SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Spinning the Data Engine: From Data Curation, to Model Validation, to Clinical Trials with an In-House Review Portal

Nicholas Molina, Penn Medicine; Kelechi Obike; Jonathan Cheng; Rafe McBeth, PhD

Briefly describe what attendees will see during your demo.
We start the application stack and load several anonymized breast-radiotherapy studies into the local imaging server, then walk the data engine lifecycle in order. Curation first, the upstream step, where a reviewer scores contour quality to assemble a clean training dataset. Then the prediction overlay: a model's raw probability map on the scan, where we tune the confidence threshold and take a developer's first look at whether the output is sane. Next, validation—a case in blinded mode, where the reviewer rates "Plan A" and "Plan B" and states a preference without knowing which is machine-generated, then the non-blinded mode of the same case, revealing the "Ground Truth" and "Artificial Intelligence" labels with per-structure Dice and Hausdorff metrics live on the image. Finally, dose display (isodose color-wash and dose-volume histogram), the downstream planning context the validated contours feed into. Throughout, attendees see the full reviewer experience, the scoring that gets captured, and how structure-naming differences between studies are reconciled automatically.

Problem & Motivation: What clinical or operational problem does your MVP address?
Building AI models in-house on our own clinical data has a huge advantage: we can compress the loop from clinical need, to model development, to validation, to modification, and get a high-quality model into the clinic faster. But that loop is only as fast as its slowest link, and today the bottleneck is review and validation. Clinical systems (treatment-planning systems, PACS) are slow to iterate in and simply don't have the features we need to interrogate a model's output. A model can be trained in days but it can take weeks for anyone to tell us whether it's actually good. Reviewing contours inside a planning system is cumbersome and biasing: overlaying an "AI" contour next to a "human" one primes the reviewer, and overlap metrics like Dice usually live in offline scripts, disconnected from the images, so a clinically unacceptable contour can hide behind a high average. We built a lightweight, DICOM-native portal to be the fast review layer that closes this loop. A clinician can compare two contour sets blinded, rate each and pick a preference, with the AI-versus-human identity hidden and randomized—then switch to a non-blinded view revealing labels alongside live metrics. The same tool curates and quality-rates the training data upstream, and extends to dose display and autocontouring prediction overlay downstream: one portal spanning curation, validation, and downstream review. Efficient, more trustworthy model iteration ultimately means better plans reaching patients sooner.

What You Built: Describe your MVP, tool, or workflow in plain language
A browser-based review portal for radiotherapy imaging. It pulls CT scans and their structure sets from a standard imaging server and shows them in synchronized axial, sagittal, and coronal views with the contours and optionally dose overlaid. It has three modes. Data curation (the original purpose): a reviewer steps through studies and rates contour/dose quality for building training datasets. Blinded comparison: for each anatomical structure, the portal shows two versions—one expert/ground-truth, one AI-generated—as neutral

"Plan A" and "Plan B," randomly flipped per study so the reviewer can't infer which is which. They rate each and record a preference. Non-blinded comparison: the same case with the labels revealed as "Ground Truth" and "AI," plus per-structure Dice and Hausdorff distance shown live on screen. Prediction Overlay: reviewers can upload paired model probability and geometry files for an individual study. The portal aligns the prediction to the CT and displays selectable organ or target channels as a mask, probability heatmap, confidence penumbra, or entropy view alongside the clinical contours. Inputs: DICOM CT plus one or more structure sets (a ground-truth set and an AI set). Outputs: structured review records—per-structure ratings, A/B preferences, and computed overlap metrics—stored in a database and exportable for analysis.

Tech Stack & Development Approach: What tools, frameworks, or methods did you use to build it?

Python and Flask on the backend; a hand-written JavaScript canvas viewer on the frontend (no heavy imaging framework). DICOM parsing with pydicom; Dice and Hausdorff computed with NumPy and SciPy. Reviews persist to SQLite. Studies are served from an Orthanc DICOM server. The whole stack, application plus imaging server, is packaged with Docker Compose and Python's uv for reproducible, one-command startup on any machine. Development was heavily agentic: built largely with Claude Code in the terminal, using a brainstorm-then-plan-then-execute workflow where implementation tasks were handed to subagents and reviewed before merging. Initially started hand-coded + ChatGPT/Claude chats on browser, then adopted latest AI tools along the way such as Cursor, and eventually fully switching to Claude Code and Codex. We kept the footprint deliberately basic and standards-based (plain DICOM, a real imaging server, containers) so it's portable and easily run elsewhere. Prediction outputs are generated with in-house trained deep-learning models and uploaded through the Flask interface as paired NumPy and JSON files. SciPy-based resampling aligns each prediction channel to the CT grid, while memory-mapped arrays support efficient slice rendering. The backend then composites the selected prediction directly into the existing JavaScript viewer.

Validation & Evidence: Have you done any informal testing or validation? If so, what did you find?

Our blinded A/B review workflow was used to run a single-institution, IRB-approved feasibility study comparing AI-generated versus manually drawn partial-breast-irradiation target contours. An in-house deep-learning model generated contours for a set of consecutively treated, non-training cases; three breast radiation oncology physicians independently reviewed AI-versus-manual structures presented blinded and in random order inside the portal—exactly the workflow we're demonstrating. Reviewers scored gross, clinical, and planning target volumes and recorded an overall preference, yielding a full set of structure-level assessments captured through the tool. The portal recorded per-structure ratings, A/B preferences, and per-structure Dice, and those exports fed the downstream statistical analysis. The Prediction Overlay has also been tested end-to-end with real model outputs on anonymized validation cases. We verified prediction upload, CT alignment, channel selection, threshold control, and all four visualization modes. Testing exposed a geometry error that visibly displaced predictions from the anatomy. Correcting the spacing and orientation handling restored alignment with the CT and clinical contours. Broader quantitative and clinician-usability validation remains in progress. For the tool, the takeaway is that the blinded workflow held up under real expert use at study scale. We separately hit, and fixed, a real failure: structure names in clinically approved plans varied widely across studies and broke pairing, which drove an alias-matching feature which resolves each expected structure to whichever name is present, so comparisons don't silently break on naming drift.

Feedback You're Seeking: What specific feedback, help, or collaboration are you hoping to get from the CAIMI26 community?

1. Is the concept valuable? Our central question: do others see an in-house review portal like this as a

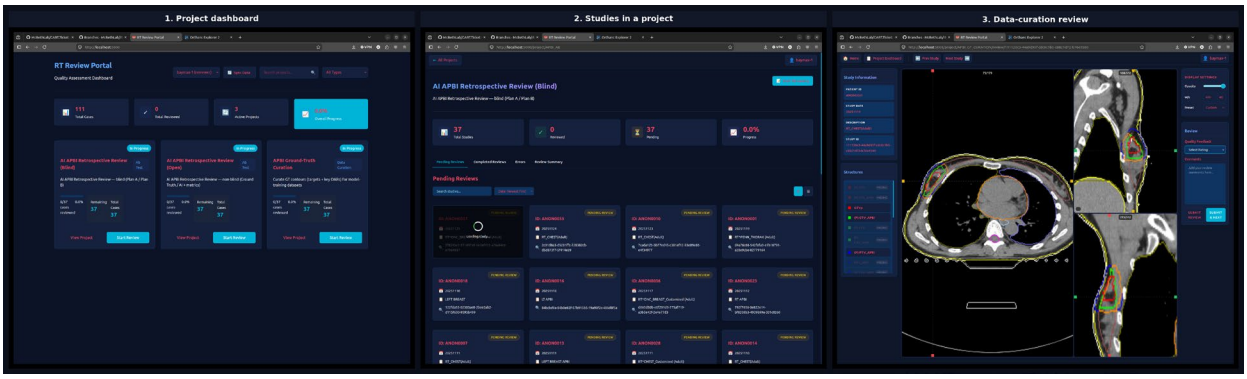
valuable way to strengthen local capability for model building, testing, and validation (i.e., to shorten the loop from clinical need to a validated, clinic-ready model)? Is "own your review layer" a direction worth investing in, or are we rebuilding something better bought? 2. Usability. Candid feedback on the reviewer experience: the blinded A/B flow, the metrics overlay, curation, and where it would slow a busy clinician down. 3. Other tools to fold in. What adjacent review workflows belong in the same portal? We're especially interested in a review layer for radiology report generation, and in what else an in-house "model engine" needs. 4. Multi-institutional validation. Sites willing to repeat the blinded comparison on their outputs and conventions, and input on a multi-reader, multi-site protocol. 5. Prediction & provenance. Which prediction displays (mask, heatmap, penumbra, entropy, disagreement) are most clinically useful, and what model/preprocessing/confidence info reviewers should see. 6. Nomenclature. How others handle structure-name variability across sites (i.e., mapping to a standard nomenclature such as the AAPM TG-263 conventions).

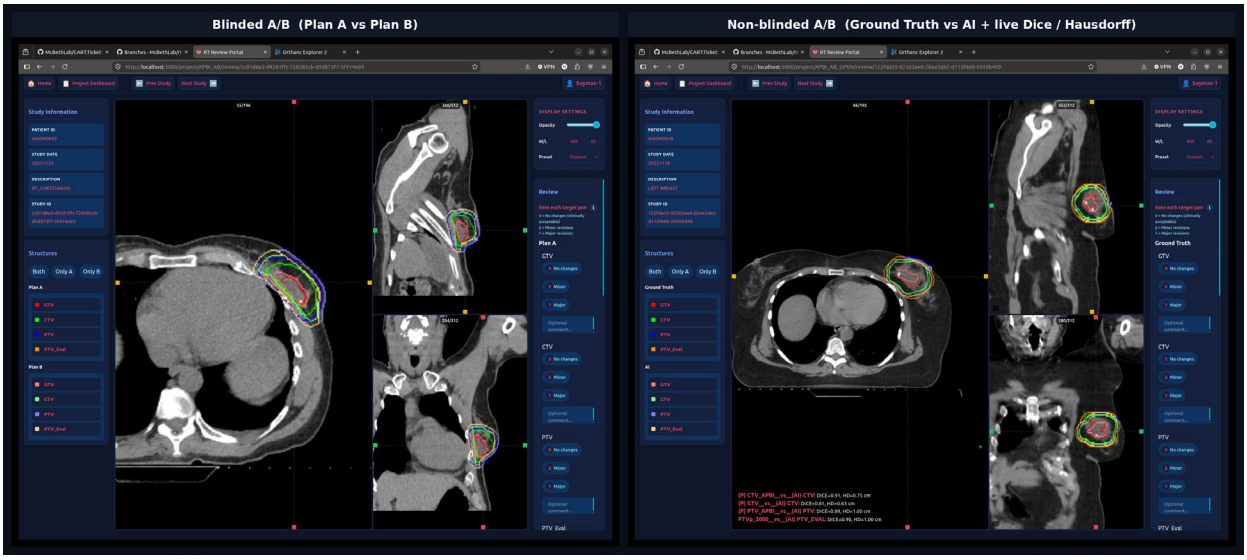
Known Limitations & Honest Failures: What isn't working yet, or what have you already tried that failed?

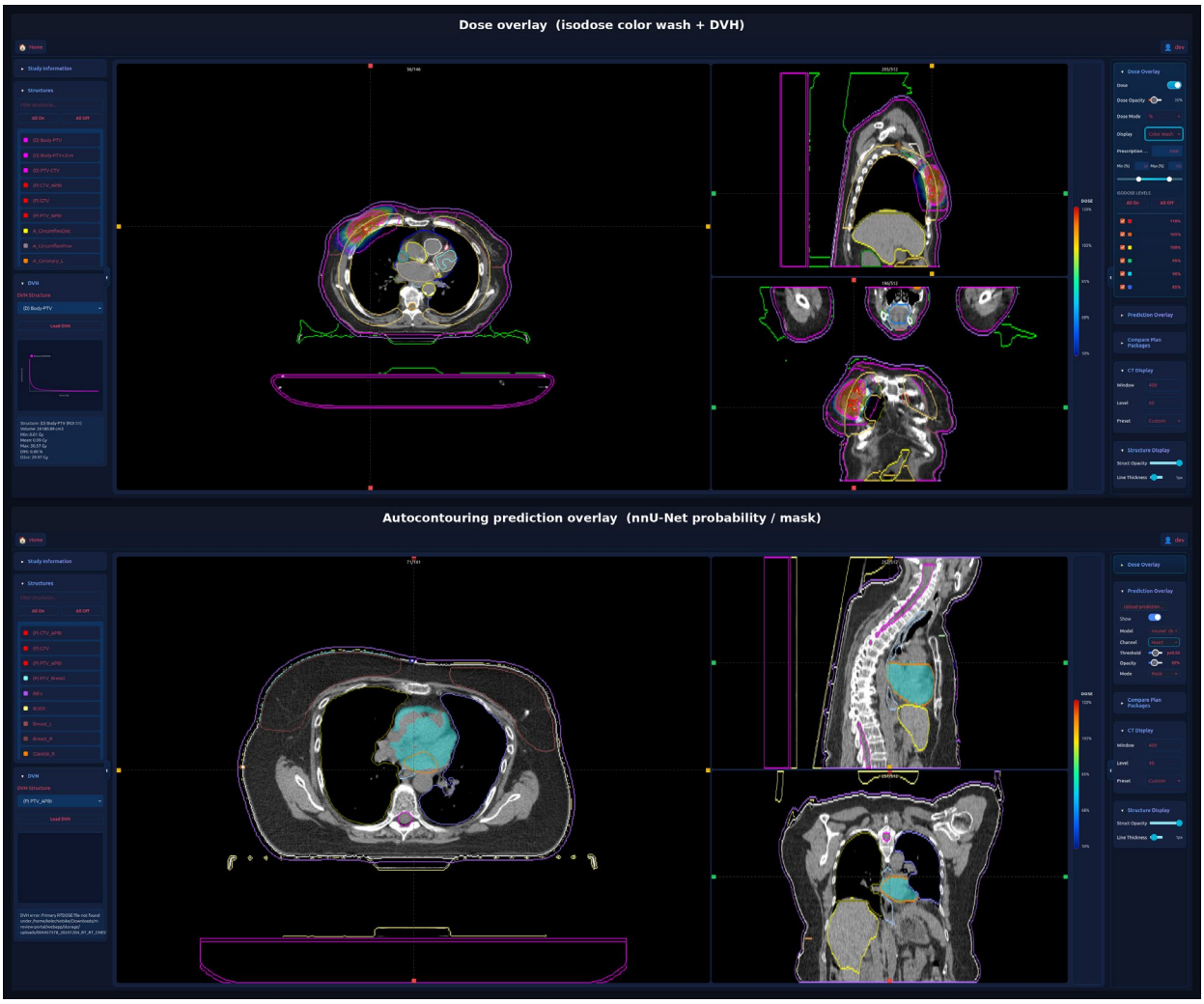
- Security depends on the institutional perimeter. Auth is not widely implemented. The metrics endpoint lacks authentication and an early build over-exposed the imaging server (since hardened: loopback-only, ingest port dropped). Today the real protection is that everything runs behind the institutional firewall on internal infrastructure. Fine for internal model work, not yet ready to expose beyond it. - Scale, speed, and memory optimizations. We haven't stress-tested many concurrent reviewers, abnormally large study volumes, or heavy interaction; large CT/prediction arrays are memory-hungry and prediction artifacts are still file-backed rather than in durable object storage. Rendering and memory management needs more robust profiling. - Single-institution validation. One site, three reviewers, one disease site (partial-breast); generalization across sites, disease sites, and reviewer pools is the next step, and metrics are only as trustworthy as the contours ingested. - Naming assumptions are local. The alias matcher is seeded from names we've seen locally; new sites need new aliases, and a principled nomenclature mapping isn't built yet.

Potential for Impact: If this MVP were fully developed and deployed, who would benefit and how?

The biggest impact is on the speed of the in-house model engine. A portal like this shortens every turn of the curate, build, validate, modify loop, so an institution can develop high-quality models faster, better meeting real clinical needs and, ultimately, benefiting patients through safer, more responsive adoption. When the review-and-validation step drops from weeks to days, in-house AI stops being a research side-project and becomes a repeatable capability. Concretely: radiation oncologists and physicists get a fast, unbiased way to quality-check automated contours with metrics anchored to the images, not a detached spreadsheet. Model developers get a structured evaluation harness plus a steady supply of quality-rated, curated training data from the same tool, and, via prediction overlay, can see exactly where a model agrees, fails, or is uncertain on the source CT, turning review into targeted model debugging. Trainees get a teaching instrument to compare their contours against expert and machine versions, blinded. Institutions get a portable, vendor-neutral capability they own outright. This isn't purely prospective, the portal has already run a real institutional blinded evaluation of an in-house model, showing it can generate the evidence a group needs to trust (or reject) a model clinically. Because it's DICOM-native and containerized, a group can stand it up without buying anything or waiting on a vendor.


1. Project dashboard
2. Studies in a project
3. Data-curation review


Blinded A/B (Plan A vs Plan B)
Non-blinded A/B (Ground Truth vs AI + live Dice / Hausdorff)


Dose overlay (isodose color wash + DVH)
Autocontouring prediction overlay (nnU-Net probability / mask)