

An integrated diffusion-weighted imaging processing and interpretation platform for MR-guided radiotherapy

Yunxiang Li^{1,2}, Yan Dai¹, Yen-Peng Liao¹, Jie Deng¹, Jill B. De Vis¹, and You Zhang^{1,*}

¹*Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, USA*

²*Department of Radiation Oncology, University of California San Francisco, San Francisco, California, USA*

**Corresponding author: You Zhang (you.zhang@utsouthwestern.edu)*

Abstract

Background: Magnetic resonance imaging-guided linear accelerators (MR-Linacs) allow diffusion-weighted imaging (DWI) to be acquired at every treatment fraction, but converting these low-signal-to-noise-ratio acquisitions into clinical decisions requires both reliable quantitative processing and an interpretation that reconciles a scattered and often contradictory literature.

Purpose: To describe and evaluate an integrated, web-based platform that carries raw MR-Linac DWI to a structured, literature-grounded clinical interpretation, and to assess its retrieval-augmented generation (RAG) interpretation module by independent expert rating.

Methods: The platform couples a deep-learning processing pipeline, comprising distortion correction, denoising, and intravoxel incoherent motion (IVIM)/apparent diffusion coefficient (ADC) fitting, with longitudinal region-of-interest analysis and a RAG interpretation agent. The agent reasons over a two-layer knowledge base of curated publications (a structured catalog index plus line-indexed full text), delegates arithmetic to deterministic tools, and is designed to trace each statement to a source document, section, and line range. One medical physicist and one physician independently rated the agent’s reports for nine longitudinal glioblastoma cases on a 1–5 scale across three metrics: clinical-reasoning soundness, literature-citation quality, and overall clinical utility.

Results: Across 54 ratings, the pooled mean was 4.65 ± 0.80 , with 93% of ratings ≥ 4 ; Metric means were 4.6 (reasoning), 4.5 (citation), and 4.8 (utility), and raters agreed within one point on 85% of paired ratings.

Conclusions: A single platform can integrate MR-Linac DWI post-processing with traceable, expert-evaluated clinical interpretation, while highlighting the safeguards needed to verify LLM-generated reasoning in radiation oncology.

Keywords: diffusion-weighted imaging; intravoxel incoherent motion; MR-guided radiotherapy; large language models; retrieval-augmented generation; glioblastoma

1 Introduction

Magnetic resonance imaging-guided radiotherapy (MRgRT) delivered on a magnetic resonance imaging-guided linear accelerator (MR-Linac) has changed not only how radiation is delivered but also what can be observed during a treatment course. Because the patient is imaged on the treatment table at each fraction, functional sequences such as diffusion-weighted imaging (DWI) can be concurrently acquired with anatomical sequences, rather than at the one or two isolated diagnostic time points typical of conventional follow-up [1, 2]. This per-fraction capability provides a longitudinal sampling density that diagnostic MRI cannot match and allows the tumor to be monitored throughout the treatment course. DWI and its intravoxel incoherent motion (IVIM) modeling are attractive functional biomarkers in this setting: the apparent diffusion coefficient (ADC) reflects tissue cellularity, while the IVIM model separates the signal into true tissue diffusion (D_t), pseudo-diffusion (D_p), and the perfusion fraction (f_p), thereby disentangling cellular from microvascular contributions [3–5]. Sampling these parameters fraction by fraction offers, in principle, an early and biologically specific measure of how a tumor is responding to therapy. In this study, we described a platform specifically developed for this acquisition regime.

Both the opportunity and the difficulty are greatest in glioblastoma (GBM) management. Despite maximal chemoradiation, median survival remains approximately 15 months [6], and post-treatment management is confounded by pseudoprogression, in which treatment-induced blood-brain barrier disruption and inflammation mimic true tumor recurrence on conventional contrast-enhanced MRI [7, 8]. This phenomenon is common in a substantial

fraction of patients overall and at especially high rates in MGMT-methylated tumors [9, 10]. Morphological response criteria cannot reliably separate these two biologically distinct processes [11], and confirmation typically requires months of follow-up, during which patients with genuine progression may be undertreated and miss the survival benefit of timely second-line therapy [12]. In a meta-analysis of 24 studies (900 patients), DWI distinguished true progression from pseudoprogression with a pooled sensitivity of 0.88 and specificity of 0.85 [13]. ADC alone, however, conflates diffusion and perfusion into a single quantity, limiting its ability to distinguish the complex biological processes in GBM, where VEGF-driven angiogenesis [14] coexists with a heterogeneous microenvironment of viable tumor, necrosis, and an angiogenic rim [15]. The multi-compartment IVIM model is better matched to this biology and has shown diagnostic and prognostic value in glioma, including tumor grading and treatment-response assessment [16–19].

Realizing the potential of DWI on an MR-Linac for response monitoring requires solving two separate problems. The first is technical. MR-Linac DWI is acquired at lower SNR than diagnostic DWI due to machine constraints, and clinically pragmatic single-shot echo-planar acquisition introduces geometric distortions; both effects destabilize quantitative fitting, especially of the perfusion-sensitive D_p and f_p parameters [5, 20]. We previously developed a dedicated processing chain for this low-SNR regime: landmark-matched B-spline implicit neural representation (INR) for distortion correction [21], band-limited INR for denoising [22], and an INR-based IVIM parameter estimator that recovers reproducible longitudinal parameter maps [23]. The second problem is interpretive and has received comparatively less attention. Even when reliable parameters are available, turning D_p , D_t , and f_p trajectories into a clinical assessment demands synthesizing evidence that is scattered across previous literature and often contradictory, because reported thresholds and parameter-outcome relationships vary with field strength, b -value protocol, vendor, tumor location, and treatment regimen [20, 24]. The three parameters may also move in opposite directions in the same patient (for example, D_t decreasing while f_p rises), and no vali-

dated framework exists to resolve such multi-parameter discordance [5]. On many occasions, the barrier to adopting IVIM for treatment monitoring outside specialized centers is this knowledge-synthesis burden rather than a shortage of data.

Software support exists for part but not the whole of the workflow. For instance, general DICOM-management tools can streamline data organization and conversion for downstream analysis [25], and several quantitative DWI methods have been validated on the MR-Linac [26, 27]. However, an integrated path progressing from raw MR-Linac DWI to an interpretation that a clinician can act on and audit is still missing. On the interpretation side, large language models (LLMs) coupled with retrieval-augmented generation (RAG) offer a natural technical approach. Rather than relying on a model’s parametric memory, which risks fluent but unsupported conclusions, RAG requires the model to retrieve published evidence before it reasons [28, 29], mirroring how a clinician consults the literature before forming an assessment. Agentic LLM systems built on this principle have begun to deliver strong diagnostic performance in other medical domains while keeping their reasoning traceable: DeepRare, an agentic rare-disease system evaluated across nine datasets spanning 2,919 diseases, achieved an average Recall@1 of 57.18% in human-phenotype-ontology-based diagnostic tasks, outperforming the next best method by 23.79%, and expert review confirmed the validity of 95.4% of its reasoning chains [30].

In this work, we develop and evaluate an integrated, web-based platform that aims to address both the technical and interpretive challenges of DWI’s clinical application. The platform carries MR-Linac DWI from raw DICOM through distortion correction, denoising, IVIM/ADC fitting, and registration, to longitudinal, region-of-interest (ROI)-based parameter trajectories, and then submits those trajectories to a RAG-based clinical interpretation agent that returns a structured report in which each statement is anchored on a specific passage of the source literature. We first present the system architecture, its data connectivity, the deep-learning processing components, and the design of the interpretation agent. We then report an evaluation of the agent’s reports on nine longitudinal GBM cases, in-

independently rated by a medical physicist and a physician across reasoning, citation, and utility metrics. Rather than proposing a new imaging algorithm, this work demonstrates that processing and interpretation can be unified in one auditable clinical tool, and reports where expert reviewers found the tool trustworthy and where it needs further improvement.

2 Methods

2.1 System overview and architecture

The platform is implemented as a single-page web application backed by a Python server. The backend uses the Flask framework with Flask-SocketIO for bidirectional, real-time communication, so that long-running processing jobs report progress to the browser as they execute [31]. The front end is an HTML/JavaScript client. All image processing runs locally on institutional hardware with GPU acceleration, and no image data leaves the local environment. Only the derived numerical summaries used by the interpretation agent are sent to the HIPAA-compliant language-model deployment described in Section 2.5. The deep-learning components are built on PyTorch [32]; medical-image input/output and geometry are handled by pydicom [33], nibabel, and SimpleITK [34], with additional transform modules from MONAI [35]. Table 1 summarizes the principal software components and their roles.

The application is organized around four user-facing capabilities, exposed from a common landing page: an ADC & IVIM processing pipeline; an MRI analysis module for longitudinal, ROI-based quantification and visualization; a clinical-interpretation agent; and a DICOM export utility that returns processed maps to clinical systems. Internally, processing is expressed as a numbered sequence of steps that share a common on-disk data hierarchy keyed by patient identifier and study date, so that intermediate results are inspectable and any step can be re-run independently. The overall data flow, from raw DICOM through quantitative maps to a traceable interpretation, is shown in Figure 1. Representative views of the

Table 1: Principal software components of the platform and their roles. Versions reflect the validated deployment configuration.

Component	Version	Role in the platform
Python	3.10	Implementation language
Flask / Flask-SocketIO	2.3 / 5.5	Web server and real-time progress streaming
PyTorch	2.6 (CUDA 12.6)	Deep-learning models (distortion, denoising, fitting)
pydicom	3.0.1	DICOM reading and metadata extraction
nibabel	5.3.2	NIfTI input/output
SimpleITK	2.4.1	Reorientation, resampling, geometry
MONAI	1.4.0	Medical-image transforms
NumPy / SciPy	1.26 / 1.15	Numerical computation and optimization
LangChain / LangGraph	n/a	Agent orchestration and tool calling
Azure OpenAI (GPT-5.1)	n/a	Large-language-model reasoning (HIPAA-compliant)

processing, analysis, and interpretation interfaces are presented alongside the corresponding modules below. A defining feature of the platform is that these stages form a single application rather than separate tools: a study can be carried from DICOM import to a clinical report without leaving the browser or writing any code.

2.2 Data connectivity and DICOM handling

Clinical interoperability is handled entirely through DICOM. Raw exports are first organized into a deterministic hierarchy (`patient/date/series`). Anatomical sequences (T2-weighted, FLAIR, T1-weighted) are converted to NIfTI directly; DWI series are parsed for their diffusion b -value and written as one volume per b -value, rendering a multi- b -value stack required for IVIM/ADC fitting. During conversion, slices are sorted by their physical position, and volumes are reoriented to a consistent voxel grid, preserving the affine geometry needed for later registration. Segmentation objects (DICOM-SEG and RT-STRUCT) are converted to binary mask volumes that are spatially matched to their referenced anatomical image through the DICOM frame-of-reference metadata, which makes clinical target volumes (e.g., GTV, CTV) and organs at risk directly available as ROIs for analysis. After processing,

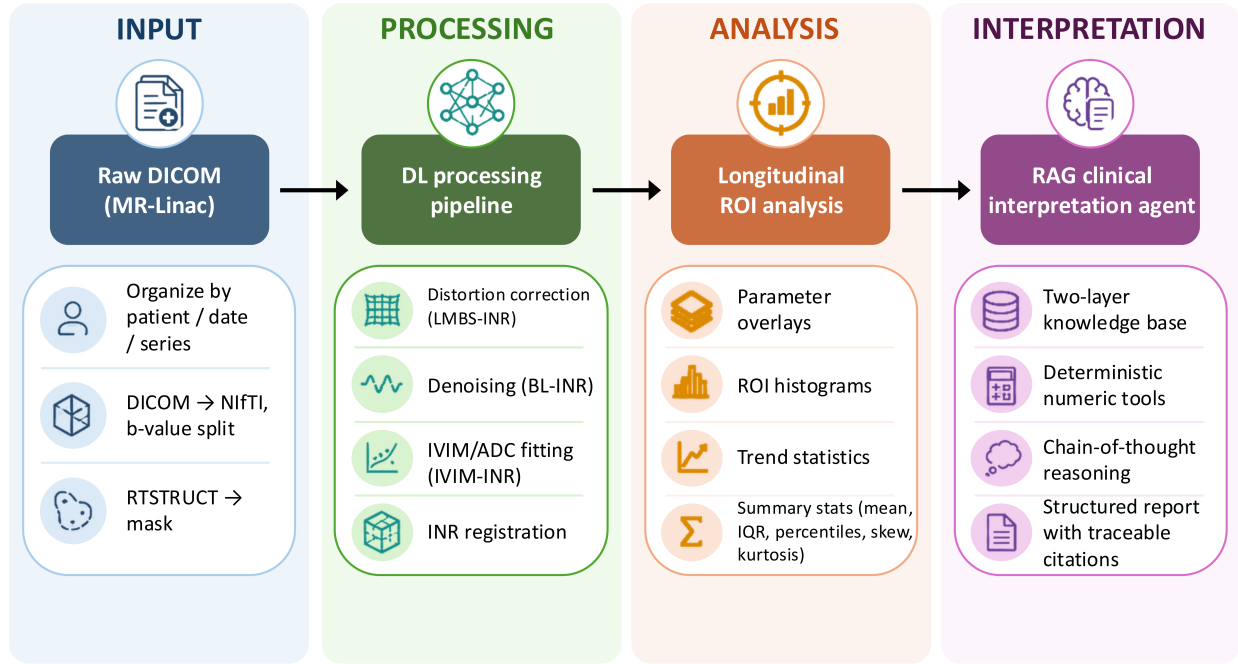


Figure 1: End-to-end architecture and data flow of the platform. Raw MR-Linac DICOM data are organized and converted to NIfTI (with b -value separation for DWI and mask extraction for segmentations), processed through a deep-learning pipeline into IVIM/ADC parameter maps, analyzed longitudinally over regions of interest, and finally interpreted by a retrieval-augmented generation (RAG) agent that is designed to trace evidence-supported statements to line-indexed source literature passages. Processed maps can be exported back to DICOM.

parameter maps can be written back to DICOM, preserving patient and study metadata so that quantitative outputs can be reviewed alongside the planning images in clinical systems.

2.3 Image-processing pipeline

The core ADC & IVIM pipeline executes as an ordered sequence of steps (Figure 2). Quantitative processing is performed at the native DWI resolution, with only the final parameter maps resampled to match the higher-resolution anatomical images. This substantially reduces fitting time compared with fitting upsampled DWI volumes while ensuring that parameter estimation is based on the originally acquired signal. The deep-learning components, summarized below, have been described and validated in detail elsewhere and are used here as integrated modules.

Anatomical downsampling. The high-resolution anatomical reference is resampled to the DWI grid to serve as a structural prior for distortion correction and denoising.

Distortion correction (LMBS-INR). Geometric distortion from B_0 inhomogeneity is corrected by deformably registering the DWI volume to the anatomical reference. Our approach combines cross-modality landmark matching from a modality-invariant image-matching model [36] with a B-spline implicit neural representation of the deformation field, overcoming the local optima of conventional intensity-only optimization while enforcing smooth, anatomically plausible deformations. On brain GBM data, the method achieved a mean Dice coefficient of 0.919, exceeding established diffeomorphic and learning-based baselines, which ensures that ROI-based parameter measurements correspond to consistent anatomy across fractions [21].

Denoising (BL-INR). The low SNR of MR-Linac DWI is addressed with a self-supervised, band-limited implicit neural representation that restricts the network’s learnable frequency content to suppress high-frequency noise, while using the multi- b -value

signal-decay relationship and cross-modality structural consistency as physics-informed constraints. The method requires no paired training data and, on clinical DWI, reduced the noise standard deviation to 40% of its original level while preserving the signal mean, achieving a structural similarity index of 0.933 [22].

IVIM/ADC parameter fitting (IVIM-INR). Voxel-wise IVIM fitting is highly sensitive to noise, particularly for f_p and D_p . Treating each voxel independently ignores spatial continuity and yields noisy maps with poor longitudinal consistency. We instead frame fitting as spatial-function learning: a periodic-activation (SIREN) implicit neural representation [37] models each parameter map as a continuous function of position, so that structural context and inter-voxel correlation are leveraged intrinsically [23]. The network fits the IVIM bi-exponential model (Eq. 1),

$$\frac{S(b)}{S_0} = f_p e^{-bD_p} + (1 - f_p) e^{-bD_t}, \quad (1)$$

and, on the low- b /high- b partition, the mono-exponential ADC model (Eq. 2),

$$\frac{S(b)}{S_0} = e^{-b\text{ADC}}, \quad (2)$$

with physically motivated bounds ($D_p \in [3.5 \times 10^{-3}, 0.1]$, $D_t \in [1 \times 10^{-5}, 3 \times 10^{-3}]$, $f_p \in [0, 0.99]$, diffusivities in mm^2/s). Across multi-time-point MR-Linac data, this approach improved longitudinal reproducibility for all three IVIM parameters over least-squares and recent deep-learning baselines (for example, raising the f_p intraclass correlation coefficient from 0.29 to 0.56), which is essential for detecting small but clinically meaningful changes over the course of treatment. [23].

Registration and resampling. The pipeline includes INR-based rigid and deformable registration. After fitting, native-resolution DWI and parameter maps are upsampled to the

anatomical reference, and inter-fractional deformable registration brings all fractions into a common anatomical space, so that a fixed ROI can be propagated across the treatment course for longitudinal comparison. For these inter-fractional alignments, a mask-based rigid pre-alignment (excluding tumor-bearing regions so that disease change does not bias the transform) is performed before the deformable refinement.

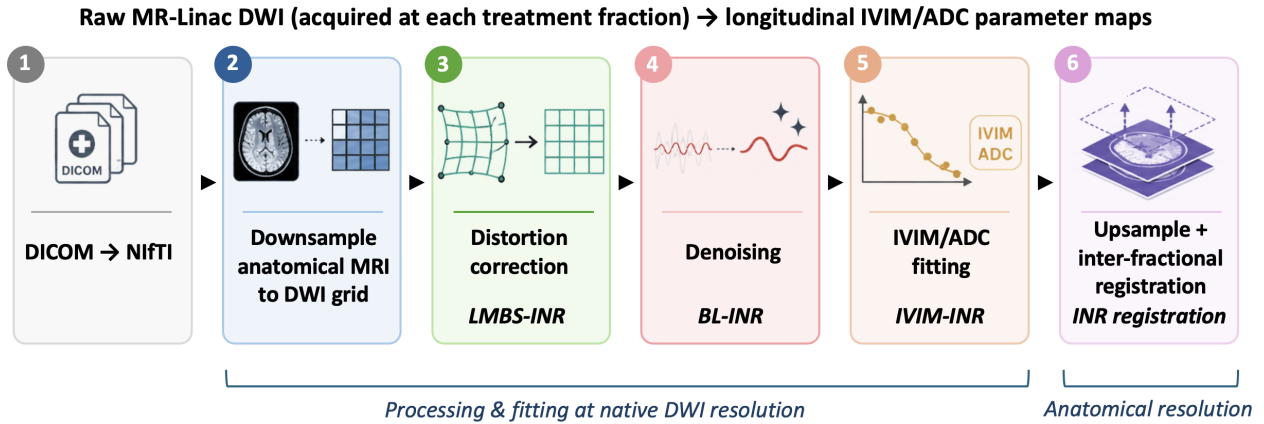
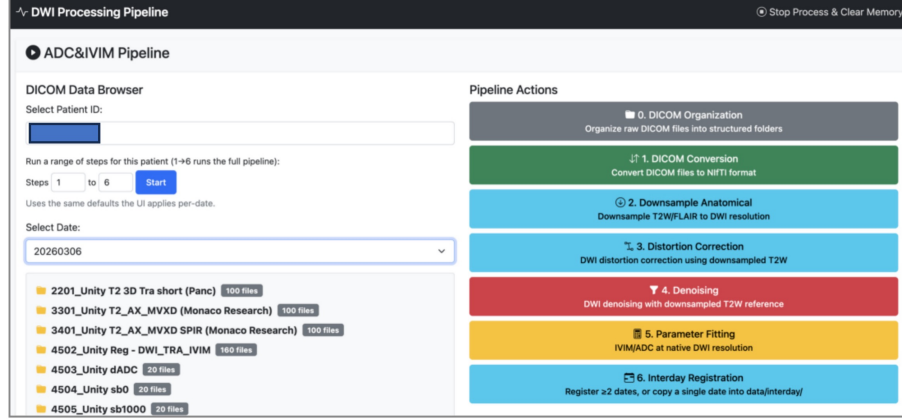


Figure 2: The DWI/IVIM processing pipeline. Raw multi- b -value MR-Linac DWI, acquired at each treatment fraction, is processed through six steps; quantitative fitting is performed at native DWI resolution (steps 2–5) and only the final maps are resampled to anatomical resolution before inter-fractional registration (step 6).

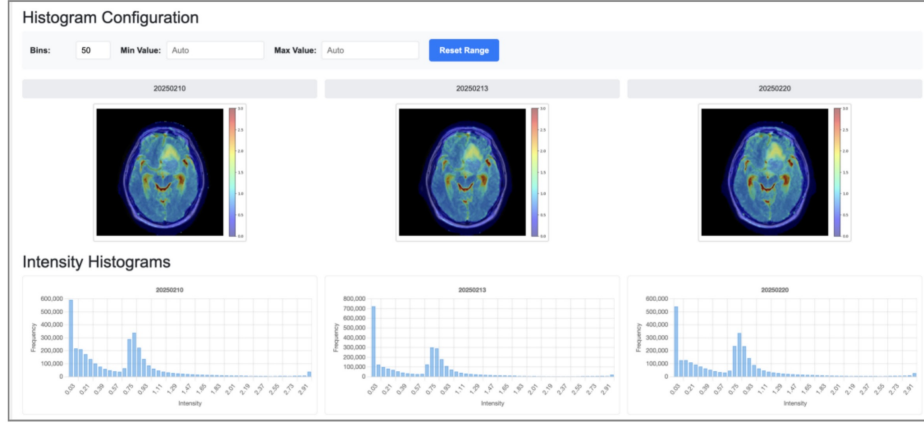
2.4 Longitudinal analysis and visualization

The analysis module turns registered parameter maps into quantitative trajectories that drive interpretation. For a user-selected ROI imported from RT-STRUCT, the module overlays each parameter map on the anatomical image at every time point, computes ROI-based intensity histograms (Figure 3b), and reports a panel of longitudinal summary statistics (mean, median, interquartile range, selected percentiles, skewness, and kurtosis) as functions of fraction date (Figure 3c). A fixed first-fraction ROI can be applied to all time points, or date-specific ROIs can be used. The resulting per-parameter trajectories and their percentage changes constitute the structured numerical input to the interpretation agent, ensuring that the agent reasons over the same auditable quantities a physicist/physician would inspect manually.

(a) ADC & IVIM processing pipeline: DICOM data browser (left) and numbered processing steps (right)



(b) Analysis tool: per-fraction parameter-map overlays and ROI intensity histograms (brain GBM case)



(c) Longitudinal trend statistics (mean, median, IQR, percentiles, skewness, kurtosis) across fractions



Figure 3: The web platform interface, with a brain GBM case shown. (a) The ADC & IVIM processing pipeline, with the DICOM data browser (patient/date/series, left) and the numbered processing steps (right). (b) The analysis tool, showing per-fraction parameter-map overlays on the anatomical image and the corresponding ROI intensity histograms. (c) Longitudinal trend statistics (mean, median, interquartile range, percentiles, skewness, and kurtosis) computed within a fixed ROI across treatment fractions.

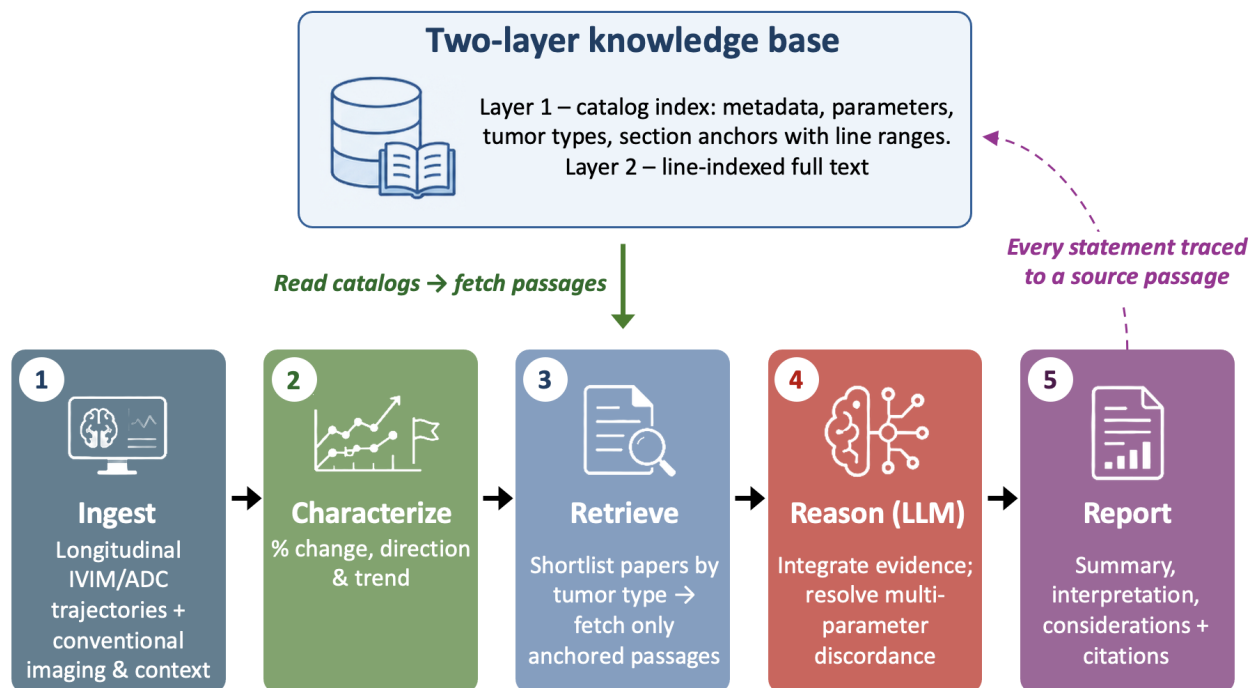


Figure 4: The retrieval-augmented interpretation agent. Longitudinal IVIM/ADC trajectories, together with the conventional imaging context, are ingested (stage 1) and characterized (stage 2). Catalog-guided retrieval then shortlists tumor-type- and parameter-matched publications from the two-layer knowledge base and fetches only the anchored passages (stage 3). The LLM integrates the evidence and resolves multi-parameter discordance (stage 4), and a structured report is produced in which evidence-supported statements are traced back to a source document, section, and line range (stage 5).

2.5 Retrieval-augmented clinical-interpretation agent

The interpretation agent is the platform’s key component. It is built with the LangChain/LangGraph orchestration framework [38] and uses a GPT-5-class LLM (GPT-5.1) served through our institution’s HIPAA-compliant Microsoft Azure OpenAI deployment. Three design choices (Figure 4) constrain the agent to ground its output in the retrieved literature rather than generate free-form text: a structured knowledge base, two-stage retrieval, and tool-assisted calculations with passage-level citation.

Two-layer knowledge base. At the time of this study, the evidence base comprised 28 peer-reviewed publications on DWI/IVIM (and related quantitative MRI) in oncology and MRgRT, curated to span the tumor types and parameters relevant to the platform (Table 2). Each publication is stored in two layers. Layer 1 is a structured catalog: a machine-readable record of the paper’s metadata (title, authors, journal, year, DOI), the tumor types and treatments studied, the IVIM/ADC and other parameters it reports, and a section-level content map in which every section is summarized and anchored to an explicit line range in the source text. Layer 2 is the paper’s full text, stored as line-indexed plain text so that any anchored passage can be retrieved word by word on demand. This mirrors how humans usually read—scanning a table of contents before turning to the relevant section. The strategy also allows the agent to reason over a concise overview of the entire corpus before accessing any full-text documents. New papers can be continuously added through the interface by uploading a PDF (Figure 5b), which is parsed into both layers automatically.

Two-stage retrieval and tool-augmented reasoning. The user selects the target ROIs and launches the analysis from the interpretation interface (Figure 5a). Given a patient’s longitudinal parameter trajectories, the agent then proceeds in stages. It first characterizes the parameter changes (percentage change, direction, and trend for each of ADC, D_p , D_t , and f_p), delegating these computations to deterministic Python tools rather than perform-

ing calculations itself, which removes a well-known LLM failure mode. It then performs catalog-guided retrieval: reading the Layer-1 catalogs, it shortlists publications whose tumor type, parameter scope, and reported change directions match the patient’s pattern, explicitly filtering for tumor-type compatibility (for example, restricting to GBM/brain studies for a brain case), and retrieves only the specific Layer-2 passages indicated by the matching catalog anchors. Finally, it integrates the tool-computed findings, user-provided prompt, and retrieved passages through chain-of-thought reasoning to identify multi-parameter discordances and resolve them by reference to biological mechanisms supported by the retrieved evidence.

Traceable report generation. The agent returns a structured report with a parameter summary, an evidence-based interpretation, and clinical considerations. Citation is implemented through explicit passage tracking rather than relying on the LLM’s recollection: each retrieved Layer-2 passage carries its source document, section, and line range, which are propagated to any statement derived from it, so that a reader can verify each claim against the exact lines of the cited paper (Figure 4). This traceability makes the system auditable and is the key property that our expert evaluation was designed to assess.

Table 2: Composition of the retrieval knowledge base (28 publications). Counts of tumor types and reported parameters exceed 28 because individual papers may span multiple categories.

Attribute	Categories (examples)	Coverage
Tumor type	Glioblastoma / glioma / brain metastases	predominant
	Pancreatic adenocarcinoma; head and neck; prostate	secondary
Parameters	ADC	28 papers
	IVIM (f_p , D_t , D_p)	13 papers
Setting	MR-Linac / MR-guided radiotherapy; chemoradiation	
Evidence	Original studies, systematic reviews/meta-analyses, position statements	mixed

(a) One-click AI clinical interpretation: target-ROI selection and saved analysis sessions

(b) Retrieval knowledge base (literature library): per-paper tumor types and parameters;

Figure 5: The clinical-interpretation interface. (a) The one-click AI clinical-interpretation view, where the user selects the target ROIs (e.g., GTV/CTV/PTV imported from RT-STRUCT) and launches an analysis. Prior analysis sessions are retained for review. (b) The retrieval knowledge base (“literature library”), which lists each indexed publication with its tumor types and reported parameters and accepts new papers by PDF upload.

2.6 Expert evaluation of the interpretation agent

We evaluated the clinical quality of the agent’s reports on nine longitudinal GBM cases whose MR-Linac studies were processed end-to-end through the platform. For each case, longitudinal DWI was processed into IVIM/ADC trajectories within the clinical gross tumor volume (GTV), and the agent generated a structured interpretation report. Two domain experts (one medical physicist and one physician with neuro-oncology imaging expertise) independently reviewed all nine reports.

Each report was rated on a 1–5 Likert scale along three pre-specified metrics defined in a shared rubric: (i) clinical-reasoning soundness: whether the reasoning is rigorous and correctly analyzes concordant and discordant multi-parameter relationships; (ii) literature-citation quality: whether citations are relevant, accurate, and traceable to a specific section and line range matching the case; and (iii) overall clinical utility: whether the report would aid clinical decision-making. Anchors were defined for each level (5 = excellent; 4 = good; 3 = acceptable; 2 = deficient; 1 = poor), and raters could add free-text comments. Ratings were collected independently, with no communication between the reviewers.

Given the small pilot-scale cohort, the analysis is descriptive. We report per-metric mean $\pm$ standard deviation for each rater and pooled across raters, the proportion of ratings at or above each scale anchor, and, as a simple inter-rater agreement check, the rate of exact and within-one-point agreement and the mean absolute difference between raters across the 27 paired score cells (9 cases $\times$ 3 metrics). We avoid inferential agreement statistics (e.g., weighted kappa or intraclass correlation), which can be unstable and easily over-interpreted at this sample size, particularly when one rater’s scores exhibit near-zero variance. Free-text comments were reviewed qualitatively to identify recurring strengths and failure modes.

3 Results

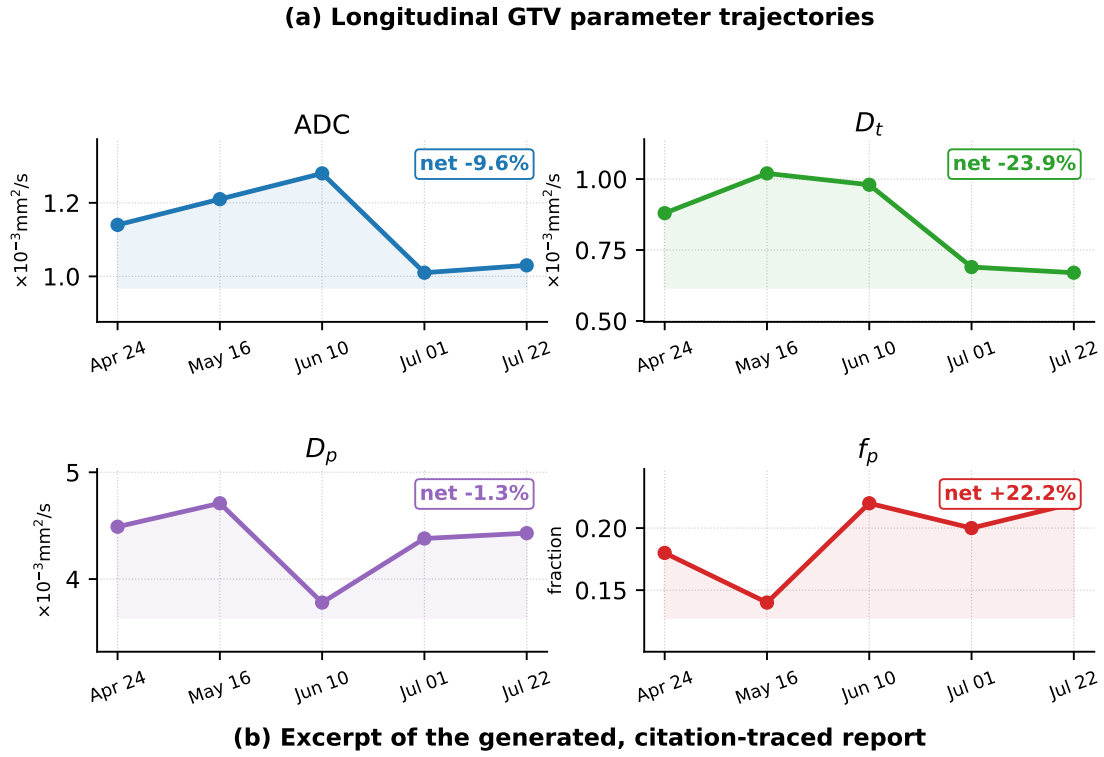
3.1 Output of a representative case

The platform processed MR-Linac studies end-to-end, from raw DICOM to registered, longitudinal IVIM/ADC maps and a clinical interpretation, within the local environment and without manual scripting. Native-resolution fitting kept per-study processing tractable, and the analysis module produced ROI-based parameter trajectories suitable both for direct review and for use as agent input.

A representative case illustrates the output (Figure 6). For a GBM patient imaged at five longitudinal MR-Linac time points spanning the treatment course, the GTV delineated at the first time point was propagated to all subsequent scans. From the first to the last time point, ADC, D_t , D_p , and f_p changed by -9.6% , -23.9% , -1.3% , and $+22.2\%$, respectively, with ADC and D_t showing an early increase followed by a marked late decline. The agent’s report first summarized these longitudinal changes and then interpreted them in phases. It attributed the early increase in diffusivity to an expected treatment response, consistent with reduced cellularity and expansion of the extracellular space, while interpreting the subsequent decline in D_t with preserved or rising f_p as a pattern warranting attention for possible tumor regrowth. The report also explicitly highlighted the discordance between decreasing diffusivity and stable or increasing perfusion. Each interpretive statement carried a citation to a specific section and line range of a matched literature source. For example, the treatment-response reading was anchored to longitudinal-ADC findings during chemoradiation, and the perfusion analysis to MRgRT brain-tumor literature [13, 39–41]. Both reviewers rated this case 5/5/5 for the three metrics.

3.2 Expert evaluation results

Across the 54 ratings (9 cases $\times$ 3 metrics $\times$ 2 raters), the pooled mean score was 4.65 ± 0.80 , and 93% of ratings were ≥ 4 (50/54), with 96% ≥ 3 (52/54). Pooled metric means were



Agent report (excerpt)

Summary. Within the fixed GTV, ADC and D_t rose early (Apr-Jun) and then declined; D_t fell 23.9% overall while f_p rose 22.2% and D_p stayed essentially stable.

Interpretation. The early diffusivity increase is consistent with a treatment effect (reduced cellularity, enlarged extracellular space); the later D_t decline with preserved/rising f_p is a discordant pattern that warrants attention for possible regrowth.

Cited: Moore-Palhares et al. (2025), Results, Lines 142-171

→ traceable to the source passage

Figure 6: A representative longitudinal GBM case. (a) GTV-mean trajectories of ADC, D_t , D_p , and f_p across five longitudinal time points, showing an early diffusivity rise followed by a late decline with preserved/rising perfusion fraction. (b) Excerpt of the agent's report, in which the interpretation is anchored to specific lines of matched source literature. This case received the maximum rating on all three dimensions from both reviewers.

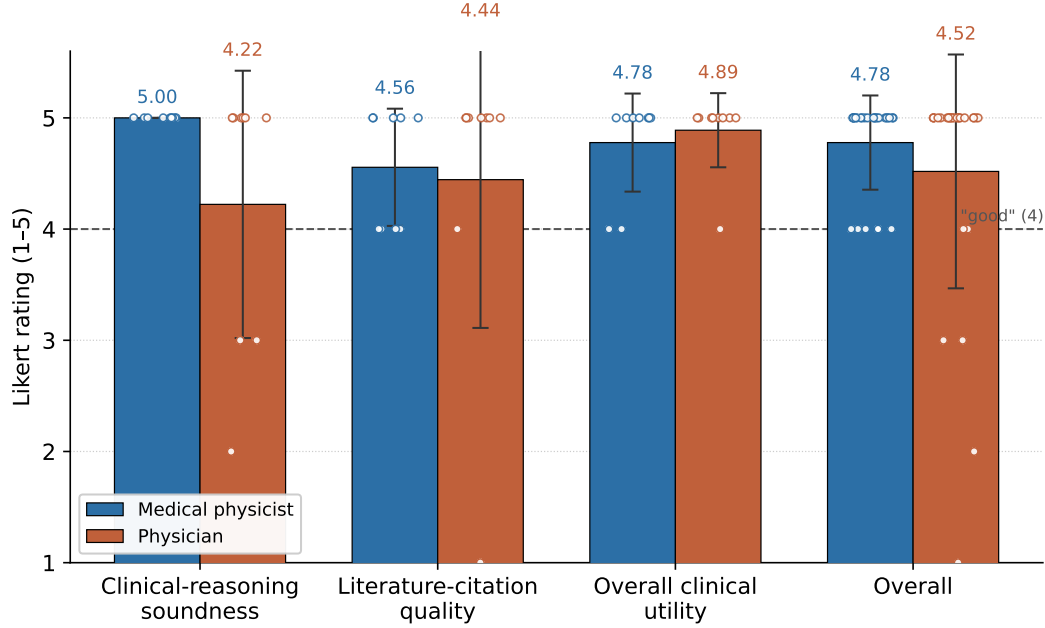


Figure 7: Independent expert ratings of agent-generated reports by metric and rater (mean $\pm$ SD; 1–5 Likert). The dashed line marks the “good” scale anchor (4). Overall clinical utility received the highest and most consistent ratings from both reviewers. The physician’s ratings showed greater variability, driven by a few atypical cases (reasoning scores of 2, 3, and 3 in three cases and one citation score of 1), whereas the physicist assigned a reasoning of 5 to all cases.

4.61 ± 0.92 for clinical-reasoning soundness, 4.50 ± 0.99 for literature-citation quality, and 4.83 ± 0.38 for overall clinical utility (Table 3, Figure 7). The physician’s overall mean was 4.78 ± 0.42 (reasoning 5.00 ± 0.00 , citation 4.56 ± 0.53 , utility 4.78 ± 0.44) and the physician’s was 4.52 ± 1.05 (reasoning 4.22 ± 1.20 , citation 4.44 ± 1.33 , utility 4.89 ± 0.33). The two reviewers agreed exactly on 59% of the 27 paired ratings and within one point on 85%, with a mean absolute difference of 0.63 points. The few larger gaps came from the physician’s lower reasoning and citation scores on a small number of cases (detailed below); the remaining disagreements were one-point differences in both directions.

The free-text comments complemented the numerical scores. The physicist recorded numerical ratings without written remarks, while the physician annotated several cases. These annotations both affirmed the reports, noting for example that a borderline case (with changes between stable disease and possible recurrence) was nonetheless interpreted correctly, and localized failure modes that recurred across three reports. In one case (clinical-

Table 3: Expert evaluation of agent-generated reports on nine longitudinal GBM cases. Values are mean $\pm$ standard deviation on a 1–5 Likert scale (higher is better). The pooled column combines both raters ($n = 18$ per metric).

Dimension	Physicist ($n = 9$)	Physician ($n = 9$)	Pooled ($n = 18$)
Clinical-reasoning soundness	5.00 ± 0.00	4.22 ± 1.20	4.61 ± 0.92
Literature-citation quality	4.56 ± 0.53	4.44 ± 1.33	4.50 ± 0.99
Overall clinical utility	4.78 ± 0.44	4.89 ± 0.33	4.83 ± 0.38
Overall	4.78 ± 0.42	4.52 ± 1.05	4.65 ± 0.80

reasoning score 2, citation score 1), the agent cited additional GBM cohorts as supporting evidence without establishing their relevance to the index patient’s outcome and suggested a supplementary analysis (examining histograms or percentiles) without providing supporting citation. In two additional cases (clinical-reasoning score 3), the agent inferred the acquisition timeline (labeling scans as “baseline,” “early treatment,” or “one month after radiation,” and even estimating fraction numbers) from typical treatment schedules reported in the literature rather than from the study metadata. In one of these cases, the agent also introduced diffusivity measures derived from very high b -values (e.g., $b = 2500$ and 3000 s/mm^2) without explaining how these measures should be interpreted or weighed. These failure modes reflect limitations in grounding rather than fluency, highlighting cases in which the platform’s passage-tracing design did not fully achieve its intended purpose. We discuss the implications of these findings further in the Discussion.

4 Discussion

This work combined two capabilities needed to reliably use per-fraction DWI on the MR-Linac: reliable processing, and interpretation of the resulting parameter trajectories against the published evidence. The platform described in this study unifies them in a single auditable tool, taking raw DICOM through validated deep-learning distortion correction, denoising, and IVIM/ADC fitting to longitudinal, ROI-based parameter maps, and then sub-

mitting those maps to a RAG agent that returns a structured, citation-traced interpretation. Domain experts judged these interpretations to be clinically usable. Pooled across 54 independent ratings, reports scored 4.65 ± 0.80 on a 1–5 scale, with overall clinical utility rated highest (4.83 ± 0.38) and 93% of all ratings at “good” or above.

Three design decisions likely contributed to these ratings. First, traceability: by linking citations directly to retrieved passages rather than relying on the LLM’s internal knowledge, the system enables reviewers to verify individual claims against the corresponding source text. Both experts viewed this capability favorably, and it represents an important distinction from a standalone LLM. However, citation received the lowest ratings in cases where claims were not adequately supported by matched evidence, indicating that traceability is only effective when grounding is consistently maintained. Second, separation of computation from reasoning: assigning percentage-change and trend calculations to deterministic tools minimizes arithmetic errors and ensures that subsequent interpretation is based on reproducible numerical results. Third, tumor-type-aware retrieval from a curated corpus: restricting retrieval to a vetted, line-indexed knowledge base and filtering evidence for case compatibility helps base the interpretation on clinically relevant literature. Together, these design choices address a key barrier to IVIM adoption: the burden of integrating and interpreting heterogeneous quantitative evidence [20, 24], rather than limitations in imaging performance or processing.

The evaluation also revealed the system’s limits. The physician’s lower scores corresponded to specific, identifiable failure modes rather than rating noise. The agent, when not given explicit acquisition metadata, inferred scan timing from typical treatment schedules in the literature; and in at least one case it drew on GBM cohorts that were not matched to the index case on the relevant outcome, and suggested an analysis without providing a supporting reference. These behaviors represent the residual risk that LLM-based interpreter carries, and point to implementable safeguards: (1). passing verified study metadata (true fraction numbers and dates) to the agent rather than letting it infer them; (2). constraining retrieval

and cohort comparisons to studies with protocols and outcome definitions that are compatible with the case, while explicitly flagging any mismatches; (3). requiring that every analysis carry a citation or be labeled as general clinical reasoning; (4). restricting the report to a pre-specified parameter set, with any auxiliary metric (e.g., very-high- b diffusivity) included only when its interpretive relevance and appropriate weighting are explicitly stated; and (5). reporting a confidence level that is reduced for atypical or under-specified cases rather than forcing a definitive interpretation. The fact that two reviewers, independently applying the same rubric, could pinpoint the specific statements in which these issues occurred provides evidence that the traceable-report design supports the intended level of auditability.

This platform differs in scope from existing tools. Research-oriented DICOM managers streamline data curation for downstream analysis [25], and a growing body of work has established the accuracy and repeatability of quantitative DWI on the MR-Linac [2, 26, 27]. The platform builds on this foundation by extending the workflow beyond quantification to interpretation and, consistent with broader shift toward agentic, evidence-grounded medical AI [30], generates reports whose statements can be directly verified against their supporting sources. Although the validated clinical use case is longitudinal GBM, the architecture itself is organ-agnostic. The knowledge base already spans head-and-neck, prostate, and pancreatic disease, so the same processing-to-interpretation path can be potentially generalized by extending the corpus.

This study has several limitations. The evaluation is a single-institution, pilot-scale study with nine cases, two raters, and a single underlying LLM configuration, which supports feasibility and clinical-quality assessment but not deployment-ready generalization. The reports were rated for quality, reasoning, and utility, not against patient outcomes. In addition, we did not evaluate whether the agent’s IVIM-based assessment distinguishes progression from pseudoprogression more accurately than an ADC-only baseline, because ground-truth outcome labels based on pathology or ≥ 6 -month imaging follow-up were beyond the scope of this study.

5 Conclusions

We present an integrated, web-based platform that takes MR-Linac diffusion-weighted imaging from raw DICOM data to a structured, literature-grounded clinical interpretation by combining validated deep-learning-based distortion correction, denoising, and IVIM/ADC fitting with a retrieval-augmented interpretation agent that links its statements to specific passages in the source literature. In an independent evaluation of nine longitudinal glioblastoma cases, a medical physicist and a physician rated the agent-generated reports as clinically usable, with an overall score of 4.65 ± 0.80 on a 1–5 scale and the highest ratings for clinical utility. The same evaluation identified the system’s remaining limitations (primarily inferred acquisition metadata and the use of imperfectly matched supporting evidence), and localized them to specific, correctable statements, illustrating how the platform’s traceable design can make LLM-based reasoning auditable in radiation oncology. By integrating quantitative processing and literature-grounded interpretation within a single verifiable platform, this work represents a step toward broader use of multiparametric IVIM for treatment response monitoring beyond specialized centers.

Acknowledgments

This research was supported by the National Institutes of Health (NIH) Grants R01 EB034691, R01 CA240808, R01 CA258987, and R01 CA280135.

Ethics Approval

All datasets were retrospectively collected from an approved study at the UT Southwestern Medical Center, under an umbrella IRB protocol 082013-008 (Improving radiation treatment quality and safety by retrospective data analysis). This is a retrospective analysis study and not a clinical trial. No clinical trial ID number is available.

Conflict of Interest

The authors have no relevant conflicts of interest to disclose.

Data Availability Statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

References

- [1] Simon Boeke, Jonas Habrich, Sarah Kübler, et al. Longitudinal assessment of diffusion-weighted imaging during magnetic resonance-guided radiotherapy in head and neck cancer. *Radiation Oncology*, 20(1):15, 2025. doi: 10.1186/s13014-025-02589-9.
- [2] Liam S. P. Lawrence, Rachel W. Chan, Hanbo Chen, et al. Accuracy and precision of apparent diffusion coefficient measurements on a 1.5 T MR-Linac in central nervous system tumour patients. *Radiotherapy and Oncology*, 164:155–162, 2021. doi: 10.1016/j.radonc.2021.09.020.
- [3] Denis Le Bihan, Eric Breton, Denis Lallemand, M.-L. Aubin, J. Vignaud, and M. Laval-Jeantet. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*, 168(2):497–505, 1988. doi: 10.1148/radiology.168.2.3393671.
- [4] Mami Iima and Denis Le Bihan. Clinical intravoxel incoherent motion and diffusion MR imaging: past, present, and future. *Radiology*, 278(1):13–32, 2016. doi: 10.1148/radiol.2015150244.
- [5] Denis Le Bihan. What can we see with IVIM MRI? *NeuroImage*, 187:56–67, 2019. doi: 10.1016/j.neuroimage.2017.12.062.

- [6] Roger Stupp, Warren P. Mason, Martin J. van den Bent, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005. doi: 10.1056/NEJMoa043330.
- [7] Dieta Brandsma, Lukas Stalpers, Walter Taal, Peter Sminia, and Martin J. van den Bent. Clinical features, mechanisms, and management of pseudoprogression in malignant gliomas. *The Lancet Oncology*, 9(5):453–461, 2008. doi: 10.1016/S1470-2045(08)70125-6.
- [8] Dieta Brandsma and Martin J. van den Bent. Pseudoprogression and pseudoresponse in the treatment of gliomas. *Current Opinion in Neurology*, 22(6):633–638, 2009. doi: 10.1097/WCO.0b013e328332363e.
- [9] Alba A. Brandes, Enrico Franceschi, Alicia Tosoni, et al. MGMT promoter methylation status can predict the incidence and outcome of pseudoprogression after concomitant radiochemotherapy in newly diagnosed glioblastoma patients. *Journal of Clinical Oncology*, 26(13):2192–2197, 2008. doi: 10.1200/JCO.2007.14.8163.
- [10] Michael Jonathan Kucharczyk, Sameer Parpia, Anthony Whitton, and Jeffrey Noah Greenspoon. Evaluation of pseudoprogression in patients with glioblastoma. *Neuro-Oncology Practice*, 4(2):120–134, 2017. doi: 10.1093/nop/npw021.
- [11] Patrick Y. Wen, David R. Macdonald, David A. Reardon, et al. Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *Journal of Clinical Oncology*, 28(11):1963–1972, 2010. doi: 10.1200/JCO.2009.26.3541.
- [12] Francesca Nava, Irene Tramacere, Andrea Fittipaldo, et al. Survival effect of first- and second-line treatments for patients with primary glioblastoma: a cohort study from a prospective registry, 1997–2010. *Neuro-Oncology*, 16(5):719–727, 2014. doi: 10.1093/neuonc/not316.

- [13] Charalampos Tsakiris, Timoleon Siempis, George A. Alexiou, et al. Differentiation between true tumor progression of glioblastoma and pseudoprogression using diffusion-weighted imaging and perfusion-weighted imaging: systematic review and meta-analysis. *World Neurosurgery*, 144:e100–e109, 2020. doi: 10.1016/j.wneu.2020.07.218.
- [14] Karl H. Plate, Georg Breier, Herbert A. Weich, and Werner Risau. Vascular endothelial growth factor is a potential tumour angiogenesis factor in human gliomas in vivo. *Nature*, 359(6398):845–848, 1992. doi: 10.1038/359845a0.
- [15] Dolores Hambardzumyan and Gabriele Bergers. Glioblastoma: defining tumor niches. *Trends in Cancer*, 1(4):252–265, 2015. doi: 10.1016/j.trecan.2015.10.009.
- [16] Pejman Jabejdar Maralani, Sten Myrehaug, Hatef Mehrabian, et al. Intravoxel incoherent motion (IVIM) modeling of diffusion MRI during chemoradiation predicts therapeutic response in IDH wildtype glioblastoma. *Radiotherapy and Oncology*, 156:258–265, 2021. doi: 10.1016/j.radonc.2020.12.037.
- [17] Josep Puig, Javier Sánchez-González, Gerard Blasco, et al. Intravoxel incoherent motion metrics as potential biomarkers for survival in glioblastoma. *PLOS ONE*, 11(7):e0158887, 2016. doi: 10.1371/journal.pone.0158887.
- [18] Dan Liao, Yuan-Cheng Liu, Jiang-Yong Liu, Di Wang, and Xin-Feng Liu. Differentiating tumour progression from pseudoprogression in glioblastoma patients: a monoexponential, biexponential, and stretched-exponential model-based DWI study. *BMC Medical Imaging*, 23(1):119, 2023. doi: 10.1186/s12880-023-01082-7.
- [19] Hechuan Luo, Ling He, Weiqin Cheng, and Sijie Gao. The diagnostic value of intravoxel incoherent motion imaging in differentiating high-grade from low-grade gliomas: a systematic review and meta-analysis. *The British Journal of Radiology*, 94(1121):20201321, 2021. doi: 10.1259/bjr.20201321.

- [20] Otto M. Henriksen, María del Mar Álvarez-Torres, Patricia Figueiredo, et al. High-grade glioma treatment response monitoring biomarkers: a position statement on the evidence supporting the use of advanced MRI techniques in the clinic, and the latest bench-to-bedside developments. part 1: Perfusion and diffusion techniques. *Frontiers in Oncology*, 12:810263, 2022. doi: 10.3389/fonc.2022.810263.
- [21] Yunxiang Li, Yen-Peng Liao, Yan Dai, Jie Deng, and You Zhang. Landmark matching and B-spline implicit neural representations for diffusion-weighted imaging distortion correction. *Physics in Medicine & Biology*, 71(4):045005, 2026. doi: 10.1088/1361-6560/ae4162.
- [22] Yunxiang Li, Yan Dai, Yen-Peng Liao, Jie Deng, and You Zhang. Band-limited implicit neural representations for diffusion-weighted imaging denoising. *Physics in Medicine & Biology*, 71(1):015033, 2026. doi: 10.1088/1361-6560/ae2a9e.
- [23] Yunxiang Li, Yen-Peng Liao, Yan Dai, Jie Deng, and You Zhang. Accurate estimation of intravoxel incoherent motion parameters based on implicit neural representation. *Medical Physics*, 53(8):e70599, 2026. doi: 10.1002/mp.70599.
- [24] Emmanuel Mesny, Benjamin Leporq, Olivier Chapet, and Olivier Beuf. Intravoxel incoherent motion magnetic resonance imaging to assess early tumor response to radiation therapy: review and future directions. *Magnetic Resonance Imaging*, 108:129–137, 2024. doi: 10.1016/j.mri.2024.02.008.
- [25] Austen Maniscalco, Yang Kyun Park, Andrew Godley, Mu-Han Lin, Steve Jiang, and Dan Nguyen. MedicalDataHandler, a research-oriented graphical user interface for DICOM data management. *Medical Physics*, 53(1):e70240, 2026. doi: 10.1002/mp.70240.
- [26] Jonas Habrich, Simon Boeke, Marcel Nachbar, et al. Repeatability of diffusion-weighted magnetic resonance imaging in head and neck cancer at a 1.5 T MR-Linac. *Radiotherapy and Oncology*, 174:141–148, 2022. doi: 10.1016/j.radonc.2022.07.020.

- [27] Ernst S. Kooreman, Petra J. van Houdt, Marlies E. Nowee, et al. Feasibility and accuracy of quantitative imaging on a 1.5 T MR-linear accelerator. *Radiotherapy and Oncology*, 133:156–162, 2019. doi: 10.1016/j.radonc.2019.01.011.
- [28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.
- [29] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, 2021. doi: 10.18653/v1/2021.findings-emnlp.320.
- [30] Weike Zhao, Chaoyi Wu, Yanjie Fan, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784, 2026. doi: 10.1038/s41586-025-10097-9.
- [31] Miguel Grinberg. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2014.
- [32] Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8024–8035, 2019.
- [33] Darcy Mason. SU-E-T-33: pydicom: an open source DICOM library. *Medical Physics*, 38(6):3493, 2011. doi: 10.1118/1.3611983.
- [34] Bradley C. Lowekamp, David T. Chen, Luis Ibáñez, and Daniel Blezek. The design of SimpleITK. *Frontiers in Neuroinformatics*, 7:45, 2013. doi: 10.3389/fninf.2013.00045.
- [35] M. Jorge Cardoso, Wenqi Li, Richard Brown, et al. MONAI: an open-source framework

- for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. doi: 10.48550/arXiv.2211.02701.
- [36] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkan Liang, Xin Zhou, and Xiang Bai. MINIMA: modality invariant image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [37] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 7462–7473, 2020.
- [38] Harrison Chase. LangChain. <https://github.com/langchain-ai/langchain>, 2022. Open-source framework for building applications with large language models.
- [39] Daniel Moore-Palhares, Liam S. P. Lawrence, Sten Myrehaug, et al. Temporal apparent diffusion coefficient changes during chemoradiation: an imaging biomarker for tumor response monitoring and spatial recurrence prediction in glioblastoma. *International Journal of Radiation Oncology, Biology, Physics*, 122(3):592–604, 2025. doi: 10.1016/j.ijrobp.2025.03.028.
- [40] Liam S. P. Lawrence, Rachel W. Chan, Hanbo Chen, et al. Diffusion-weighted imaging on an MRI-linear accelerator to identify adversely prognostic tumour regions in glioblastoma during chemoradiation. *Radiotherapy and Oncology*, 188:109873, 2023. doi: 10.1016/j.radonc.2023.109873.
- [41] Danilo Maziero, Michael W. Straza, John C. Ford, Joseph A. Bovi, Tejan Diwanji, Radka Stoyanova, Eric S. Paulson, and Eric A. Mellon. MR-guided radiotherapy for brain and spine tumors. *Frontiers in Oncology*, 11:626100, 2021. doi: 10.3389/fonc.2021.626100.