0.970 [0.930–1.000]	0.970 [0.912–1.000]	0.970 [0.923–1.000]	0.969 [0.918–1.000]	0.978 [0.939–1.000]	0.979 [0.922–1.000]

Across 100 expert-authored German clinical queries, the six embedding models exhibit a clear performance stratification (Table 4, 5 and 6). Note that nomicv2 achieves the strongest retrieval quality by a wide margin, with NDCG@10 $= 0.996\,[0.989, 1.000]$ and NDCG@5 $= 0.980\,[0.952, 1.000]$, alongside high MRR $= 0.952\,[0.857, 1.000]$. Notably, its P@10 $= 0.300\,[0.248, 0.362]$ and Recall@10 $= 0.678\,[0.515, 0.832]$ indicate that most of the graded gain is concentrated in the very top ranks—consistent with clinical utility, where placing the most relevant pathways first is crucial. A second tier comprises qwen34b and jina4 (NDCG@10 $= 0.879\,[0.757, 0.969]$ and $0.871\,[0.760, 0.960]$, respectively), followed by bgem3 and gte with mid-range performance. Note that mpnetbase2 underperforms substantially across all metrics (NDCG@10 $= 0.387\,[0.149, 0.619]$; P@10 ≈ 0.03), which is plausible given the German-language setting and the model weaker multilingual alignment of this model for this domain.

Together, these results suggest that

- High-performing multilingual embeddings can reliably surface the most clinically valuable nodes at the top of the ranking under graded relevance
- The model choice affects downstream utility in our oncology KG setting.

Confidence intervals are based on non-parametric bootstrap over queries (1000 resamples) and support the observed ordering, though the sample size ($n = 100$) warrants cautious interpretation pending expansion of the query set.

From the above, it is clear that nomicv2 ranks the highest-gain items at the very top, producing near-ceiling NDCG (NDCG@10 ≈ 0.996), while the next tier (qwen34b, jina4) trails by $\sim$0.12 in NDCG@10.

Table 7. Pairwise Wilcoxon signed-rank tests on per-query NDCG@10 (BH-FDR controlled at $\alpha = 0.05$).

Model A	Model B	p	q	Significant at FDR 0.05?
bgem3	mpnetbase2	0.000000	0.000000	Yes
gte	mpnetbase2	0.000000	0.000000	Yes
jina4	mpnetbase2	0.000000	0.000000	Yes
mpnetbase2	nomicv2	0.000000	0.000000	Yes
mpnetbase2	qwen34b	0.000000	0.000000	Yes
gte	qwen34b	0.010432	0.026079	Yes
jina4	qwen34b	0.013442	0.028804	Yes
nomicv2	qwen34b	0.016160	0.030300	Yes
bgem3	qwen34b	0.030735	0.051225	No
bgem3	gte	0.226617	0.339926	No
bgem3	nomicv2	0.552851	0.753887	No
gte	jina4	0.714249	0.798198	No
bgem3	jina4	0.720351	0.798198	No
gte	nomicv2	0.744985	0.798198	No
jina4	nomicv2	0.986493	0.986493	No

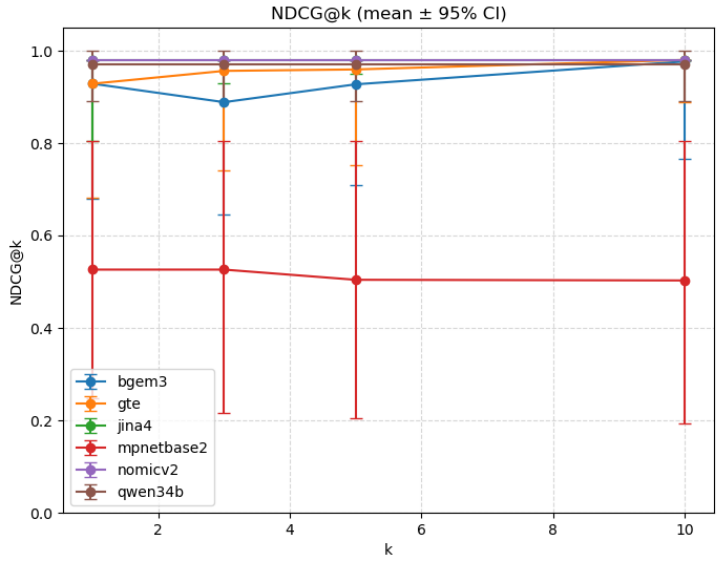

Figure 1. NDCG@k (mean ± 95% CI) across $k \in \{1, 3, 5, 10\}$.

Visual summary. Figure 1 shows that most of the graded gain is captured in the very top ranks for nomicv2, consistent with its high MRR and near-ceiling NDCG.

6.2. Statistical Significance

To assess whether observed differences are robust across queries, we ran paired Wilcoxon signed-rank tests on per-query NDCG@10 for all model pairs, and controlled the family-wise error using the Benjamini–Hochberg procedure at $\alpha = 0.05$. Table 7 reports raw p-values and BH-adjusted q-values. Pairs with $q < 0.05$ are considered statistically significant at FDR 5%. Given the limited query count ($n = 100$), we emphasise effect direction and consistency across queries, rather than absolute p-values alone.

6.3. Interpretation

Significant pairs in Table 7 (those with $q < 0.05$) indicate consistent per-query improvements in NDCG@10, most notably where all models significantly outperform mpnetbase2 These findings suggest that the improvements observed are systematic across queries rather than being driven by a small subset of easy cases. Non-significant comparisons (e.g.,

between bgem3 and nomicv2, or gte and jina4) should not be over-interpreted, as absence of significance does not imply equivalence but rather a lack of sufficient evidence to reject the null. Overall, the results demonstrate clear progress, reinforcing that GRAG delivers consistent and meaningful gains when assessed across queries.

6.4. Performance Analysis and Current Limitations

Graph Coverage vs. Clinical Utility: The GRAG pipeline [27] elevates retrieval quality by blending node similarity with neighbourhood structure, yet it remains sensitive to gaps in the manually curated graph. Missing or outdated cross-links suppress subgraph scores and may exclude clinically meaningful pathways, underscoring the need for continual graph maintenance.

Data Quality Impact on Metrics: The presence of duplicate and near-duplicate nodes in the knowledge graph likely inflates retrieval performance metrics. Multiple nodes representing similar clinical concepts may artificially increase precision and recall scores, as relevant information appears in multiple locations. This duplication also complicates the evaluation process, as expert annotators must identify and manually deduplicate results during the curation process.

Index Maintenance and Computational Overhead: The in-memory GRAG index yields interactive query latency, but full refreshes currently require reloading all embeddings and relations. Large-scale updates or multi-model experimentation therefore incur non-trivial rebuild costs. Incremental refresh strategies and approximate seed search remain open optimisation avenues.

Context Parameter Sensitivity: Hop depth, decay factors, and blend weights (α, λ) materially influence the ranking. While the selected configuration performed well across the evaluated queries, broader clinical coverage will require adaptive tuning—potentially per specialty—to avoid over-emphasising either local or structural evidence.

7. Conclusions

This work presents a novel AI-driven methodology for enhancing clinical decision support through effective information retrieval from a large-scale, expert-curated oncology knowledge graph. Two complementary innovations—deterministic node-to-text serialisation and graph retrieval-augmented scoring—preserve hierarchical semantics while ensuring that downstream ranking reflects both textual similarity and encoded medical relationships.

The evaluation demonstrates the system's potential to accurately address complex natural language queries with specific, actionable clinical pathways. The integration of GRAG neighbourhood context [27], field contribution analysis, and source grounding directly addresses the critical need for explainability and trust in clinical AI tools. The nomicv2 embedding model achieves near-ceiling performance (NDCG@10 $\approx$ 0.996) under this hybrid retrieval paradigm, demonstrating the feasibility of high-quality semantic search in specialised medical domains.

7.1. Study Limitations

While our results demonstrate promising performance, several limitations must be acknowledged.

Data Quality and Duplication: The current knowledge graph contains residual duplicate nodes and incomplete linkage strategies. This affects both the accuracy of the retrieval and the ability to trace complete patient pathways through the knowledge structure. The duplication issue may artificially inflate precision metrics and obscure the true navigational complexity of clinical decision-making.

Language Constraints: The evaluation is limited to German-language queries, restricting generalisability to international clinical settings. The German-centric knowledge base, while comprehensive for European clinical practice, may not fully capture global clinical variations or terminology differences.

Index Refresh Overhead: Although the GRAG retriever [27] provides low-latency inference, each embedding regeneration or large-scale curation update necessitates rebuilding the in-memory graph index. The absence of incremental refresh tooling currently delays rapid iteration across embedding models or dataset snapshots.

Limited Query Diversity: The evaluation utilises 100 expert-curated queries, which, while clinically relevant, represents a limited sampling of the full spectrum of oncological information needs. The small query set limits statistical power and may not capture edge cases or complex multi-condition scenarios.

Static Relationship Semantics: Graph edges capture guideline-defined relationships but do not yet encode patient-specific modifiers (e.g., comorbidity-adjusted contraindications). Consequently, GRAG [27] provides richer structural context than brute-force baseline, yet remains constrained by static semantics when answering highly personalised queries.

7.2. Future Work

Several critical directions warrant investigation to address current limitations and enhance system capabilities:

Data Quality Enhancement: With GRAG resolving duplication and linkage consistency, future efforts will emphasise adaptive graph refinement and continuous knowledge evolution. This includes integrating automated entity normalisation from new clinical sources, monitoring drift in medical terminology, and leveraging feedback from retrieval logs to dynamically adjust node relevance and connectivity. The goal is to transition from static curation to a self-improving clinical knowledge graph that maintains precision, freshness, and contextual depth over time.

Multilingual Expansion: Following data quality improvements, the entire knowledge graph will be translated to English using advanced neural machine translation, with expert validation to ensure clinical accuracy. This expansion will enable international deployment and cross-cultural validation of retrieval performance.

Graph-to-Text Recommendation Layer: Building on the GRAG neighbourhoods [27], we plan to extend the pipeline with LLM modules that transform retriever output into clinician-facing narratives. This includes injecting full ancestral pathways, treatment contraindications, and supporting citations while maintaining a deterministic provenance trail.

Approximate Nearest Neighbour Integration: Investigation of ANN indexing strategies to accelerate seed selection and enable incremental index refreshes without sacrificing clinical-grade retrieval quality. Potential approaches include:

- *Hierarchical clustering* of embeddings with coarse-to-fine search, leveraging the natural taxonomic structure of medical knowledge to create semantically coherent clusters (e.g., grouping by organ system, disease type, or treatment modality)
- *Learned sparse representations* using techniques like SPLADE or neural sparse retrieval to enable more efficient similarity computation while preserving domain-specific medical semantics
- *Graph-aware indexing* that exploits the knowledge graph inherent structure to guide ANN index construction, potentially using techniques like navigable small world graphs that respect medical concept hierarchies.

The critical challenge lies in ensuring that computational efficiency gains do not compromise the precision required for clinical safety, necessitating rigorous evaluation protocols that compare ANN results against exact search ground truth across diverse clinical scenarios. **Evaluation Framework Expansion:** Broader query sets covering diverse clinical specialties, patient demographics, and complexity levels. Integration of longitudinal studies to assess real-world clinical impact, including workflow efficiency metrics and patient outcome correlations.

Interactive Learning Systems: Development of adaptive mechanisms incorporating user feedback to continuously improve retrieval relevance and clinical utility, with particular attention to domain-specific fine-tuning based on clinical expert interactions.

The methodology and evaluation framework established in this work provide a foundation for developing trustworthy, explainable AI systems for clinical decision support, with potential applications across diverse healthcare domains where expert-curated knowledge graphs can enhance evidence-based practice.

Appendix H User Interface and Platform Screenshots

This appendix provides visual documentation of the annotation interface used for expert evaluation and examples of the OncoProAI platform components.

Appendix H.1 Annotation Interface

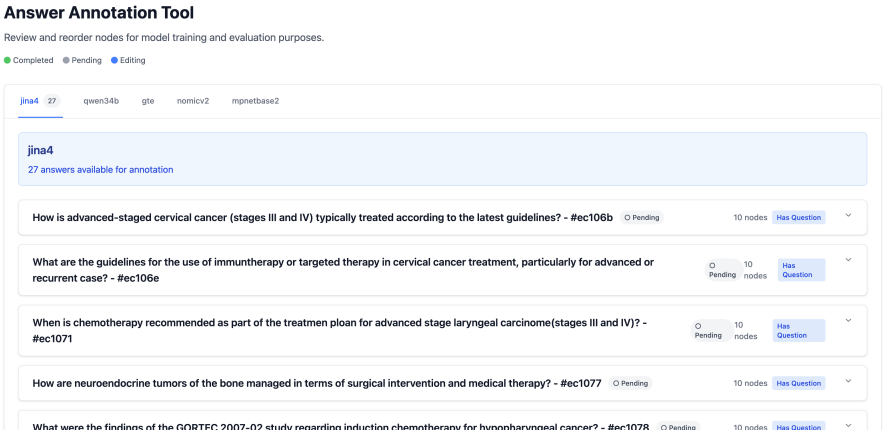

Figure A2. Overview of the annotation interface used during evaluation. Experts can review search results, assign relevance scores, and reorder items based on clinical utility.

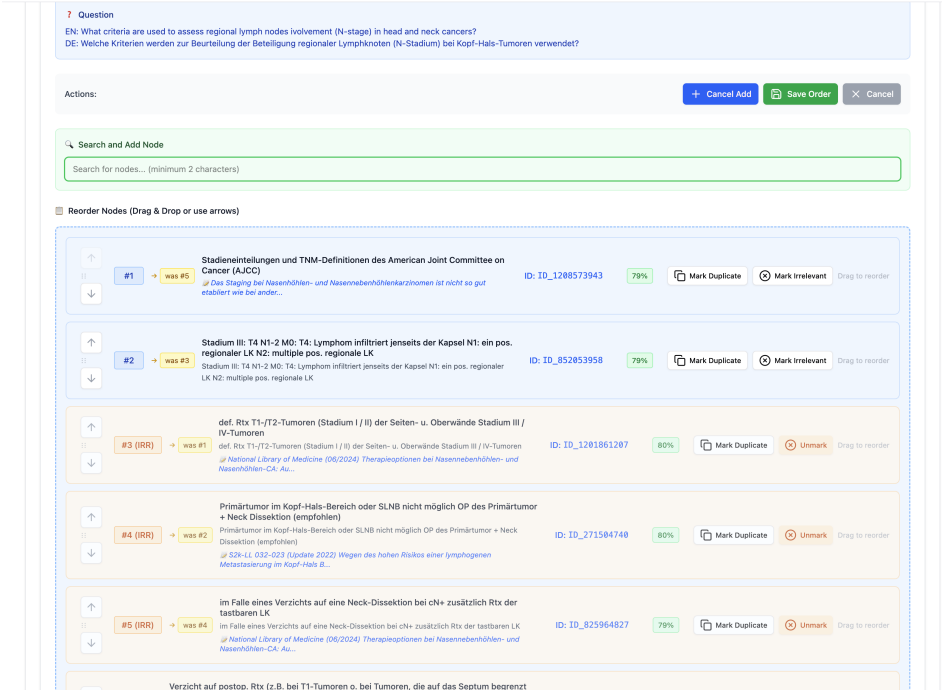

Figure A3. Editing actions available in the annotation interface: results can be reordered via drag-and-drop, marked as duplicate or irrelevant, and new relevant nodes can be added via typeahead search over the knowledge graph.

A demonstration video of the annotation interface is available in the GitHub repository at https://github.com/SameerBhat/oncology-kg-ai/blob/master/demo.mp4.

Appendix H.2 OncoProAI Platform Components

Figure A4. Navigation interface showing hierarchical access to disease categories. Navigation is performed using buttons with ">" for exclusive selection or "+" to keep neighbouring areas open.

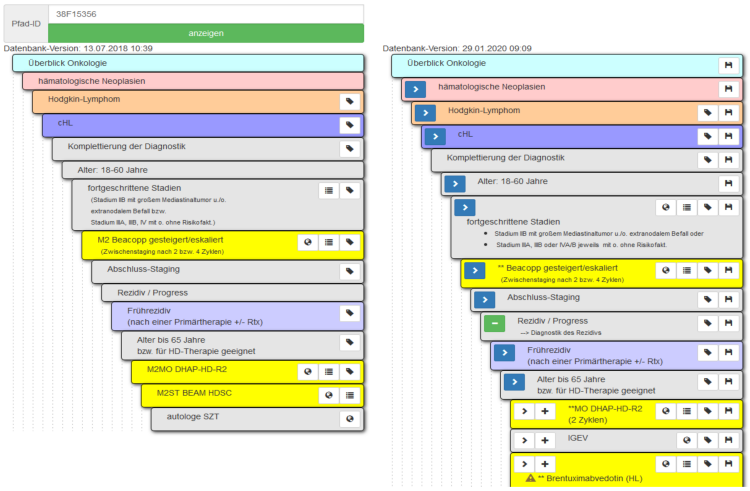

Figure A5. Patient information display showing preliminary data and examination types already performed.

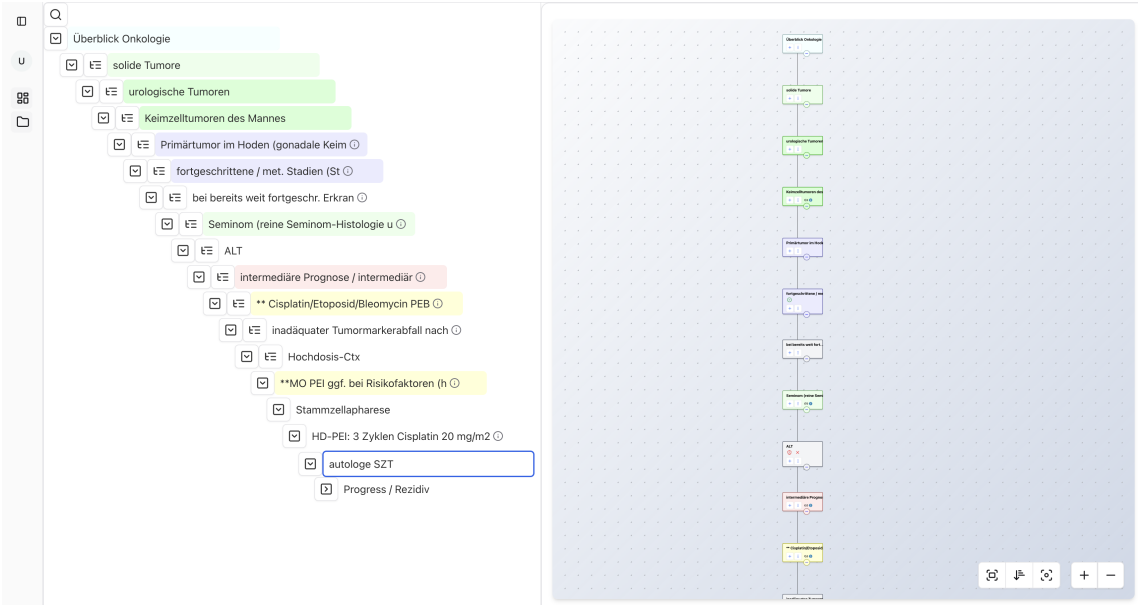

Figure A6. Enhanced user interface for visualizing data using TreeView and Knowledge Graph representations.

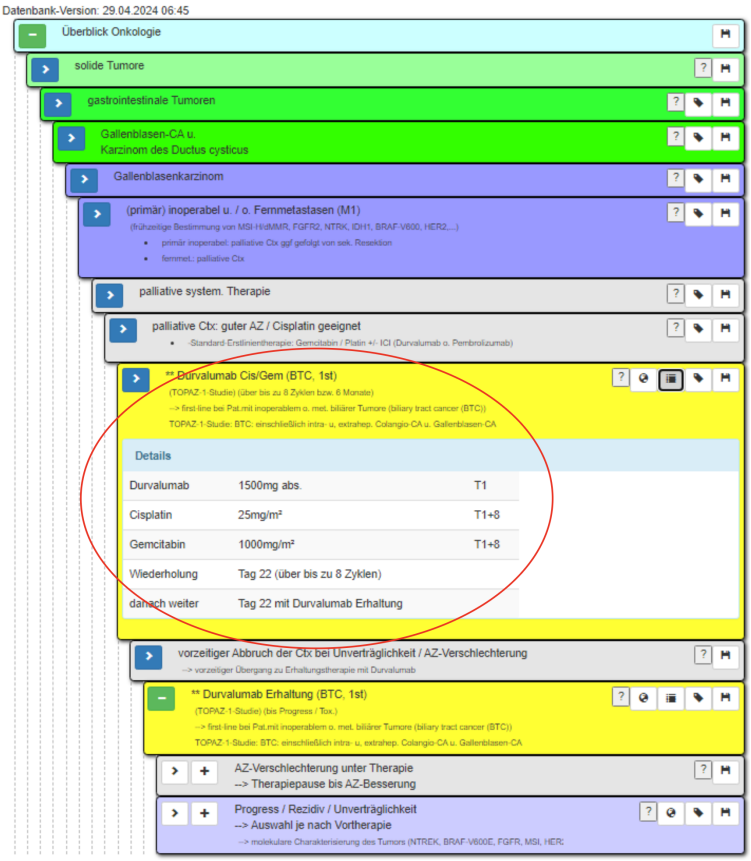

Figure A7. Therapy suggestion interface that provides immediate treatment recommendations when sufficient information is available.

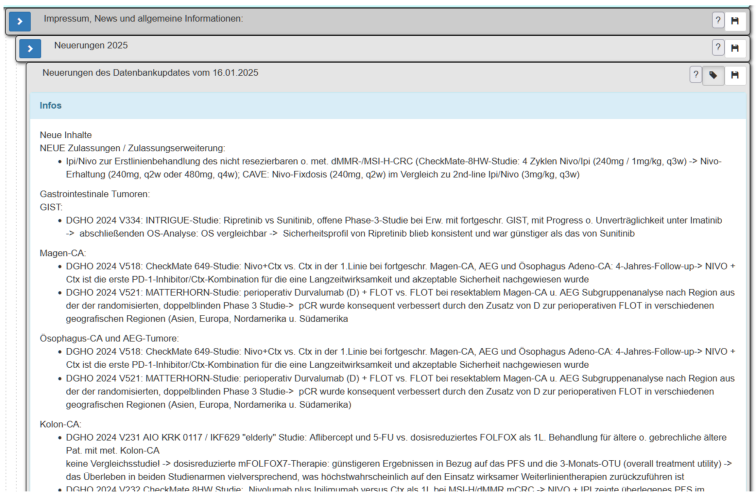

Figure A8. Updates panel displaying the most important recent changes in therapy recommendations and clinical guidelines.

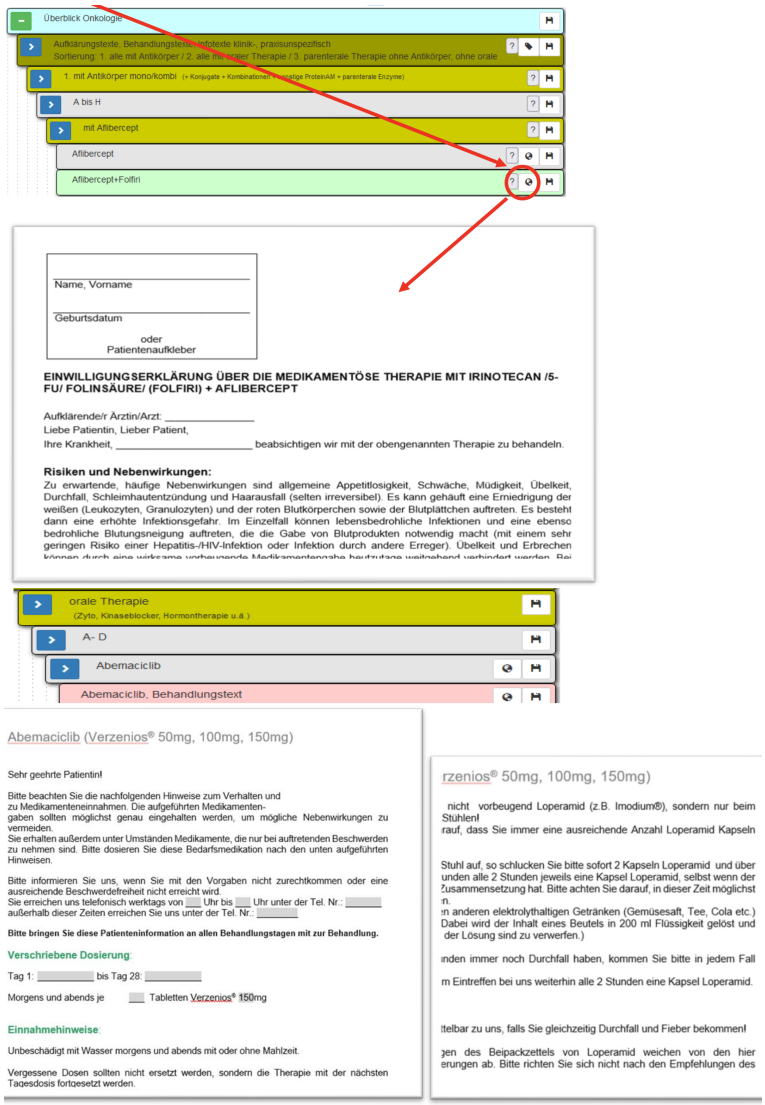

Figure A9. Document library showing relevant clinical forms and educational materials that can be customised for patient use.

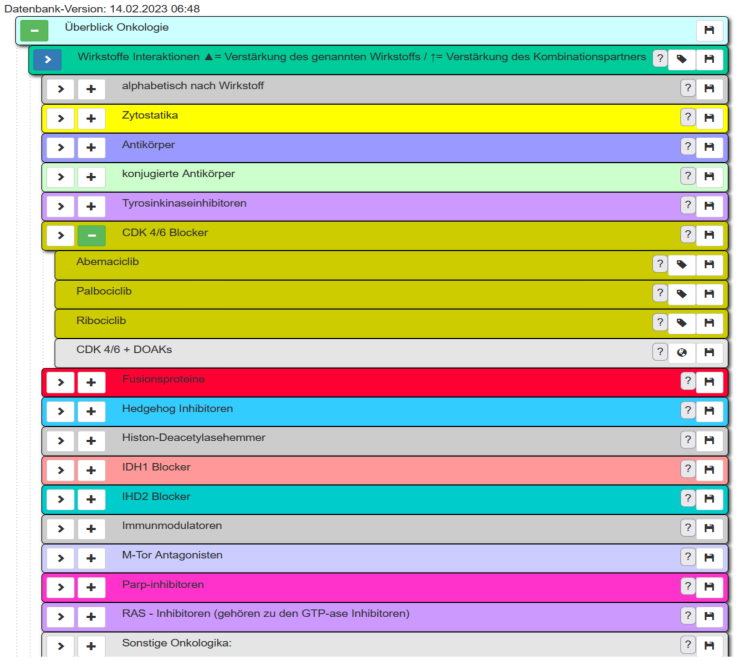

Figure A10. Case-specific supportive information panel displaying relevant drug interactions and contraindications.

Figure A11. Detailed interaction contraindications display for specific medications, highlighting potential adverse effects.

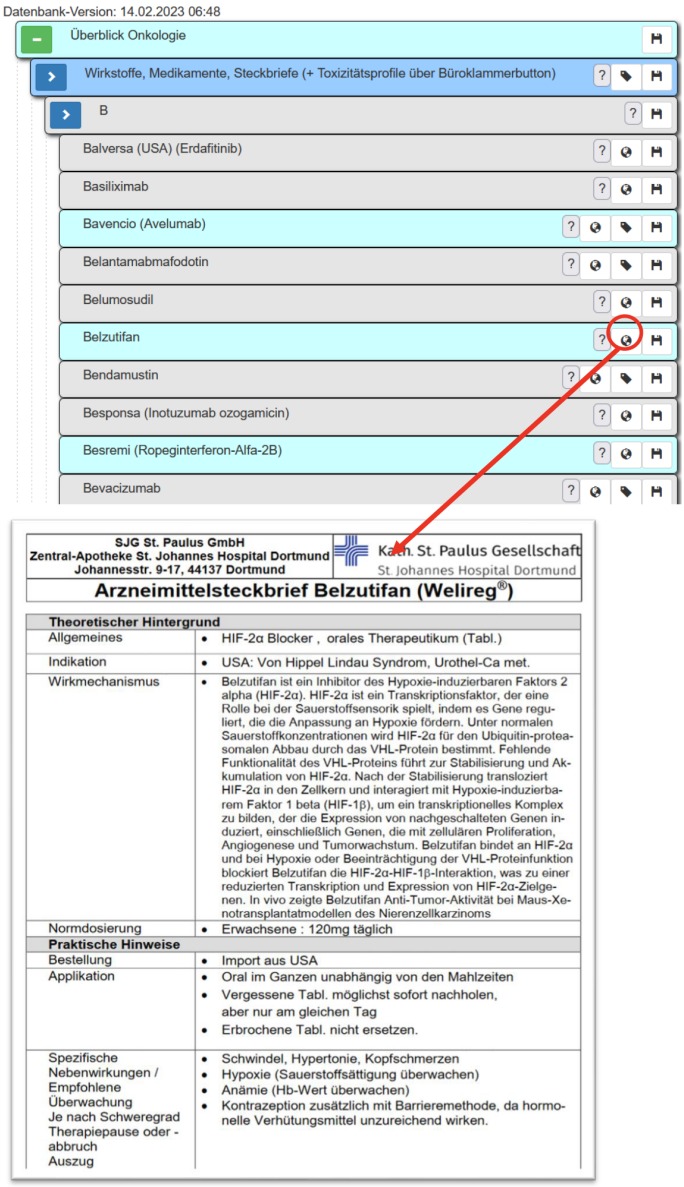

Figure A12. Comprehensive medication information summary providing practical details about each pharmaceutical agent accessible to clinical practitioners.

Author Contributions: Conceptualisation S.B. and M.T; methodology, S.B.; software, S.B; validation, S.B.; investigation, S.B., M.T. and N.B.; resources, S.B; data curation, S.B; writing—original draft preparation, M.T , S.B; writing—review and editing, N.B; visualisation, S.B; supervision, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Edge Hill University (protocol code ETH2425-0274, 7th July 2025).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is not publicly available due to commercial constraints.

Acknowledgments: The authors would like to thank the staff at OncoProAI for their support throughout this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rudie, J.D., Rauschecker, A.M., Bryan, R.N., Davatzikos, C. and Mohan, S., 2019. Emerging applications of artificial intelligence in neuro-oncology. Radiology, 290(3), pp.607-618.
2. Hamamoto, R., Suvarna, K., Yamada, M., Kobayashi, K., Shinkai, N., Miyake, M., Takahashi, M., Jinnai, S., Shimoyama, R., Sakai, A. and Takasawa, K., 2020. Application of artificial intelligence technology in oncology: Towards the establishment of precision medicine. Cancers, 12(12), p.3532.
3. Shimizu, Hideyuki and Nakayama, Keiichi I, Artificial intelligence in oncology, Cancer science, volume 111, 1452–1460, 2020.
4. Aftab, M., Mehmood, F., Zhang, C., et al. (2025). AI in Oncology: Transforming Cancer Detection through Machine Learning and Deep Learning Applications. arXiv preprint arXiv:2501.15489.
5. Madabhushi, A. (2016). Image analysis and machine learning in digital pathology: challenges and opportunities. Elsevier.
6. Collins, C., Baker, S., Brown, J., et al. (2025). Text mining for contexts and relationships in cancer genomics literature. Bioinformatics, 40(1), btae021.
7. Tran, C., Li, S., Wang, H., et al. (2025). In-Context Learning for Label-Efficient Cancer Image Classification in Oncology. ResearchGate.
8. Elucidata. (2024). Transformative Trends: Manual and Automated Curation Approaches in Biopharma Research. Retrieved from https://www.elucidata.io/blog/manual-and-automated-curation-approaches-in-biopharma-research
9. Sarrouh, M., Torkar, R., Paterno, F. (2024). Ontology-Based AI Design Patterns and Constraints in Cancer Registry Data Validation. Cancers, 15(24), 5812.
10. Luschi, A., Monreale, A., Panigutti, C., et al. (2023). Semantic Ontologies for Complex Healthcare Structures: A Scoping Review. FLORE.
11. Corbucci, L., Monreale, A., Panigutti, C., et al. (2023). Semantic Enrichment of Explanations of AI Models for Healthcare. ResearchGate.
12. OncoProAI. *OncoProAI website*. Available at: https://www.oncoproai.com. Accessed 16 Aug 2025.
13. Sameer Bhat. *oncology-kg-ai* (GitHub repository). Available at: https://github.com/SameerBhat/oncology-kg-ai. Accessed 16 Aug 2025.
14. Kann, B.H., Thompson, R., Thomas Jr, C.R., Dicker, A. and Aneja, S., 2019. Artificial intelligence in oncology: current applications and future directions.Oncology,33(2), pp.46-53.
15. Abdel Razek, A.A.K., Alksas, A., Shehata, M., AbdelKhalek, A., Abdel Baky, K., El-Baz, A. and Helmy, E., 2021. Clinical applications of artificial intelligence and radiomics in neuro-oncology imaging. Insights into imaging, 12(1), p.152.
16. Farina, E., Nabhen, J.J., Dacoregio, M.I., Batalini, F. and Moraes, F.Y., 2022. An overview of artificial intelligence in oncology. Future science OA, 8(4), p.FSO78
17. Vicini, S., Bortolotto, C., Rengo, M., Ballerini, D., Bellini, D., Carbone, I., Preda, L., Laghi, A., Coppola, F. and Faggioni, L., 2022. A narrative review on current imaging applications of artificial intelligence and radiomics in oncology: Focus on the three most common cancers. La radiologia medica, 127(8), pp.819-836.
18. Alsharif, F., 2024. Artificial Intelligence in Oncology: Applications, Challenges and Future Frontiers. International Journal of Pharmaceutical Investigation, 14(3)
19. Matsui, Y., Ueda, D., Fujita, S., Fushimi, Y., Tsuboyama, T., Kamagata, K., Ito, R., Yanagawa, M., Yamada, A., Kawamura, M. and Nakaura, T., 2025. Applications of artificial intelligence in interventional oncology: An up-to-date review of the literature. Japanese Journal of Radiology, 43(2), pp.164-176.
20. Li, C., Zhang, Y., Weng, Y., Wang, B. and Li, Z., 2023. Natural language processing applications for computer-aided diagnosis in oncology. Diagnostics, 13(2), p.286.
21. Silva, M.C., Eugénio, P., Faria, D. and Pesquita, C., 2022. Ontologies and knowledge graphs in oncology research. Cancers, 14(8), p.1906.
22. Nicholson, N., Giusti, F. and Martos, C., 2023. Ontology-based AI design patterns and constraints in cancer registry data validation. Cancers, 15(24), p.5812.
23. Ritchie, J.B., Frey, L.J., Lamy, J.B., Bellcross, C., Morrison, H., Schiffman, J.D. and Welch, B.M., 2022. Automated clinical practice guideline recommendations for hereditary cancer risk using chatbots and ontologies: system description. JMIR cancer, 8(1), p.e29289.
24. Liao, J., Li, X., Gan, Y., Han, S., Rong, P., Wang, W., Li, W. and Zhou, L., 2023. Artificial intelligence assists precision medicine in cancer treatment. Frontiers in oncology, 12, p.998222.
25. SNOMED International. (2024). SNOMED CT International Edition. Available online: https://www.snomed.org/snomed-ct/international-edition (accessed on 14 August 2025).
26. The Gene Ontology resource: 20 years and still GOing strong. Nucleic Acids Research, 47(D1), D330-D338. https://doi.org/10.1093/nar/gky1055

27. Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C. and Zhao, L., 2025. GRAG: Graph Retrieval-Augmented Generation. arXiv preprint arXiv:2405.16506.
28. Westbury, S. K., E. Turro, D. Greene, C. Lentaigne, A. M. Kelly, T. K. Bariana, I. Simeoni, et al. 2015. "Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders". Genome Medicine 7 (1): 36. doi:10.1186/s13073-015-0151-5.