AAPM 2026
68TH ANNUAL MEETING & EXHIBITION
JOINT AAPM | COMP MEETING
JULY 19 – 22 | VANCOUVER, CANADA

EMBRACING CHANGE.
IMPACTING PATIENT CARE.

Patient-Specific and Population-Level Insights in Radiation Oncology via an Agentic LLM Platform

Hao Jiang[1,3], Xuejun Gu[1,2], Arnold Pompos[1], Steve Jiang[1], Weiguo Lu[1]*

[1] MAIA Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX
[2] Department of Radiation Oncology, Stanford University, Stanford, CA
[3] NeuralRad LLC, Madison, WI
Contact: Hao.Jiang@UTSouthwestern.edu · Weiguo.Lu@UTSouthwestern.edu

MAIA Laboratory
UT Southwestern Medical Center
Radiation Oncology

13	47	16-way	≤ 800ms	1,152
CLINICAL DATA SOURCES	SPECIALISED AGENT TOOLS	CONCURRENT TOOL DISPATCH	FIRST REASONING TOKEN	PULSAR PATIENTS 2018 – 2026

PURPOSE

A single radiation-oncology patient's information is spread across 13+ specialised information systems spanning five distinct data domains:

- **EMR** — Epic Clarity + Caboodle (demographics, encounters, notes, impressions, pathology, labs).
- **OIS** — MOSAIQ + Aria (treatment management, scheduling, prescriptions, machine logs, delivery history).
- **TPS** — Monaco + Eclipse (3-D dose grids, plans, contours, CT / MR registration, MLC / beam parameters).
- **Imaging DB** — VarianMirror + ImagingDB (DICOM CT / MR / PT studies, RT structure sets, RT-dose objects, C-FIND / C-MOVE).
- **Operational & analytical** — OneSign (signed documents), Equator (machine QA), DataWarehouse (departmental analytics), MedLever (trial protocols), QGenda (scheduling), RTLS (location), Qdrant (vector search), MongoDB.

Each system has its own schema, authentication, and query language. Assembling a single patient's complete picture takes hours of manual navigation; population-level analyses — clinical-trial matching, cohort discovery, treatment-eligibility screening, care-pattern audit — conventionally require weeks of chart review.

Objective: deliver a production-grade agentic harness — a coding-agent core (Azure OpenAI GPT-5.2) wrapped in radiation-oncology skills, specialised data-discovery tools, parallel multi-source dispatch, persistent self-learning, and read-only guardrails — that, in response to one natural-language prompt, autonomously extracts, fuses, and reasons across the entire UTSW radiation-oncology data landscape, powering single-patient lookups and population-scale cohort discovery at clinician-actionable latency.

METHOD

A coding-agent harness (Fig. 1). At its core the platform is a coding agent — Azure OpenAI GPT-5.2 writes SQL queries, Python / Node.js scripts, DICOM C-FIND requests, and structured JSON tool calls on the fly to answer each clinical question. Around it we built an LLM-agent harness: a Go-based REST + WebSocket backend that loads radiation-oncology skills at startup, dispatches up to 16 tools in parallel, runs a four-step Reason → Act → Observe → Respond loop for up to 25 iterations, persists self-learned knowledge between sessions, and enforces read-only guardrails on every interaction.

Specialised data-discovery tools. Because each of the 13 clinical databases has its own schema, the harness ships dedicated discovery tools — list_tables, describe_columns, search_schemas, sample_data, check_values — that let the agent autonomously explore an unfamiliar database, ask the system what's there, sample a few rows, and synthesise a correct query. The agent reasons like a human analyst, not a hard-coded ETL pipeline, so new data sources are onboarded with no Python plumbing.

Tool taxonomy (47 total). Database query (12 — SQL across MOSAIQ / Clarity / OneSign / Equator / DataWarehouse / MongoDB), Schema discovery (5), REST API (5 — Aria / MedLever / QGenda / RTLS), DICOM (4), Monaco TPS (4), File operations (3), Sandboxed Python / Node.js (2), Visualisation (4), Vector search (1).

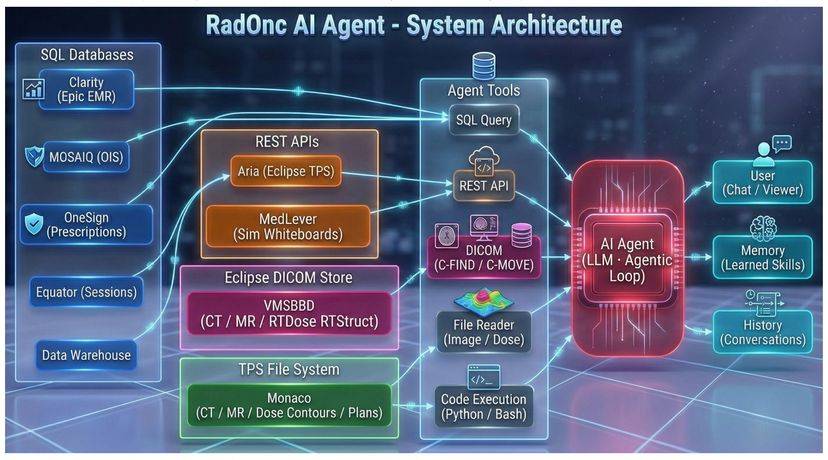


Fig. 1. System architecture — 13 read-only clinical data sources (left) → 47 specialised agent tools (centre) → GPT-5.2 AI agent core with persistent memory (right) → user chat / viewer.

METHOD (cont.)

Agentic reasoning loop (Fig. 2). Every iteration runs a four-step pattern: Reason — analyse current state, previous tool results, and remaining objectives; Act — select and invoke one or more tools (parallel execution supported, up to 16-way); Observe — process tool outputs, extract relevant information, update understanding; Respond — when enough information is gathered, stream a final response over WebSocket. The agent decides autonomously when no further tool call is needed.

Persistent self-learning. After every interaction that touches a database, up to 3 silent LLM calls extract new knowledge from the conversation; a TF-IDF filter deduplicates against existing skill files; consolidated facts are written to capped (8 000-character) Markdown skill files (mosaiq_learned.md, clarity_learned.md, etc.) plus per-patient caches (mrn/{MRN}.md). Updated knowledge is loaded into the system prompt at the next session — the agent measurably improves over time, with no manual curation.

Embedded interactive visualisation. The chat interface integrates: (a) a CT/MR viewer with togglable structure overlays and windowing controls; (b) dose-colorwash overlay with configurable opacity and a jet colormap (0 → 62.3 Gy); (c) an interactive DVH explorer with point-specific dose-volume metrics (e.g. $D_{0.035cc}$ = 10.7 Gy); (d) GPU-accelerated Monte Carlo dose recalculation via the integrated gDPM engine for independent QA verification and adaptive replanning.

Read-only guardrails + auditability. All write SQL (INSERT / UPDATE / DELETE / DROP / ALTER / CREATE / TRUNCATE / EXEC / MERGE) is blocked at the parser; dangerous shell commands are sandboxed (120-s timeout, 30 k-character output cap, blocklist for rm -rf, mkfs, dd, fork bombs, chmod 777); WebSocket sessions use per-user isolation; AE-title verification gates DICOM associations; every tool call, intermediate inference, and final verdict is journaled to an auditable session ledger. Deployment is live at http://rop589298.swmed.org against production UTSW clinical mirror databases.

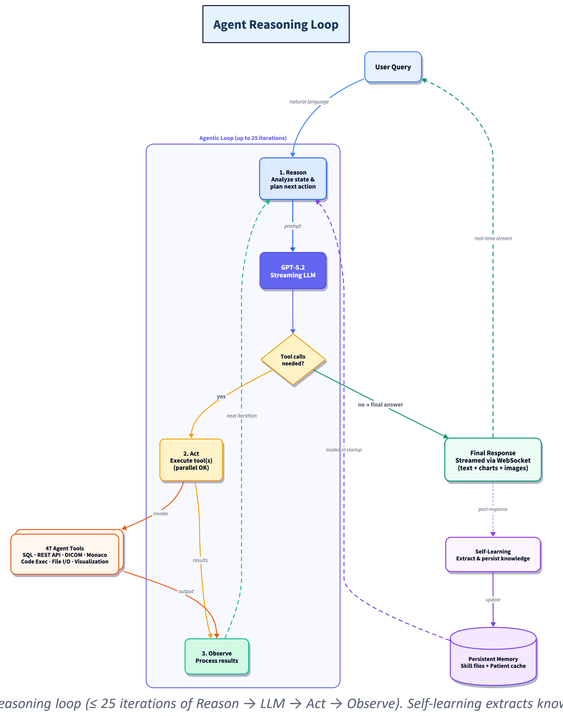


Fig. 2. Agentic reasoning loop (≤ 25 iterations of Reason → LLM → Act → Observe). Self-learning extracts knowledge into persistent skill files post-response.

> ***"We have been having great luck with your platform. We are finding patients who were otherwise going to be missed using RT history. Identified new HGG patients otherwise would be missed via platform — asked to overlap HGG treated with PULSAR > 80 % precise cohort (65 patients). One prompt."***
>
> *— Michael Dohopolski, M.D., UT Southwestern Dept. of Radiation Oncology*

RESULTS

Production deployment. The platform is live at the UTSW Department of Radiation Oncology, integrating 13 heterogeneous data sources behind a single natural-language interface. The same architecture powers the Web chat UI, an external REST API verified against the production server via a documented Python example client (POST /api/chat and POST /api/tool-dispatch), and an SSE patient-load surface that fans every applicable data source out in parallel for a single MRN.

PULSAR cohort discovery (Fig. 3). Asked to find every UTSW PULSAR patient since the technique was introduced, the agent autonomously queried MOSAIQ, mined 6 039 free-text schedule + site notes, and parsed site names against an ultra-hypofractionation rubric — identifying 1 152 unique patients across 2018 – 2026 (≈ 4 × growth from 43 patients in 2018 to 165 in 2025). The same run inferred anatomical category for 867 patients (brain 180, spine 115, prostate 86, thorax/lung 61, pelvis 42, pancreas 41, kidney/adrenal 31, liver 28, bone/limb 27, rectum 15, cervix/uterus 14), flagged 251 SIB treatments, and surfaced 286 patients with > 1-year follow-up windows — candidate cohorts for re-irradiation review and outcome studies.

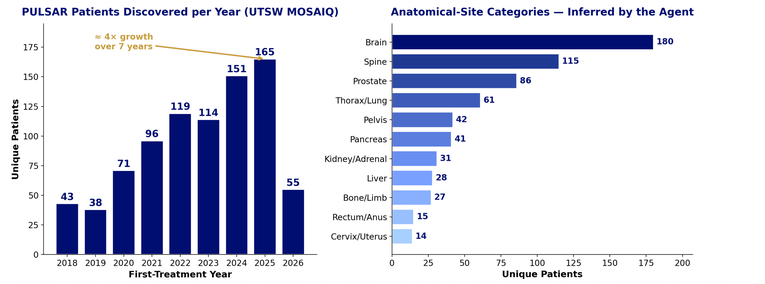
HGG ∩ PULSAR sub-cohort. From the full 1 152-patient cohort the agent then filtered for high-grade-glioma patients also treated with PULSAR at ≥ 80 % target precision, returning a ranked 65-patient candidate list within minutes — surfacing previously-missed re-irradiation candidates.

Cross-database patient summary. A physician requests a comprehensive treatment summary for a patient by MRN. The agent autonomously queries MOSAIQ for treatment courses + machine parameters, Clarity for clinical notes + lab results, Aria for plan details, and Monaco for the dose grid + structures — rendering an inline dose colorwash with PULSAR-eligibility checks. End-to-end ≈ 5 s.

Population-level survival analysis. Queried for survival status of pancreatic-cancer patients treated on Unity1/Unity2 MR-Linac systems over the past 6 months, the agent joined MOSAIQ course data filtered by machine + time with Clarity DEATH_DATE records: 23 patients, 1 death, 95.7 % alive fraction with full per-patient breakdown.

Adaptive-therapy tracking. Monitoring PTV-volume evolution across 11 adaptive Unity MR-Linac fractions, the agent reads Monaco plan data, computes ΔVolume vs. the reference (ADT01), and renders both a tabular summary (121.36 → 77.41 cc, Δ = −43.95 cc) and a longitudinal trend chart via sandboxed Python execution — entirely from the chat interface.

Interactive auditable view (Fig. 4). Every finding is inspected through an embedded canvas with 3-D dose colorwash on CT, structure-set toggling, DVH explorer with point-specific dose-volume metrics, and a side panel listing every tool the agent invoked — full provenance from query to verdict.



1,152 unique PULSAR patients identified across 2018-2026 · 867 with categorisable anatomical site · 251 with SIB treatments · 286 with > 1-year follow-up

Fig. 3. Autonomous PULSAR cohort discovery — 1 152 unique patients 2018-2026 mined from MOSAIQ + 6 039 free-text notes (left) with inferred anatomical-site categorisation (right).

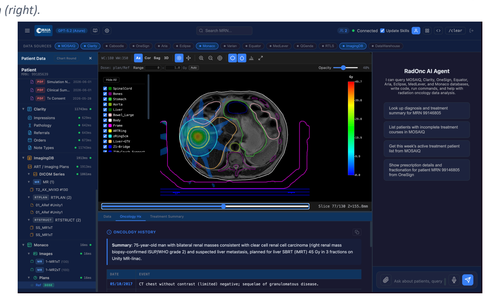


Fig. 4. Clinical canvas in production — 3-D dose colorwash on CT (centre) + interactive AI agent panel with suggested clinical queries (right).

CONCLUSIONS

An autonomous multi-database AI agent — combining a GPT-5.2-driven iterative reasoning loop, 47 specialised tools fronting 13 heterogeneous read-only clinical data sources, persistent self-learning skill files, GPU-accelerated Monte Carlo dose calculation, embedded interactive visualisation, and multi-layer read-only guardrails — delivers patient-specific and population-level insights at clinician-actionable latency. Real-world deployment at UTSW Radiation Oncology has surfaced clinical-trial and treatment-eligibility candidates that prior manual chart-review workflows would have missed, with validated production use cases including HGG ∩ PULSAR cohort discovery, cross-database patient summaries, population-level survival analysis, adaptive-therapy tracking, and automated PDF report generation. The framework generalises beyond the current MOSAIQ / Clarity / Aria-Eclipse / Monaco footprint and is ready for expansion to predictive analytics, prospective decision support, and clinical-trial enrolment workflows.

REFERENCES

[1] Jiang H. et al. Chat with Oncology Information System Via Large Language Model. AAPM 2025.
[2] OpenAI GPT-5.2, Azure OpenAI Service. 2026.
[3] Microsoft Phi-4 Technical Report. December 2024.
[4] gDPM — GPU-based Dose Planning Method (CUDA Monte Carlo).
[5] LangGraph multi-agent orchestration — github.com/langchain-ai/langgraph.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants R01 CA235723-01, R01 CA280135, R01 EB034691-01, and NIH/NCI SBIR 75N91021C00031 and 75N91023C00032, and by computational resources from the UTSW MAIA Laboratory.