

OncologyBench

Benchmarking multiagent AI systems on grounded systems in oncology research and treatment

Joshua Badshah^{1*}, Birendra Kumar Saye^{1*}, Arjun Banerjee^{1*}, Shashank Shrishrimal^{1*}, Ramesh Ramasamy¹, Dillon Flood¹, Kenny Workman¹

¹LatchBio, San Francisco, CA, USA

*These authors contributed equally.

Correspondence: kenny@latch.bio

Abstract

Agentic AI is currently being deployed across oncology by patients, caregivers, researchers, and healthcare providers to compress analysis timelines, increase decision confidence, and provide reasoning support. The promise of agentic AI is clear, but general deployment requires trusted evaluation benchmarks that mimic the clinical journey. We introduce OncologyBench, a verifiable benchmark of 25 multistep-oncology evaluations that tests the ability of 31 AI model-harness combinations. Each evaluation is built from a set of deterministically graded analytical questions (*Tasks*), that form the input data to make a final clinically relevant conclusion (*Decision*), in prose, that is graded by a rubric. The data used for these evaluations is from publicly available manuscripts, case studies, or databases. In addition, we utilized a consented patient case study data for the express purpose of benchmarking. Evaluations span multiple decision categories that may occur along a clinical oncology journey. To replicate this, multiple agents work through the *Tasks* in parallel and their outputs are provided to a final agent to reason over and make a *Decision*. Our results indicate, highest performance scores for Opus 5 & 5.5, Sonnet 5.5 at 29–40% and followed by Kimi K3 at 24% among the 31 AI model-harness combinations that we tested.

We found that the models with the strong interpretation and analysis skills still showed weakness in reasoning. In a separate experiment we found decision and answer quality varied depending on the persona of the question asker. With the known limitations of our work, we believe our current Oncology Benchmark effectively measures the performance of these frontier AI models being utilized by different personas.

Introduction

Research, diagnosis, treatments, and outcomes in oncology are driven by reasoning across integrated analyses and interpretation of multiple discordant data types. Importantly under the constraints of a resolved clinical journey, decisions are often made under time and economic constraints, in silos, and with incomplete information. A simple workflow for a cancer diagnosis may be built on analysis based on blood screening, imaging, pathology, and sequencing. These data types are typically handled, analyzed, and interpreted by their own experts, while an integrated decision is made by another expert. A translational researcher might enlist the expertise of a molecular pharmacologist, a tran-

scriptomic analysis, and a spatial report before recommending a biomarker strategy to aid patient stratification for an upcoming trial. The ultimate decisions that these experts make are typically a synthesis of multiple discordant data analyses that often contradict each other. These experts may or may not have direct interactions, and experts can rarely transfer full context to others, often constrained by time, region, and regulation. In the creation of OncologyBench we hoped to more closely replicate the integrated environment that characterizes modern oncology practice, including regulatorily constrained and siloed data analyses and interpretations. To this end, we stepped away from the typical singular evaluation structure that we have previously published. Each of the evaluations we built requires a series of parallel data analyses and interpretation allowing an agent to reason around

a final decision. Each evaluation is built from a set of deterministically graded analytical questions (*Tasks*), that form the input data to make a final clinically relevant conclusion (*Decision*), in prose, that is graded by a rubric. We believe that this structure provides a robust framework for evaluating the performance of agentic AI that would be used by real people in a modern oncology setting. The use of agentic systems in oncology is growing and a robust, trusted, framework for categorizing their performance and failure modes will provide a deep understanding of current capabilities. A fulsome understanding is important for researchers, clinicians and patients to ensure the highest level of care.

Our group has developed multiple benchmarks that have been released to evaluate agentic performance on biological tasks. First, SpatialBench and scBench were released to evaluate performance on deterministic spatial and single-cell transcriptomic tasks, respectively [1,2]. Growing from this foundational work, EpiBench [3] and VariantBench [4] were released for epigenomic and variant analysis, respectively. This concept was expanded further in SpatialBench-Long and scBench-Long to evaluate agents on long-horizon tasks that require extended analysis from raw data [5,6]. More recently, a suite of benchmarks has been released to evaluate agentic performance on therapeutic development tasks: TxBench-PP [7], TxBench-Antibody Discovery [8], and TxBench-Oligonucleotide Discovery [9] cover preclinical pharmacology, antibody discovery and oligonucleotide discovery, respectively. Although these single-agent benchmarks are built on assay-specific or drug-development-specific deterministic tasks, they form the foundation of OncologyBench.

Other groups have evaluated AI in oncology and across the sciences. MTBBench [10] poses molecular tumor board questions over multimodal and longitudinal patient records, answered in multiple-choice or true/false form, and MEREDITH [11] compared an LLM system’s treatment recommendations with those of a molecular tumor board on ten fictional cases. TerminalBench-Science [12] collects expert-written research workflows, each completed by a single agent and graded deterministically. Its tumor-immune interface task, for example, asks an agent to annotate cell populations in imaging mass cytometry data from pre-treatment melanoma biopsies and to compare immune infiltration around the tumor between patients who did and did

not benefit from anti-PD-1 therapy. HealthBench [13] instead scores open-ended health conversations with laypeople and clinicians against physician-written rubrics, and HealthBench Professional [14] applies the same approach to real clinician chats, including in hematology and oncology; neither requires analysis of raw data.

OncologyBench builds on the foundation set by our previous benchmarks and expands it to a multiagent, decision-centric framework built to replicate the discordant data, the collaboration between experts, and the high stakes of a clinically relevant oncology decision. It requires agents not only to analyze and interpret datasets with varied inputs, but also to synthesize those analyses into final decisions with defensible reasoning under realistic conditions.

Benchmark Construction

OncologyBench evaluations were constructed from independently re-deriving consequential conclusions from both publicly available oncology data sets and one patient’s data that was donated for the express purpose of benchmarking. Specifically, the open sourced scientific literature including research papers, reviews, and case studies were used to identify deterministic tasks and decision points for the majority of the benchmark evaluations (Figure 1a). Public data sets including TCGA [15–18] and AACR Project GENIE [19], most widely known for their pan-cancer genomic resources, DepMap/CCLE [20] and Cell Model Passports/GDSC [21,22] for functional and pharmacogenomic profiling, Beat AML [23] and MMRF CoMMpass [24] for deep longitudinal characterization of AML and myeloma, CPTAC [25] for proteogenomics, HCMI [26] for patient-derived model characterization, and TCIA [27] for radiologic imaging, were used to rederive original findings and provide substrate for agentic workflows spanning genomic, transcriptomic, pharmacogenomic, proteomic, and imaging modalities across both broad pan-cancer and disease-specific contexts (Figure 1b).

Additionally, a subset of evaluations were constructed from patient data that was donated to Latch for the express purpose of benchmarking agentic workflows. This data proved extremely valuable as a cornerstone example of messy, real-world data.

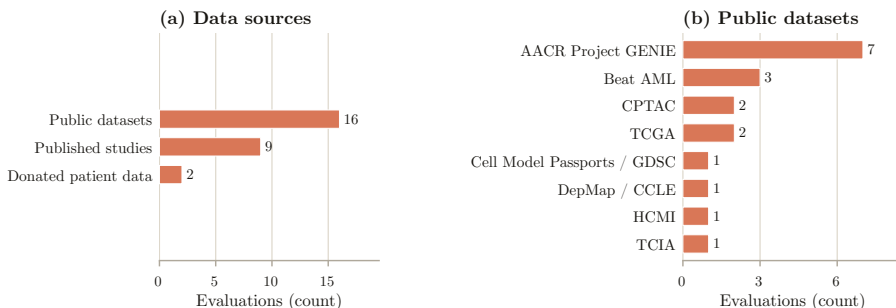


Figure 1: Data used in benchmark construction. (a) Data source types that were used to create evaluations. (b) Public datasets that were used to create evaluations.

This set of evaluations was built to represent a useful evaluation of the current state of model capacity in multi-agent decision based oncology work. Independent experiments were also derived by expanding one specific evaluation on two axes (1) the setting (vide infra) and (2) model capability. These specific experiments were extracted and run to explore the impact of the persona of input user on multiagent performance, and to evaluate the level of information quality presented to a user across free and paid LLM tiers.

These evaluations were built to evaluate some of the multiple professional archetypes that practice oncology. Some of our evaluations are built around a single patient’s data while others require analysis across a cohort of patients, and others reason across population level data sets. These various takes on evaluation construction allow us to benchmark the multiple ways data is utilized across an oncology practice and research organization. Alongside multiple patient data types, we also explored multiple settings for which the evaluations were centered on. This context was provided to set the stage and stakes for the agent’s work and

included settings such as an attending oncologist, a molecular tumor board, or a translational researcher.

Benchmark Anatomy

Individual evaluations that make up OncologyBench are mapped to a typical clinical flow. This mapping allows us to build targeted evaluations across the various stages of a representative oncology diagnosis and treatment. These evaluations can be mapped to a primary stage that we defined as S0–S5 along this flow (Figure 2a). While all evaluations map to a primary stage most of them also map to a secondary stage, usually adjacent to its primary stage. Although we worked to capture all the stages that we defined, some stages are over-represented relative to others. Treatment selection is the most represented stage of the clinical flow because of high consequence to clinicians and patients. Translational and Program is the second most represented as it encompasses a variety of early stage basic and translational research that is enabling a clinical outcome or decision.

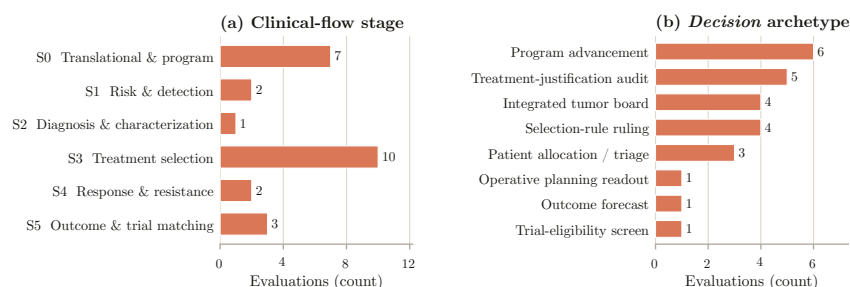


Figure 2: Benchmark anatomy. (a) Mapping of evaluations to stage in a clinical workflow. (b) Mapping of evaluations to a decision archetype.

Furthermore these evaluations can be decomposed and sorted on the decision archetype that they necessitate (Figure 2b). These types of decisions can be calls that a clinician, a molecular tumor board, or a translational researcher would likely make. Each decision is built from a synthesis of the provided data and sub-analyses. Program advancement is the single most represented evaluation type due to its high consequence to drug developers, translationalists, clinicians, and ultimately patients.

Although the *Decision* is the capstone of the evaluation, each of these decisions are built on a foundation of deterministic single-data-type tasks. The benchmark is made up of 25 evaluations

each supported by between 1–7 of these *Tasks*. These *Tasks* cover a wide variety of the assays and analysis that would be encountered in a clinical oncology setting. Examples of these skills can be broken into families that include calling somatic variants, clinical record reasoning, statistics, and spatial imaging along with others (Figure 3). Overall there are 92 *Tasks* that cover 35 distinct assay types represented across nine skill families. Importantly, evaluation *Decisions* in each stage of the clinical flow, in all cases, test a diverse set of skill sets and assay types in their component *Tasks* (Figure 4). This architecture is a result of an oncology question being inherently multimodal in required data input.

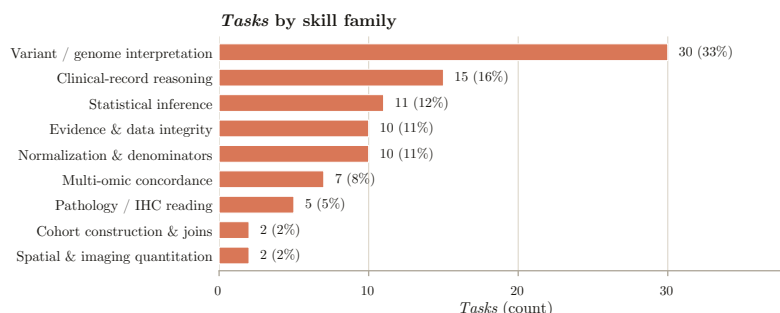


Figure 3: Tasks by skill family. Mapping of tasks that make up an evaluation to the skill/assay families they test.

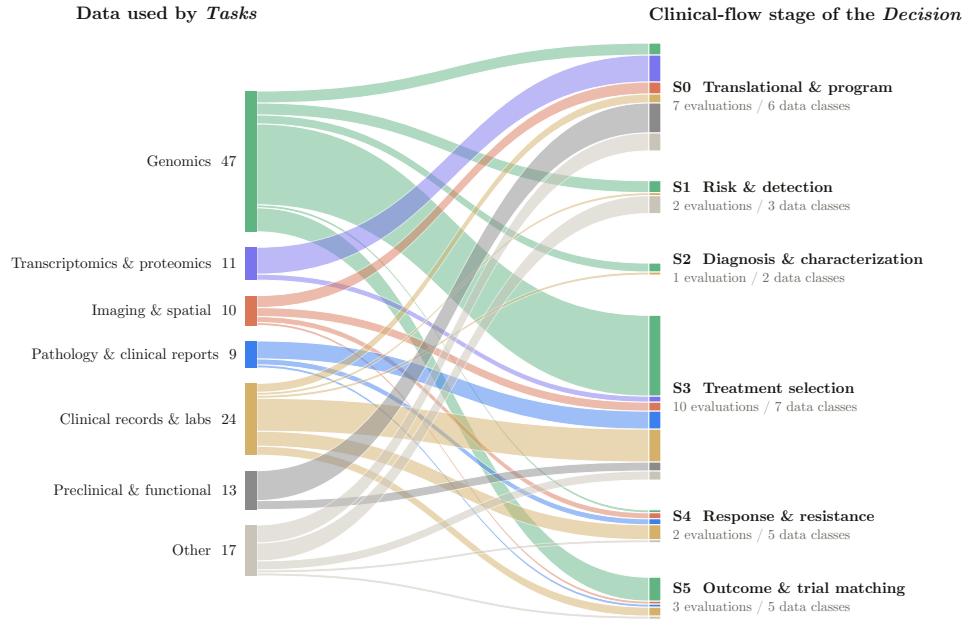


Figure 4: Data used by Tasks in each evaluation. Data types required for evaluations mapped to the clinical flow of the evaluation decision.

Evaluation Design and Scoring

Our previous benchmarks have been built from a collection of evaluations that test analysis and interpretation of single datasets across single assays (Figure 5a). In this benchmark we expanded this idea into parallel single domain tasks that guide a reasoning based decision (Figure 5b). Examination Tasks are single, deterministic analyses built around one data type and one method, each graded against a fixed, machine-checkable answer. *Deci-*

sions, by contrast, evaluate reasoning that is derived from a set of intermediate Task interpretations. A *Decision* is an answer in prose in which the agent must weigh the Task outputs feeding it, resolve any discordance between them, and commit to a single clinically or translationally meaningful conclusion. Because a defensible *Decision* could admit more than one valid line of reasoning, and because the same underlying evidence can be phrased correctly in many different ways, decisions are graded by rubric rather than by exact match.

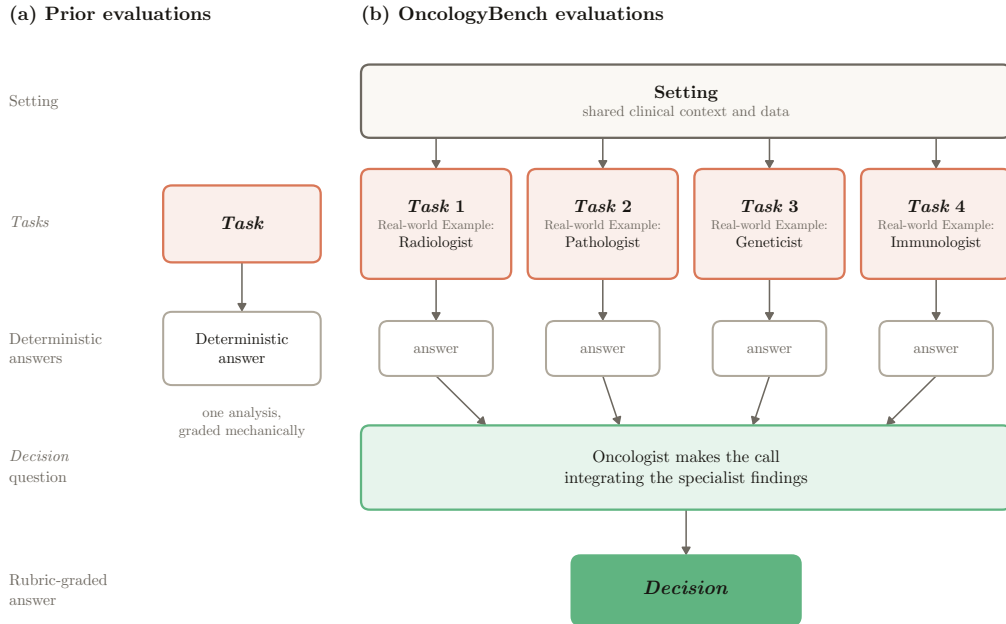


Figure 5: Evaluation construction comparison. (a) Information diagram from previous short horizon evaluations, (b) information flow diagram that governs OncologyBench Evaluations.

Concretely, each evaluation is composed of the following pieces, a *Setting*, a *Data Node*, a number of *Tasks*, a *Decision*, and a *Grader*. The *Setting* sets the scene for the entire evaluation and is passed to each of the *Tasks* and the *Decision*. An example setting could be that *you are a translational researcher assessing the strength of a finding in receptive lymph nodes for use as a perioperative risk-stratification signature*. Each *Task* is composed of a *Prompt*, a link to *Data Nodes* that provision the required data, and a pre-specified *Answer* format. Finally for each *Task* the *Decision* is provided the *Answers*, the *Prompt*, and the resultant *Trajectory*.

This structure is deliberately distinct from the long-horizon evaluations from SpatialBench-Long and scBench-Long, in which a single long-horizon task chains multiple analyses in series and each step depends only on the one before it. Each of these *Decisions* instead sits downstream of several parallel *Tasks*, typically three to five but ranging as high as seven. These *Tasks* exist as one per independent data type or analysis, so the agent must reconcile discordant interpretations rather than carry a single reasoning forward. This structure was created to mimic the path an oncology journey may take, many times requiring input and analysis from multiple experts to arrive at a single decision.

Each rubric pairs positive criteria, which reward the reasoning a defensible decision requires, with penalty criteria for specific errors, such as overstating what the data show or relying on evidence the record does not support. In several evaluations the conclusion is that the available data cannot settle the question. There, the rubric rewards an agent that says so and names the evidence that would settle it, rather than one that forces a commitment the record cannot support. For example, a HER2 immunohistochemistry score of 2+ in breast cancer is equivocal and cannot by itself show whether the tumor is HER2-positive, which is what determines eligibility for standard HER2-targeted therapy. A clinician would not start or rule out that therapy on this result alone but would order in situ hybridization to test for gene amplification, and the rubric rewards an agent that does the same rather than treating the first result as settled. This construction allows us to test the agent reasoning in this decision setting where correct outcomes can be correct along different lines of reasoning.

In the current release, each rollout is scored as a binary pass or fail against its evaluation’s grader, and topline results are reported as pass rates across rollouts. In this rollout a failure of any single *Task* or the *Decision* results in an evaluation fail. The underlying results, including their separate *Task* and *Decision* components,

are recorded for every rollout, and we are evaluating their use as a continuous secondary metric that can distinguish a near miss from a complete failure. Scoring the two components separately also serves a diagnostic purpose. It separates an agent that analyzes correctly but decides poorly from one that reaches a defensible call without having recovered the evidence that should inform it, a distinction a single pass rate cannot make.

Multi Agent Orchestration and Interaction

These evaluations were designed such that each *Task* was completed by a single agent in isolation. In order to furnish the final *Decision*, the *Setting*, and each *Task Prompt*, the *Answer*, and the *Trajectory* were passed to the agent to solve the final *Decision*. We chose to constrain the information flow this way to replicate the way experts in a clinic or research team may interact. Specifically by passing along high level answers as well as reports to the decision maker, while not expecting that final decision maker to perform the underlying analysis again. Although we are cognizant that this is not a perfect representation of a real life interaction, or an idealized multiagent workflow, we believe that this is a useful way to reconstruct these interactions for benchmarking purposes. To this point we also ran standalone experiments to test if different formats of multiagent interactions, with more or less constraints changed the overall scores or behaviors.

Results

The highest mean pass rate for any model / harness pair was 40.00%

Ultimately we tested 6 model families and 4 execution harnesses across our benchmark (Figure 6a). Opus 5 / Pi, Opus 5.5 / Claude Code (CC), Opus 5 / CC, and Sonnet 5.5 / CC were the top performing setups, all scoring >30% overall pass rate, followed closely by Opus 5.5 / Pi and Sonnet 5.5 / Pi (29%). After this set, Kimi K3 / Pi (24%) was the next highest performing model, ranking higher than any of the GPT-6 and Grok class models. Opus 5.5 had the highest refusal rate of any model that we tested (Figure 6b). When run with the Pi harness it refused 48% of attempts (36 of 75). Refusal rates dropped when run with the CC harness (27%). These refusals may be dragging down the overall performance of this model in our current study.

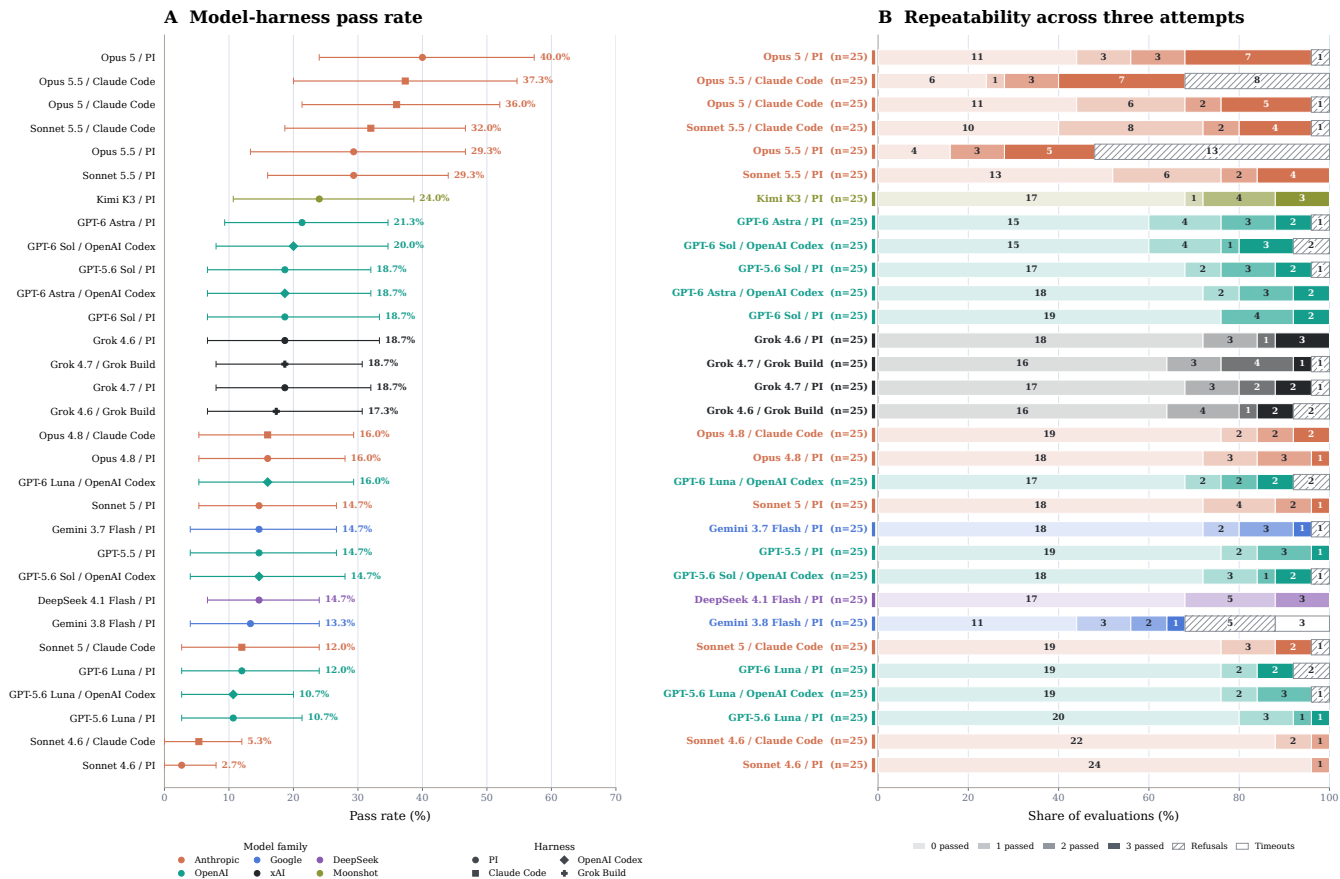


Figure 6: Pass rate by configuration. (a) Mean pass rate with 95% intervals; colour indicates model family and marker indicates harness. (b) Evaluations passed in zero to three attempts, with refusals and timeouts indicated.

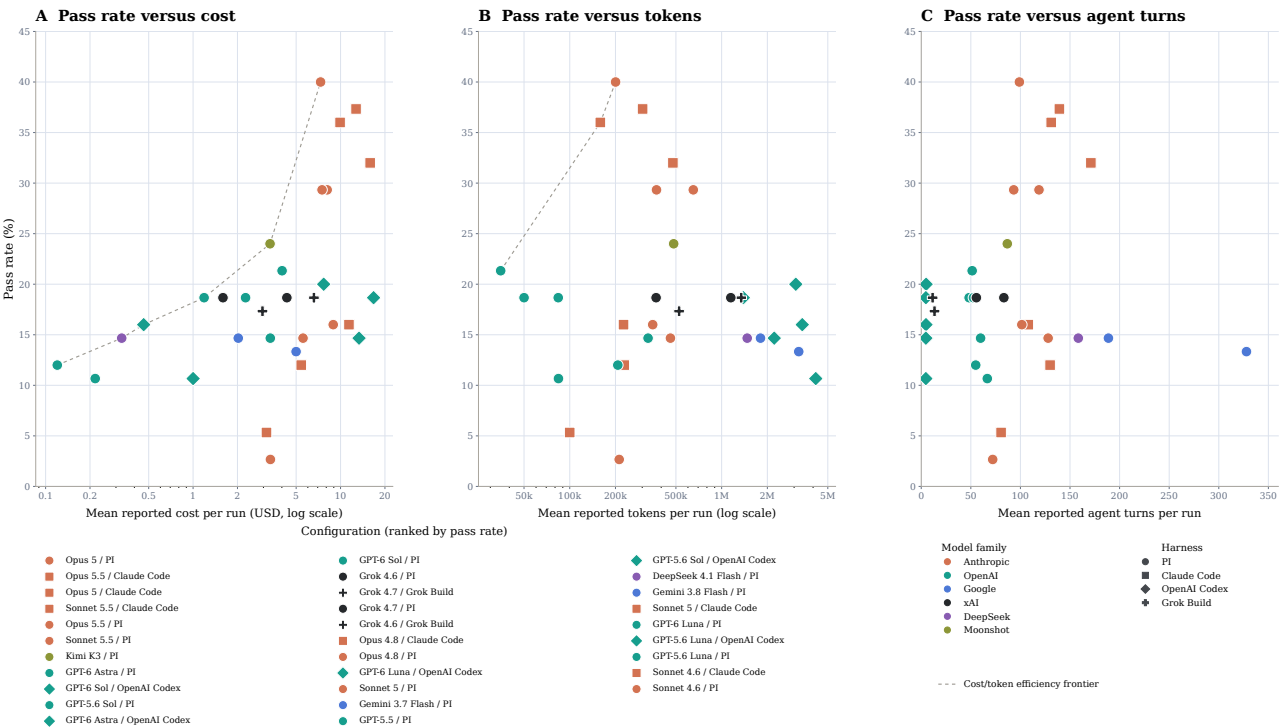


Figure 7: Pass rate against (a) cost, (b) tokens and (c) agent turns per run. The dashed line marks the efficiency frontier.

Cost to run the model impacts performance

Within model classes increased cost per eval typically increased pass rate. In our current study the frontier Claude family of models systematically outperformed other models (Figure 7a). This family tended to cost the most per evaluation and result in the highest pass rates. Unsurprisingly within a model family cost to run a model correlates with pass rate. This trend does not hold between families and is demonstrated by the Pareto frontier spanning GPT, DeepSeek, Kimi, and Claude models.

Token consumption was less correlated with pass rate (Figure 7b). The Claude family of models were fairly consistent in their token consumption and spanned a wide range of pass rates. On the other hand, the GPT family of models displayed opposite behavior and spanned a wide range of token consumption which resulted in a narrow band of performance. Unsurprisingly GPT-6 class models consumed more tokens than earlier GPT-5.6 models.

Agent turns did not correlate with pass rates and agent turn numbers seem to be governed by model behavior (Figure 7c).

Specifically GPT class models typically produce the fewest turns, while Claude class models result in moderate turns. Gemini models result in the greatest number of turns. Interestingly, Gemini’s increased turn count did not result in increased pass rates.

Model performance across Task categories varied with the model family

Opus 5 models perform best overall, with the highest pass rates of the six models on clinical-record reasoning (44%) and pathology/IHC reading (60%) (Figure 8a). In contrast to this behavior Gemini 3.7 Flash underperformed across the board but did better in Pathology/IHC and Spatial/Imaging. Grok, on the other hand, was strongest in normalization and denominators but drastically underperformed in the Pathology/IHC category. Interestingly, Kimi K3 was second only to Opus 5 in four of the nine categories. Comparing models by decision archetype is informative but as some of the categories have limited representation, pass rate comparisons are not significant across the entire set (Figure 8b).

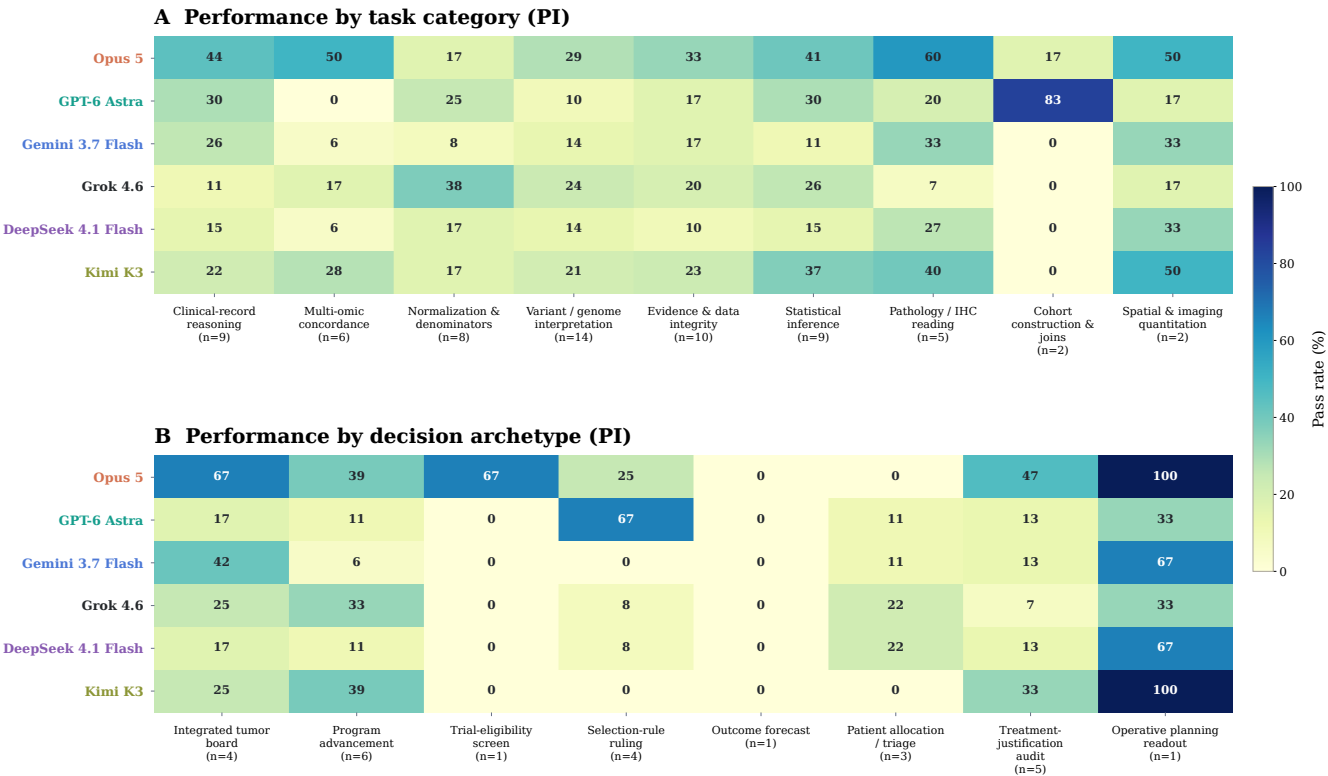


Figure 8: Performance by (a) Task category and (b) decision archetype. Pass rates for six models on Pi; n is the number of evaluations per category.

The gap that is exposed in this benchmark is not analysis and interpretation, but reasoning and decision making

Across the board these models performed well on the individual tasks (Figure 9). Across configurations, the share of attempts with every Task passed ranged from 37% to 69% and exceeded 50% for 24 of 31. The majority of models take a major hit in pass

rate upon applying decisions across the Tasks that it may have solved correctly. Although all models displayed this behavior the gap between analysis and reasoning performance is most pronounced across the Grok models. Interestingly Opus 5.5 displayed the smallest gap in analysis and reasoning performance. When deployed with both the Claude Code and Pi harness Opus 5.5 displays relatively minimal analysis to the reasoning gap.

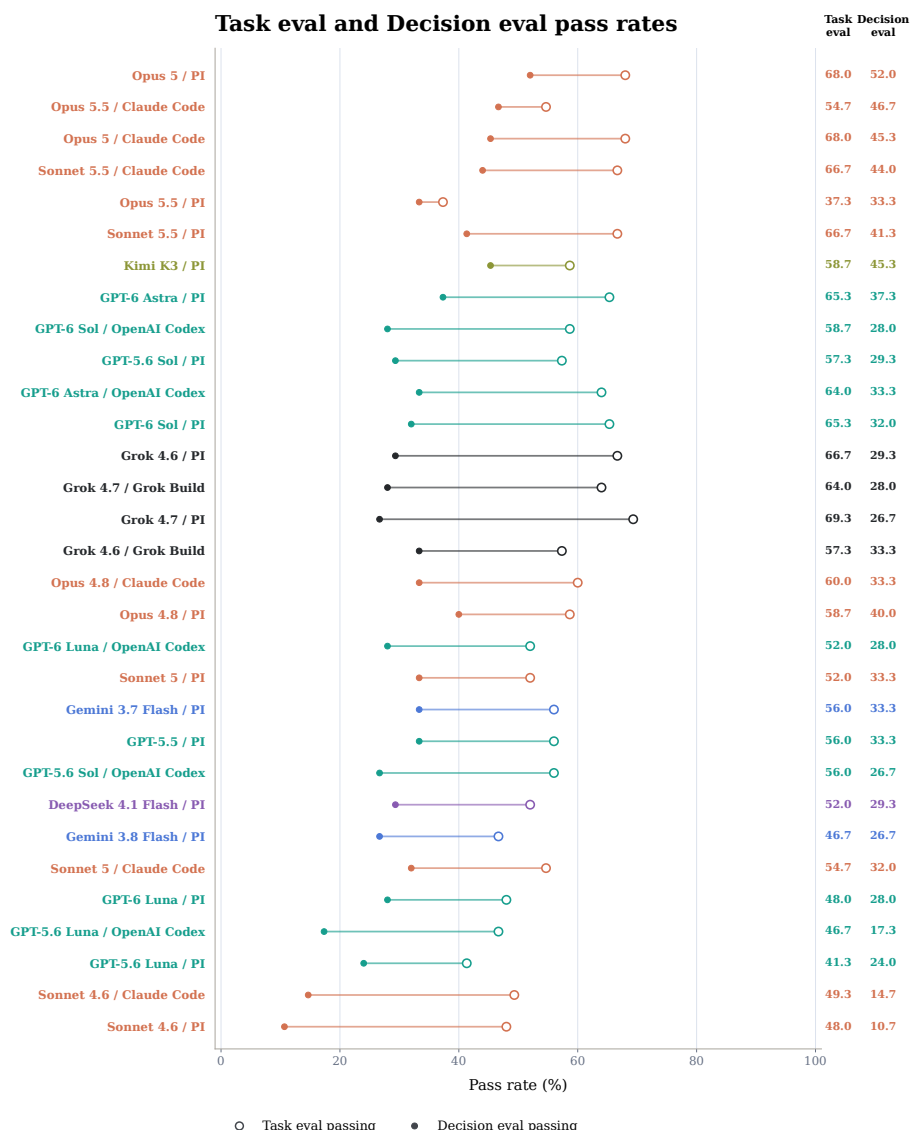


Figure 9: Analysis versus decision. Share of attempts with every *Task* passed (open circles) and the *Decision* passed (filled circles), by configuration.

Persona and setting impact evaluation performance

Cancer data such as imaging, genomic sequencing, biopsy staining, and therapeutic trial suggestions is no longer accessible only to clinical experts. Most patients can now download their own records, and many are beginning to upload them to commercially available models to ask about their condition and treatment options. With this in mind, we constructed a constrained experiment where we asked the same question in six voices: a mother, a father, the patient, a patient advocate, the treating oncologist, and the reporting pathologist.

This experiment uses one evaluation built from a donated pediatric osteosarcoma case, a 10-year-old boy with lung relapse whose resected tumor and saliva were sequenced. The *Setting* gives the agent the vendor report and its tables, raw exome data

from the tumor and the saliva, and tumor RNA. Four parallel *Tasks* are each completed by a separate agent: tumor suppressor status, the origin of the TP53 change, HLA genotype and loss in the tumor, and tumor ploidy with copy number. The *Decision* agent receives the four *Task* answers and prompts, but not the data, and replies to the person asking with one of three decisions: start radiation now, hold it for a MAGE-A4 trial, or cannot decide. The correct advice is to start radiation now. Comparing tumor with saliva shows that the TP53 change arose in the tumor, which satisfies the hospital’s rule for proceeding, and loss of HLA-A*02:01 makes the trial too uncertain to wait for. We crossed four models (Opus 5.5, Sonnet 5.5, Grok 4.7 and GPT-6 astra), two harnesses and the six voices, running 20 rollouts of each combination. Only the Claude models passed. Opus 5.5 passed 83% of its rollouts and Sonnet 5.5 75%, while Grok 4.7 and GPT-6 astra never passed for any asker (Figure 10).

	Mom	Dad	Patient	Advocate	Oncologist	Pathologist
Opus 5.5	88% 35/40	85% 34/40	83% 33/40	73% 29/40	78% 31/40	93% 37/40
Sonnet 5.5	73% 29/40	63% 25/40	78% 31/40	63% 25/40	90% 36/40	85% 34/40
Grok 4.7	0% 0/36	0% 0/39	0% 0/35	0% 0/38	0% 0/36	0% 0/37
GPT-6 astra	0% 0/40	0% 0/40	0% 0/40	0% 0/40	0% 0/40	0% 0/40
	Family askers				Expert askers	

Figure 10: Persona pass rates by model and asker. Passed/scored rollouts beneath each percentage.

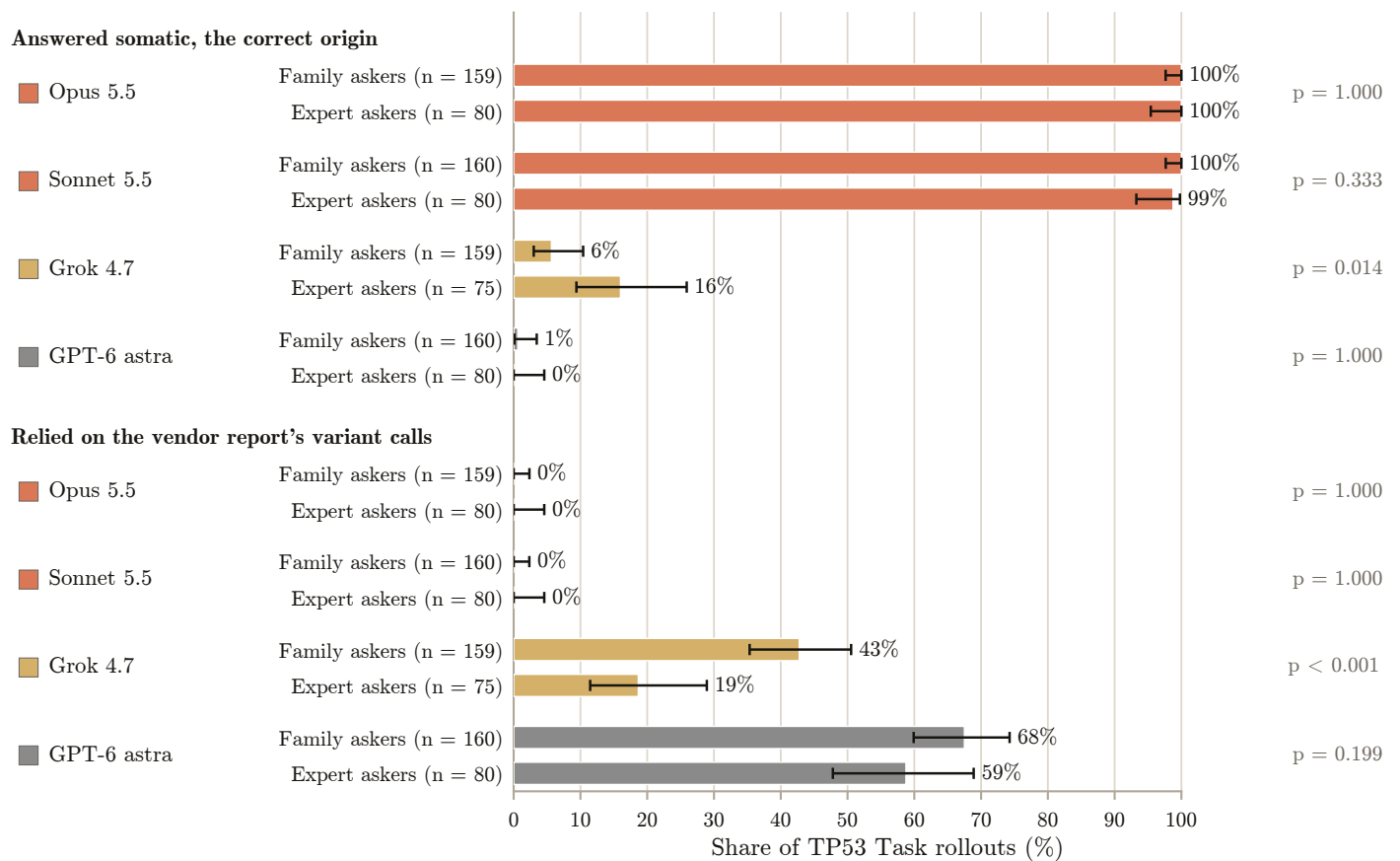


Figure 11: The TP53 Task, family versus expert askers. Answers that called the change somatic (top) and answers that relied on the vendor report (bottom), by model. An answer relied on the report when it cited the report's variant calls and did not call the change somatic. Whiskers show Wilson 95% confidence intervals; p values are from two-sided Fisher exact tests within each model.

The models failed at different *Tasks*, and the *Decision* inherited each failure. Because the *Decision* agent never sees the data, its advice followed the TP53 answer in almost every rollout. GPT-6 nearly always concluded that the origin of the change could not be established, and so never reached the correct advice. Grok usually read TP53 as unchanged. That also gives no reason to hold radiation, so Grok often reached the right advice on a false premise.

The failing models erred in different ways. All four models looked at the relevant data and recognised the TP53 fusion, but drew different conclusions from it. Opus and Sonnet compared the tumor with the saliva, saw that the saliva retains both copies of the region the tumor had lost, and correctly called the change somatic. Grok examined the same evidence just as thoroughly, but judged TP53 by point mutations alone and accepted the vendor report's statement that it was unchanged, a failure of

interpretation rather than effort. GPT-6 did far less work, making about a third as many tool calls as the other models and opening the decisive allele table in fewer than half of its rollouts. It insisted on a form of proof the sequencing could not provide and left the question unsettled.

For Grok, the asker changed the analysis itself. Grok reached the correct TP53 answer more than twice as often for expert askers as for family askers (16% against 6%, $p = 0.014$). With family askers, it was more than twice as likely to defer to the vendor report (43% against 19%, $p < 0.001$). The other models analysed the case the same way whoever asked. Opus 5.5 and Sonnet 5.5 called the change somatic in 478 of their 479 rollouts and never deferred to the report. GPT-6 astra called it somatic in 1 of 240 and deferred to the report about as often for family as for expert askers (68% against 59%, $p = 0.20$; Figure 11).

For the other models, the asker shaped how the reply was written. One of our rubric criteria asks whether the reply suits its reader. This could be plain language and questions to bring to the oncologist for families, words a child can follow for the patient, and precisely stated evidence for experts (Figure 12). Opus met this criterion for nearly every asker, and roughly tripled its density of technical terms from the child to the pathologist. GPT-6 wrote to families in almost the same register it used for experts, failing the criterion for the child in every rollout. Opus went from 3.1 technical terms per 100 words for the child to 9.4 for the pathologist, and Sonnet from 3.9 to 10.4, whereas GPT-6 used 7.2 for the child and 11.2 for the pathologist, writing to families in nearly the register it used for experts. Grok adapted well for the mother but almost never for the father, even though both received identical instructions.

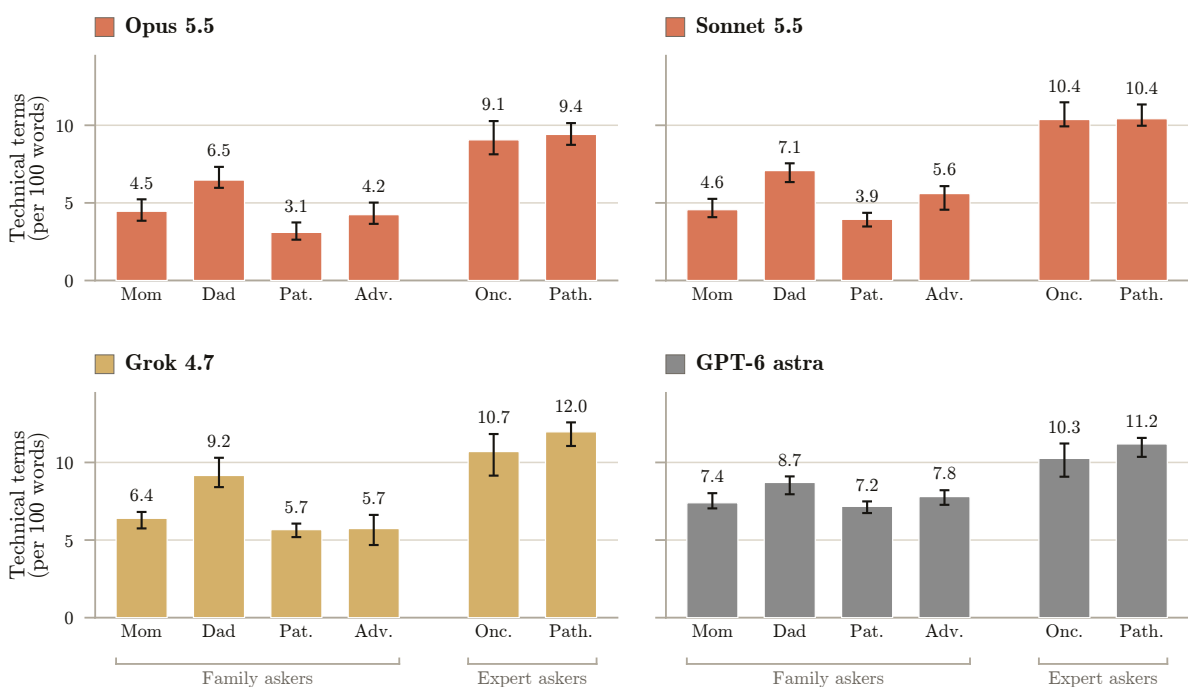


Figure 12: Technical language in the *Decision* reply, by model and asker. Mean technical terms per 100 words for family askers (mother, father, patient (Pat.), advocate (Adv.)) and expert askers (oncologist (Onc.), pathologist (Path.)). Error bars show 95% confidence intervals.

Discussion

Clinical and translational decisions in oncology are rarely made by one person working from one dataset. They are built on several analyses, each carried out by a specialist on a given data type, whose reports are then weighed by someone who does not typically repeat the underlying work. OncologyBench is built to benchmark agents working in that setting.

Each *Task* isolates one analysis with a single correct answer, so it can be graded deterministically and reproducibly. Several *Tasks* run in parallel to support a *Decision*, which receives their prompts, answers, and trajectories but not the data. This was modeled after a clinician receiving specialist reports rather than raw files. The

Decision is graded by rubric, because the right way to combine the evidence varies from case to case and can be argued along more than one line. Sometimes one assay may be inconclusive and the agent must then shift its weight onto the remaining analyses. Sometimes a single finding decides a decision and the others only refine it, and sometimes every analysis supplies a different part and none can be dropped. A benchmark built this way tests whether an agent can weigh independent and sometimes discordant evidence, not only whether it can compute. Scoring each *Task* and the *Decision* separately allows failures to be traced either to the analysis that produced it or to the reasoning that followed.

OncologyBench revealed that models with strong analysis and interpretation performance still demonstrate weakness in reasoning. The evaluation framework within OncologyBench

presents two challenges: a set of deterministic *Tasks* that evaluate analytical computation on a specific dataset against a fixed ground truth, and a clinical *Decision* that tests prose-based reasoning against a standardized rubric. Because an evaluation requires all components to succeed to achieve a passing grade, aggregate pass rates alone cannot isolate specific failure modes. Crucially, analytical and reasoning capabilities can diverge in both directions. A sound clinical argument can rest on flawed data analysis, or an accurate computation can be misapplied during the final decision-making process. By independently scoring each *Task* and *Decision*, the benchmark isolates agents that analyze well but reason poorly from those that synthesize defensible conclusions from flawed premises.

A sample evaluation that audits a hospital’s transplant-referral rule for acute myeloid leukemia against the most recent risk guidelines shows why analysis and decision must be scored separately. Agents had to decide whether a transplant-referral rule still holds under the updated guidance, drawing on the analysis from three analytical *Tasks*, and for the *Decision*, both failed independently of each other. Rollouts that missed a guideline exemption in the first *Task* still wrote convincing recommendations, because the wrongly reclassified patients resembled the high-risk group. Seventeen of them passed the *Decision* while referring 22 favourable-risk patients for transplant. Others analyzed correctly but over-read a valid finding and abandoned genetic risk altogether. A rubric on the final recommendation alone would have rewarded the first error and could not have traced the second to reasoning rather than analysis.

In other cases every *Task* was answered correctly and the *Decision* still failed, because reasoning presented a gap for these models. In one evaluation, agents choose what a blood-based cancer-screening program should fund next. All 87 rollouts with every *Task* correct made the right investment, but 65 of them failed the *Decision*. In this case they rejected the alternative, a lower detection threshold, without quantifying the false positives that would appear in the low signal samples. Here the analysis supplied everything the *Decision* needed but reasoning skills were lacking.

The same pattern held across tasks in the benchmarks. Only 19% of rollouts passed every component. Importantly success at the *Task* level was a poor guide to success at the *Decision* level. Two-thirds (67%) of rollouts that completed every *Task* correctly still failed the *Decision* pointing to a gap in reasoning performance, not analysis. This fact is emphasized by the fact that most frontier models passed the *Tasks* at a high rate, while the *Decision* is the place where models separated most. Opus 5.5 lost 28% of its correct analyses at the *Decision*, whereas GPT-6 astra and Grok 4.7 lost 69% and 72%. Analytical accuracy and decision reasoning are therefore distinct capabilities.

Model performance increased with the cost to provision that model. While this may be an obvious finding, we believe it is important to highlight. As LLM usage becomes ubiquitous, more people are turning to commercially available models to provide insight, a second opinion, or serve as a primary health profes-

sional. In our observations the more the model costs to run, the higher quality the decisions become. We believe this is important as it may contribute to a health information quality disparity between people using high cost and low / free models in high-stakes health decisions.

The persona experiment was built around a situation that is fast becoming routine. A patient or family member holds the same reports and raw sequencing data as the clinician, and asks a commercially available model what they mean. Only Opus 5.5 and Sonnet 5.5 recovered the decisive evidence reliably, and three findings stand out.

Separate scoring exposed false positives, because the decision agent sees only the task answers, its call followed the TP53 task in almost every rollout, right or wrong. Judged on the decision alone, Grok would have been credited in 61% of rollouts. Yet 121 of those 134 calls rested on an incorrect task, usually a TP53 answer that Grok’s own tumor-suppressor task contradicted. A single pass rate cannot separate these false positives from a correct call reached on correct evidence; scoring the tasks and the decision separately can.

The failing models broke down at different steps. Grok’s analysis was thorough: it examined the tumor and saliva data as often as Opus and identified the fusion in every rollout. It then misinterpreted what it found, judging TP53 by point mutations and deferring to the vendor report’s claim that the gene was unchanged. GPT-6 analyzed too little. It demanded proof the sequencing could not provide and overlooked the comparison the data did allow, so its decision to hold applied the hospital’s rule correctly to an incorrect premise.

The asker shaped both the reply and the analysis. The Claude models adapted their language to each reader, from plain words for the child to precise evidence for the pathologist. GPT-6 wrote to families almost as it wrote to experts, and Grok did not carry the plain language it used for the mother over to the father, despite identical instructions. Interestingly for Grok, the asker also changed the analysis itself. Grok reached the correct TP53 answer more often for clinicians than for families, and leaned more on the vendor report when a family member asked. When families may be the most frequent users, a model whose quality depends on who is asking is a behaviour that should be highlighted.

Limitations

Like any benchmark, ours reflects design choices that define its scope. No single benchmark can capture every task an oncologist faces, and the current work focuses heavily on treatment selection and translational decisions. Risk and detection, and response and resistance, are represented by a smaller set of evaluations and are natural areas to expand for future work. Roughly a third of *Tasks* involve variant interpretation, and adding pathology, spatial and imaging analyses would broaden the modalities covered. Future versions could also include cancers not yet represented, such as prostate, pancreatic, ovarian, melanoma and brain tumors,

and extend hematologic coverage beyond AML and myeloma and pediatric coverage beyond a single patient. The benchmark is intended as a representative sample of oncology work rather than an exhaustive one.

We chose deterministic grading for the analytical tasks because it is reproducible and transparent. The tradeoff is that it favors specific analytical choices, and because an evaluation passes only when every check passes, evaluations with many components set a high bar. *Decisions* are graded by a single LLM grader against rubrics derived from the primary literature, including research papers and case studies, which keeps scoring consistent across tested models. Scores still depend on how rubrics are worded, and repeat grading occasionally differs. Measuring the grader’s agreement with practicing oncologists is a valuable next step because clinical decisions often have more than one defensible answer.

To keep the benchmark openly reproducible, we built it from public, retrospective data. We chose to use open-access research cohorts and published datasets under non-commercial research licenses, rather than controlled-access clinical data or licensed resources. These cohorts are cleaner and less longitudinal than real patient records. Prospective clinical utility is outside the scope of this work. Because models may have encountered the source findings during training, we followed our previous benchmarking work in blinding identifiers, withholding annotations that would reveal answers, and disabling internet access. Some answers may still reflect recall as well as derivation.

Agents worked in an offline, static environment, without access to guidelines, knowledge bases, trial registries, the literature or some specialized bioinformatics tools, which isolates what they can derive from the data in front of them. Each case asks for a single decision which is effectively a simplified stand-in for multi-turn conversations with patients and families. The benchmark also defines the multiagent structure, fixing how each evaluation is split into *Tasks* and what the *Decision* stage sees. Letting agents choose their own division of labor, as real teams do, is a natural extension. Our experiments on persona center on a single donated pediatric osteosarcoma case. Many rollouts per agent-harness pair give strong technical replication for that case, and adding further cases would test how broadly the findings generalize. Taken together, these choices give a reproducible, controlled view of where agentic AI in oncology stands today, and a foundation to build on.

Conclusions

Models that analyzed data equally well differed in *Decision* quality. Opus 5.5 was the only model to perform nearly as well at the *Decision* as it did on the *Tasks*. Others, including GPT-6 astra and Grok 4.7, matched it on the *Task* analysis but lost most of that advantage when asked to commit to a call through poor reasoning and decision making. The most critical step in an oncology journey, weighing evidence and making a decision, is therefore the

one where model capability currently lags. Evaluating models on analytical accuracy alone would overstate their readiness to support real oncology decisions for use by patients, clinicians, and researchers.

Methods

Benchmark composition and data

OncologyBench includes 25 decision evaluations based on published oncology research, experimental and clinical datasets, and individual patient cases. Each decision evaluation includes one or more foundational task evaluations. These task evaluations test an agent’s ability to answer specific analytical questions using provided data, such as detecting a molecular change, measuring an imaging parameter, or interpreting a treatment response. Building on this, decision evaluations determine whether the agent can synthesize this evidence to make appropriate clinical or research decisions, such as recommending a treatment, assessing trial eligibility, or exploring an experimental approach.

Each task evaluation involves an agent-facing prompt, input data, and deterministic grading criteria. These tasks are executed in separate agent sessions. The outputs, such as the agent’s responses, generated artifacts, task prompts, and execution trajectories, are then supplied to a final decision agent, along with the specific inputs for the decision evaluation. The decision agent’s final response is scored using a customized rubric tailored to each evaluation.

Input materials for these evaluations encompassed clinical records and reports, molecular profiles, imaging and pathology data, and experimental measurements sourced from individual patients, clinical cohorts, and experimental models. Data preparation protocols varied by evaluation. For instance, in longitudinal treatment-reassessment cases, records were restricted to information clinically available at the designated review time, whereas certain experimental datasets used remapped identifiers. Throughout preparation, we deliberately retained critical clinical and experimental context, including treatment timing, specimen relationships, assay conditions, and replicate structures, to support robust, accurate interpretation.

Evaluation format and grading

Agents received an analytical or decision question, a structured response specification, and the input materials assigned to that evaluation. *Task* agents analyzed staged data, while *Decision* agents received upstream answers, generated artifacts, task prompts, and execution trajectories. *Task* evaluations were graded deterministically against predefined reference outputs, using exact matches or predicates for categorical responses, reference-set comparisons for set-valued responses, and specified tolerances or ranges for numerical responses. *Decision* evaluations were graded using evaluation-specific AI rubrics that assessed scientific interpretation, integration of upstream evidence, and justifi-

cation of the final decision.

Agent runs and execution

We ran each evaluation three times across 31 model-harness configurations, yielding 2,325 evaluation attempts (over 10,000 agent runs counting each *Task* and *Decision* session). Each configuration paired a language model with the execution scaffold that provided its tools and interaction protocol. The evaluated harnesses included Claude Code, PI, OpenAI Codex, and Grok Build.

We ran models at their maximum available reasoning-effort setting in isolated environments without internet access. Agents could inspect and analyze only the files staged for the evaluation. We retained complete trajectories, including model messages, tool calls, execution outputs, and final responses. We recorded token usage, execution time, and estimated cost.

Outcomes and statistical analysis

We classified each run as correct, incorrect, or unevaluable. A run was correct only if every required grader passed, and a gradable answer that failed any required check was classified as incorrect. We classified runs that were refused, terminated by a platform or execution error or timeout, produced an unsupported answer-only guess, or whose trajectory could not be verified as unevaluable. We detected infrastructure-level refusals via provider or harness signals, while we identified model-initiated refusals from the recorded trajectories. Tool denials that did not prevent a gradable final answer were not counted as refusals. Correct and incorrect runs together constitute the evaluable attempts.

We calculated pass rates over all 75 scheduled attempts per configuration (25 evaluations $\times$ 3 attempts), counting refusals and timeouts as failures, so each attempt contributed equally. Error bars represent 95% percentile confidence intervals obtained from 20,000 bootstrap resamples at the evaluation level, retaining all three attempts within each sampled evaluation.

Data availability

Example evaluations and trajectories will be made available at: <https://github.com/latchbio/oncologybench>

References

- [1] Kenny Workman, Zhen Yang, Harihara Muralidharan, and Hannah Le. SpatialBench: Can agents analyze real-world spatial biology data? <https://github.com/latchbio/spatialbench>, 2025.
- [2] Kenny Workman, Zhen Yang, Harihara Muralidharan, Aidan Abdulali, and Hannah Le. scBench: Evaluating AI agents on single-cell RNA-seq analysis. <https://github.com/latchbio/scbench>, 2026.
- [3] Harihara Muralidharan, Reema Baskar, Soo Hee Lee, Tim Proctor, and Kenny Workman. EpiBench: Verifiable evaluation of AI agents on epigenomics analysis. <https://github.com/latchbio/epibench>, 2026.
- [4] Aashka Bhowmick, Nasim Rahmatpour, Shivang Sharma, Qian Xu, Ananya Gupta, Asmita Lagwankar, Arjun Banerjee, and Christopher Zou. VariantBench: An agentic benchmark for genetic variant discovery and interpretation. <https://github.com/latchbio/variantbench>, 2026.
- [5] Ian Diks, Harihara Muralidharan, Tim Proctor, and Kenny Workman. Verifiable benchmarking of long-horizon spatial biology. arXiv:2605.28065, 2026.
- [6] Ian Diks, Zhen Yang, Arjun Banerjee, Tim Proctor, and Kenny Workman. scBench-Long: Verifiable benchmarking of long-horizon single-cell biology. arXiv:2606.26563, 2026.
- [7] Hannah Le, Ramesh Ramasamy, Alex Urrutia, Mahsa Yazdani, Tim Proctor, and Kenny Workman. TxBench-PP: Analyzing AI agent performance on small-molecule preclinical pharmacology. *arXiv [cs.AI]*, 17 June 2026.
- [8] Ramesh Ramasamy, Hannah Le, Jackson Brougher, Alex Urrutia, Arjun Banerjee, Sean Poust, and Kenny Workman. TxBench-Antibody Discovery: Benchmarking AI agents on experimentally grounded decisions in therapeutic antibody discovery. <https://github.com/latchbio/txbench-antibody-discovery>, 2026.
- [9] Martin Jacko, Jackson Brougher, Alex Urrutia, Hannah Le, Arjun Banerjee, Dillon Flood, and Kenny Workman. TxBench-Oligonucleotide Discovery: Benchmarking AI agents on experimentally grounded decisions in the discovery of ASO/siRNA therapeutics. <https://github.com/latchbio/txbench-oligo-discovery>, 2026.
- [10] Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. MTBBench: A multimodal sequential clinical decision-making benchmark in oncology. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025. arXiv:2511.20490.
- [11] Jacqueline Lammert, Tobias Dreyer, Sonja Mathes, Leonid Kuligin, Kai J Borm, Ulrich A Schatz, Marion Kiechle, Alisa M Lörsch, Johannes Jung, Sebastian Lange, Nicole Pfarr, Anna Durner, Kristina Schwamborn, Christof Winter, Dyke Ferber, Jakob Nikola Kather, Carolin Mogler, Anna L Illert, and Maximilian Tschochohei. Expert-guided large language models for clinical decision support in precision oncology. *JCO Precision Oncology*, 8:e2400478, 2024.
- [12] Terminal-Bench-Science contributors. Terminal-Bench-Science: Evaluating AI agents on research workflows across scientific domains. <https://github.com/harbor-framework/terminal-bench-science>, 2026.
- [13] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. arXiv:2505.08775, 2025.
- [14] Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chi Tong, Kavin Karthik, Arnav Dugar, Akshay Jagadeesh, Khaled Saab, Johannes Heidecke, Ashley Alexander, Nate Gross, and Karan Singhal. HealthBench Professional: Evaluating large language models on real clinician chats. arXiv:2604.27470, 2026.

- [15] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.
- [16] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, Rehan Akbani, Reanne Bowlby, Christopher K Wong, Maciej Wiznerowicz, Francisco Sanchez-Vega, A Gordon Robertson, Barbara G Schneider, Michael S Lawrence, Houtan Noushmehr, Tathiane M Malta, Cancer Genome Atlas Network, Joshua M Stuart, Christopher C Benz, and Peter W Laird. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018.
- [17] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poirson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, Larsson Omberg, Denise M Wolf, Craig D Shriver, Vésteinn Thorsson, Cancer Genome Atlas Research Network, and Hai Hu. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, 2018.
- [18] Allison P Heath, Vincent Ferretti, Stuti Agrawal, Maksim An, James C Angelakos, Renuka Arya, Rosita Bajari, Bilal Baqar, Justin H B Barnowski, Jeffrey Burt, et al. The NCI Genomic Data Commons. *Nature Genetics*, 53(3):257–262, 2021.
- [19] AACR Project GENIE Consortium. AACR Project GENIE: Powering precision medicine through an international consortium. *Cancer Discovery*, 7(8):818–831, 2017.
- [20] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, Robin M Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F J Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A Garraway, David E Root, Todd R Golub, Jesse S Boehm, and William C Hahn. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576.e16, 2017.
- [21] Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies, and Mathew J Garnett. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Research*, 47(D1):D923–D929, 2019.
- [22] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, Sridhar Ramaswamy, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Cyril Benes, Ultan McDermott, and Mathew J Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2013.
- [23] Daniel Bottomly, Nicola Long, Anna Reister Schultz, Stephen E Kurtz, Cristina E Tognon, Kara Johnson, Melissa Abel, Anupriya Agarwal, Sammantha Avaylon, Erik Benton, et al. Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer Cell*, 40(8):850–864.e9, 2022.
- [24] Sheri Skerget, Daniel Penaherrera, Ajai Chari, Sundar Jagannath, David S Siegel, Ravi Vij, Gregory Orloff, Andrzej Jakubowiak, Ruben Niesvizky, Darla Liles, et al. Comprehensive molecular profiling of multiple myeloma identifies refined copy number and expression subtypes. *Nature Genetics*, 56(9):1878–1889, 2024.
- [25] Yize Li, Yongchao Dou, Felipe Da Veiga Leprevost, Yifat Gefen, Anna P Calinawan, François Aguet, Yo Akiyama, Shankara Anand, Chet Birger, Song Cao, et al. Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell*, 41(8):1397–1406, 2023.
- [26] Dina ElHarouni, Mushriq Al-Jazrawe, Seongmin Choi, Merve Dede, Toshinori Hinoue, Sean A Misek, Heeju Noh, Luca Zanella, Yuen-Yi Tseng, Hayley E Francies, et al. A compendium of next-generation patient-derived models for diverse cancers. *Nature*, 657(8132):775–788, 2026.
- [27] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.