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

A Rule-Validated Editing Tool for Faster, More Reliable Oncology Radiology Report

Igor Austin Martins, MD, MBA, UT Southwestern Medical Center

Briefly describe what attendees will see during your demo.
Short walkthrough video on a synthetic cancer report. First I show the manual pain: opening a prior report and counting everything that has to be updated by hand when the next scan arrives, measurements becoming priors, image references shifting, interval descriptors needing recalculation. Then I paste the report into the tool, which splits it into discrete findings. I update a few lesions using shorthand, the way I would speak at the workstation, and the tool regenerates each finding with the prior moved into place, the image reference reformatted, and the interval descriptor recalculated. I include a more complex case with vascular involvement to show it holds up beyond simple examples. All data shown is synthetic.

Problem & Motivation: What clinical or operational problem does your MVP address?
Radiology longitudinal comparison is fundamental to oncologic imaging. Every restaging report has to state whether disease grew, shrank, or stayed stable, and that assessment drives treatment decisions. Producing these comparisons is repetitive and error prone. For each tracked lesion, the radiologist carries forward the prior measurement, updates the image reference because the series and slice differ between studies, and recalculates the interval descriptor against new numbers. None of this is a typo. It is arithmetic and bookkeeping done by hand, under time pressure and volume. A spell checker does not catch it, and a general purpose language model does not reliably catch it either, because it is not a language problem. A single stale prior or a mislabeled descriptor, "increased" when it decreased, can misrepresent treatment response. I do these comparisons dozens of times a day and wanted to remove the manual steps most likely to go wrong.

What You Built: Describe your MVP, tool, or workflow in plain language
A tool that takes a free text oncology report and restructures it into discrete, editable findings, one per line. The radiologist updates a finding with shorthand, for example "now 1.5 cm, new location," and the tool regenerates that finding correctly: the previous measurement moves into the prior slot, the image reference is reformatted, and the interval descriptor is recalculated from the numbers. The prior handling follows explicit clinical logic rather than blind text replacement. With one prior, the previous value is treated as obsolete and replaced. With two or more, the tracked chain is preserved because that pattern means someone is following the lesion deliberately. Input is the current report text plus the radiologist's shorthand updates. Output is a corrected, comparison ready report. The rule set lives outside the code so it can be edited per subspecialty.

Tech Stack & Development Approach: What tools, frameworks, or methods did you use to build it?
Built as an application using a large language model API for language generation, paired with a separate deterministic rule layer that handles the measurement logic, prior counting, unit normalization, image reference formatting, and interval descriptor calculation. The design choice is deliberate: the language model

drafts, and rule based code checks and enforces the parts that must be correct, rather than trusting the model to get the arithmetic right. The comparison rules are stored as editable configuration outside the application code, not learned from data, so they can be adjusted per subspecialty. Developed independently, on my own time, over evenings and weekends, iterating against real report formats. No institutional data or resources were used.

Validation & Evidence: Have you done any informal testing or validation? If so, what did you find?
Informal only. In a single case test, updating a five structure comparison went from roughly thirty five manual actions to about ten, with the largest reduction in mouse clicks, which are the actions that pull attention away from the images. That is one test on one case by one person, not a time motion study, and I am treating it as a directional signal, not evidence. Functionally, the tool handles the prior logic and descriptor recalculation correctly on the synthetic cases I have run, including several complex cases.

Feedback You're Seeking: What specific feedback, help, or collaboration are you hoping to get from the CAIMI26 community?
Two things. First, help define the right endpoint to validate this. The candidates are reduced error rate, time saved, and downstream concordance in management decisions, and I want input on which is most meaningful and feasible to measure in practice. Second, potential clinical partners in longitudinal reporting that could help test it on a larger, de-identified set of real comparisons. I would also value discussion on a design question with governance implications: who should own the rule set in a department, the individual radiologist, the subspecialty division, or the institution, since shared rules mean shared reporting language. Finally, I am curious whether others have applied the same pattern, a deterministic validator auditing a probabilistic generator, to other reporting tasks.

Known Limitations & Honest Failures: What isn't working yet, or what have you already tried that failed?
No PACS or dictation system integration, which is the single biggest gap. Right now it is a separate window, and the realistic path to adoption is embedding it in the dictation workflow, ideally a right click action inside the existing report rather than another application to switch to. That integration almost certainly requires vendor cooperation or API access I do not currently have, and I expect it to be the hardest part. No validation study yet, so the efficiency numbers are anecdotal. It is single user with no persistence or deployment infrastructure. The language model can still produce awkward phrasing, which is exactly why the measurement and comparison logic is handled deterministically rather than left to the model. I have not yet stress tested it against messy or nonstandard report styles, where I expect it to struggle.

Potential for Impact: If this MVP were fully developed and deployed, who would benefit and how?
If fully developed and deployed, radiologists doing high volume oncologic comparison would spend less time on manual bookkeeping and make fewer transcription and interval errors, while keeping their eyes on the images instead of navigating text. Referring oncologists would get more reliable interval language, which directly informs whether a therapy continues or changes. Patients benefit because response assessment is a real decision point in cancer care, and a stale prior or wrong descriptor can distort it. Trainees and international medical graduates would benefit from consistent reporting language and structure. At the department level, an editable shared rule set could standardize comparison language across a division, and the broader pattern of pairing a language model with a deterministic validator could apply to other structured reporting tasks beyond oncology.