Protege Double Blind Evaluation of Oncology Prediction Systems

Standard Model Bio and frontier models

Protege

October 2026

TLDR

Protege assessed Standard Model Bio (SMB) against three frontier models using records from 412 patients with breast and lung cancer. Each system received multimodal patient information, and its predictions on two oncology workflows were scored against ground truth drawn from the electronic medical record packets. Ground truth was also extended to patient records in the North American Association of Central Cancer Registries (NAACCR), where hospitals submit final staging and diagnostic attributes of a cancer.

SMB ran locally on infrastructure hosted by Protege. The frontier models were accessed through their providers' APIs under zero data retention settings. Protege tested multiple workflows and ablations for the frontier models and reported the strongest results for each.

The evaluation was double blind in a specific sense: model builders neither selected nor saw the hidden test set, and Protege conducted the assessment independently of model development, without knowledge of the exact model architecture. SMB had the right to silent exit throughout the study. Upon seeing the results, SMB chose to co-publish them in this report but had no input into the results.

The patient records were embargoed under data exclusivity requirements traced back to the source hospitals. Protege is responsible for data provenance and contractual obligations that guarantee these records were previously unused in model training. In this report, we describe the framework governing these embargoed data, where it is weakened by patients' ability to request and upload their own records to frontier models, and how we aim to address this specific issue by strengthening our embargo framework.

This report describes what we assessed, the infrastructure we used, the steps we followed, and what SMB and Protege decided before and after the assessment. Our findings suggest that SMB across tested cancer types had the lowest reported cost and fastest turnaround. The SMB model needed approximately 2 seconds and less than one cent per prediction. This is compared with 108–166 seconds and \$0.32–\$1.53 for the frontier SOTA models. There were heterogeneous performance results across cancer types. The frontier models scored higher on average in breast cancer, while SMB had higher accuracy, and lower cost and time to prediction in lung cancer.

Broadly this assessment has two main goals. One, to reduce the information gap between what models can do and what users can independently verify through a third-party evaluator with healthcare expertise, and two, to extend evaluation beyond benchmark leaderboards to tasks and workflows that occur in practice. Through independent assessments of vertical AI builders, we aim to connect gains in prediction accuracy, faster inference, and lower execution costs to the economic KPIs of organizations using these systems, such as cost per patient, staff productivity, and care delivery capacity. Establishing those effects, and understanding how wider adoption changes resource allocation and outcomes across the healthcare and health sciences system, remains a goal for future work.
For organizations considering local deployment of the SMB model, these specific results below aim to shed light on the tradeoffs between accuracy, cost, and speed. The practical question is not which model is fastest or cheapest. It is how much additional predictive accuracy an organization gains from a slower, more expensive system, and whether that gain justifies the added resources in its intended workflow. Our Protege Scoring Framework so far measures prediction performance and operating efficiency at a high level; it does not establish what faster diagnosis or better predictions contribute in further downstream general equilibrium outcomes. This is an extension to future work we aim to pursue.

Finally, we acknowledge here that parts of this report were co-authored with AI research copilot tools that assisted in code reading and writing of technical steps.

1 Key findings

We evaluated SMB against three frontier models on two time-to-event prediction tasks using records from 412 patients with breast and lung cancer. The comparison focused on frontier models and did not include other specialized AI developers. We examined three aspects of productivity: prediction accuracy, inference

time, and execution cost.

SMB, a 3-billion-parameter world model, had the lowest reported execution cost and median inference time. Predictive performance varied by cancer type: frontier models had higher mean outcome-ranking scores in breast cancer, while SMB had the highest mean in lung cancer.

We report confidence intervals to show the uncertainty around the performance estimates. We pre-registered the entire assessment internally (i.e., we decided ex ante how to score refusals and other failures).

For organizations considering these AI systems, the results present a tradeoff between predictive performance and the cost and time required to obtain a prediction. The economic value of these differences will depend on how each system is integrated into an organization's workflow, including the value of additional accuracy and whether faster, cheaper predictions translate into useful gains in practice.

Scope of the evidence

The completed comparison covers two workflows: overall survival (OS) prediction and time to next treatment (TTNT) prediction. We also had initially scoped stage prediction, which will be addressed in future reports. Generally, stage prediction is a saturated task for most models. It was ultimately deprioritized as a result.

The task sections below describe the OS and TTNT workflows in the data. For each workflow, we explain how the task was constructed from the patient records and how ground truth was established, providing a basis for evaluating other models on the same workflows.

1.1 Standard Model Bio and its intended uses

Standard Model Biomedicine develops a multimodal foundation model that represents a patient's condition mathematically, using biomedical data as its input. Applications include prognosis, disease progression, trial simulation, enrollment planning, and trial eligibility criteria. These uses extend beyond the outcomes assessed here.[2]

The study evaluated SMB's 3B parameter world model, pretrained on the clinical records of approximately 50,000 patients. A separate linear prediction layer was trained for each task. SMB reported to us that the world model itself was not specifically trained for performance on OS or TTNT.

What is being compared As with most benchmarks, the results reflect the underlying model, the data, and the choices made during implementation. We therefore evaluate each submitted system as a whole, including any fitted components, input processing, prompts, and the steps used to produce a prediction.

Protege's assessment framework aims to make these comparisons under comparable conditions by controlling the inputs, prompts, and prediction setup. Before running the evaluation, we agree on the scoring method and define the test set. We also specify how refusals will be handled, which ablations may be run, and how many runs are permitted. These decisions are made before any results are seen.

The comparison asks how well each system performs on a dataset withheld from the model builders. Our assurance about prior access applies to this specific dataset. We cannot establish that information about the same patients never entered a model's training data through another hospital or source during their care.

To support that assurance, we selected records that the participating health systems and cancer centers had never previously licensed, along with pathology slides that had never been scanned. Protege then arranged for those slides to be digitized through an independent scanning vendor. These provenance controls provide the basis for treating the evaluation dataset as embargoed, while leaving the limits of patient-level verification explicit.

2 Workflow Impact in Oncology

Cancer, often called the emperor of all maladies, accounted for 619,876 of the 3,072,666 deaths recorded in the United States in 2024—approximately one in five.[4]

Oncology decisions depend on information gathered across specialties, including radiology, pathology, laboratory testing, and the patient's clinical history. That information accumulates over a series of appointments, tests, and reviews. Each step can also involve administrative work: scheduling, insurance authorization, transferring records, and coordinating care between teams.

The opportunity for AI is to help clinicians bring this information together and interpret it sooner. If a system can reliably process clinical records and relevant images, it could reduce the time needed to reach an informed decision about diagnosis, prognosis, or treatment. Its practical value depends on whether faster processing actually removes a bottleneck in care.

The time involved can be substantial. A U.S. claims study of 458,818 patients newly diagnosed with cancer in 2018–2019 reported a mean diagnostic interval of 156.2 days and a median of 118 days. The interval began with the first diagnostic test identified by the study's algorithm, or a qualifying preceding office visit, and ended at diagnosis. These figures describe a commercially insured and Medicare Advantage population; they do not establish a national average from first suspicion to treatment initiation.[6]

Treatment delays also carry potential consequences. Hanna and colleagues found an association between each four-week delay and higher mortality across several cancer treatment settings. For breast cancer surgery specifically, the estimated hazard ratio was 1.08, corresponding to an 8% higher mortality hazard per four-week delay. Figure 1 shows selected estimates.[5]

The workflows evaluated here predict overall survival and time to next treatment. Reliable survival estimates could support discussions between physicians and patients about prognosis, downstream treatment, and clinical trial participation. They could also help trial sponsors characterize patient risk and inform enrollment planning alongside clinical eligibility criteria.

Time-to-next-treatment estimates could help care teams anticipate when a patient may need another treatment and prepare for that transition, including reviewing trial options. Predicting this interval, however, does not itself establish that treatment should begin sooner or that a delay can be prevented. The economic value of both workflows will depend on whether their predictions improve decisions and coordination in practice. This comparison evaluates predictive performance, cost, and speed; it does not measure reductions in treatment delays or improvements in patient outcomes.

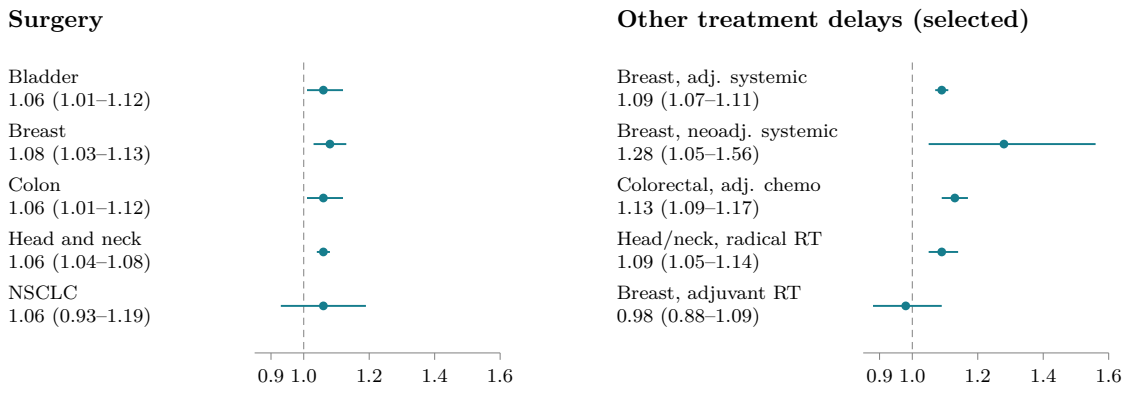

Figure 1. Mortality associated with a four-week treatment delay. Points show hazard ratios; bars show 95% confidence intervals. Both panels use the same scale; the dashed line marks no association. Adj.: adjuvant; neoadj.: neoadjuvant; RT: radiotherapy; NSCLC: non-small-cell lung cancer. [5]

3 Independent evaluation and embargoed data

3.1 Definition of a Double Blind Evaluation by Protege

In a double-blind clinical trial, participants and the investigator or sponsor staff involved in treatment and clinical evaluation do not know which treatment each participant receives. This separation reduces the risk that expectations influence care or the assessment of outcomes. Treatment assignments are protected through controlled randomization procedures and restricted access to allocation codes.[7]

Protege applies a related principle to AI evaluation: separating model development from the selection and scoring of hidden test cases. Builders specify the intended tasks and workflows and submit their systems. Protege controls which patients enter the test set, executes the agreed runs and ablations, holds the reference labels, and scores the predictions. Protege does not design or train models for foundation model providers or specialized AI developers, nor does it perform post-training for enterprise models.

This separation allows Protege to act as an independent evaluator between the model builder and the data provider. It is intended to limit opportunities for "hill climbing," where repeated exposure to a benchmark informs changes that improve performance on that particular test set. Related risks include tailoring a benchmark to a model's architecture or selecting favorable test cases after reviewing results. Defining the test set, scoring rules, and permitted ablations before execution helps constrain these practices. Builders cannot view the protected test records or labels, including after scoring.

Independent evaluation also requires infrastructure trusted by both model builders and data providers. For this study, Protege as an independent third party built the frontier inference containers. SMB ran locally on Protege-hosted infrastructure, while the frontier models were accessed through their providers' APIs under zero data retention settings. These arrangements were designed to protect proprietary model assets and restrict the use and retention of embargoed patient information.

The AI builder's identity remains known to Protege. However, under the right to silent exit, a participating builder's identity remains confidential to the public unless the builder elects to proceed with disclosure. The cost of running the assessment is agreed upon prior to results and is a standard fixed fee (covering compute mainly) independent of the results. To summarize, "double blind" here refers specifically to restricted access to test data and the separation of model development from evaluation responsibilities.

3.2 Development split

Protege provided Standard Model Bio (SMB) with a development split before the benchmark. It contained 200 unique patients: 100 with breast cancer and 100 with lung cancer. None appeared in the test split. Protege determined the allocation of patients to the development and test splits; SMB did not choose which patients or patient characteristics were retained in either group.

The development records came from the same healthcare systems as the test records. The results therefore measure performance on held-out patients from those sources. They do not establish how well SMB would generalize to patients from other healthcare systems.

Access included structured clinical records, clinical notes and reports, pathology whole-slide images, radiology images, and reference labels for overall survival (OS) and time to next treatment (TTNT). SMB was permitted to use these development data as it wished to prepare its system before test-set evaluation.

The development split allows providers to check input processing, inspect prediction errors, compare preprocessing and inference configurations, develop prompts, and fit or select survival prediction components. SMB used the development labels to train task-specific linear probes. The reported results therefore reflect a system with components fitted on the development split and evaluated on a separate, hidden test split.

3.3 Definition and boundaries of the embargo

We use **embargoed evaluation data** to describe specified patient records and reference labels reserved for evaluation under time-limited access and use agreements. Participating healthcare systems and data providers enter into exclusive agreements with Protege that prohibit providing the protected records to any entity for AI training during the embargo period. These agreements designate Protege as the authorized custodian for evaluation and define the permitted uses, execution environments, and release conditions.

The embargo applies to the protected test records and labels. Builders may not use them for training, probe fitting, prompt selection, calibration, or model selection. The separately authorized development split is governed by the development permissions described above.

These contractual controls do not establish that every piece of information about a patient has remained outside all prior model training. A patient could, for example, obtain their records and upload them to an AI service whose terms permit their use for training. Records about the same patient could also exist at another institution outside the agreements governing this evaluation.

Our current standard is that the protected records were not deliberately solicited or supplied for AI training through the participating sources, and that the agreements reserve them for evaluation during the embargo period. This is an assurance grounded in provenance and contractual controls, rather than an empirical guarantee that no model has ever encountered information about these patients.

We plan to strengthen these controls by prioritizing more recent records and, where verifiable, records that patients have not yet requested from the participating sources. These measures may reduce opportunities for prior exposure, but they cannot eliminate every route by which patient information could reach a model. Maintaining the embargo will require continued verification and refinement as data-sharing practices change.

Other Reported Controls for this Evaluation

Model inputs were read-only and versioned. Test labels were hidden from model developers prior to evaluation, and were never on the machine at the same time as model weights or processes during evaluation. Inference was run with full network isolation whenever possible, and with strict network restrictions when not.

4 Evaluation architecture

Protege controls which patient records each AI system receives and keeps the grading answers separate. Predictions are saved before the answers are introduced for scoring. Figure 2 summarizes the assessment process we followed; the separation of records, model runs, and grading follows strict controls we will later report on in detail outside the scope of this brief.

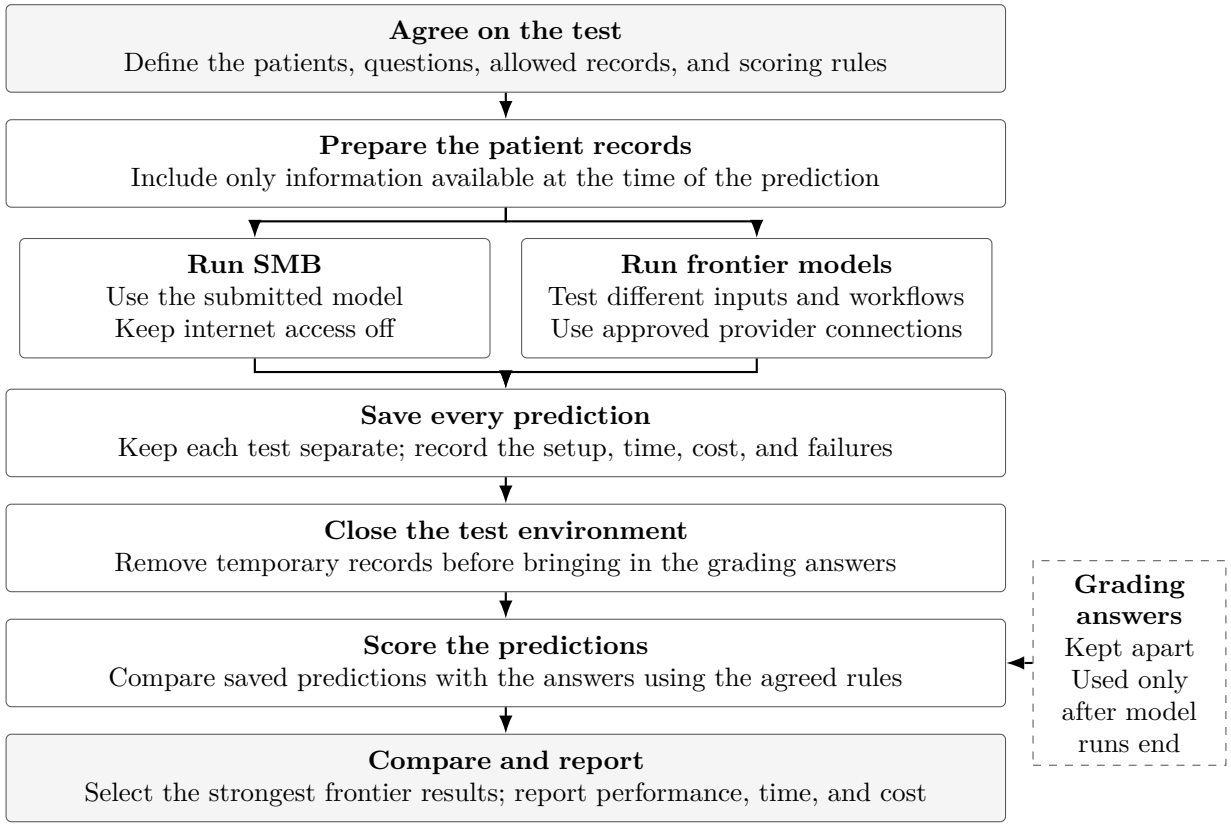

Figure 2. Summary of the Protege Scoring System. Shown here is how we scored SMB (hosted by Protege) against frontier models under zero data retention.

Testing frontier workflows Requires Ablations. These Ablations must also be Controlled

Much of what is reported in the sections below used AI to describe exact code steps and infrastructure choices. Human edits were made for clarity. This does mean that the sections below are a mixture of AI and human interpretation of code workflows.

Protege ran ablations across different frontier-model harnesses and selected the top-performing results for reporting. A harness is the software that presents patient information to a model and collects its answer. Ablations test how changes in the inputs or workflow affect performance. This approach evaluates how well the frontier models perform under the tested setups.

Comparability depends on what each model actually receives. For each model, the proposed API implementation prepares medical images (pathology and radiology), divides large records into batches, and asks the same model to combine its findings. We did several A/B testing studies to determine the best methods to insert the image context into frontier models.

5 Methods

5.1 Task design and prediction timing

The evaluation included two time-to-event tasks: overall survival (OS) defined as time to death, and time to next treatment (TTNT) defined as time to change in line of cancer therapy or death.

For each task, the model received data up to and including the start of the patient's first line of systemic therapy for the indicated cancer. We defined this landmark as the earliest treatment timestamp with line-of-therapy value equal to 1, as indicated by the healthcare providers. For TTNT, the next treatment was defined as the earliest timestamp where line-of-therapy was equal to 2 in the source data. The model predicted time from this landmark to the relevant event.

This design restricted the evaluation to patients who started treatment and excluded patients who died before treatment began.

Both tasks assessed prognosis among treated patients. Neither task was designed to estimate the causal effect of choosing one treatment over another.

5.2 Task methods

The completed evaluation includes 271 breast-cancer and 141 lung-cancer patients. The task definitions below describe the evaluation methods, and Table 1 summarizes the reported system configurations.

5.3 System specifications and adaptation

Submitted system	Reported configuration
Standard Model Bio	Frozen representations and fitted task-specific linear probes; local GPU execution. Inference container with SHA-256 digest prefix `f1518c278fe`
GPT-5.6-sol	API system; maximum reasoning on the source slides
Gemini 3.7 Flash	thinking HIGH (the maximum) through Vertex AI
Claude Opus 5	API system; maximum reasoning on the source slides

Table 1. Names and settings are reproduced as study labels, not independently verified API identifiers.

6 Task 1: Overall Survival (OS)

Question Given information available at the prediction landmark, how well does the system forecast subsequent all-cause mortality?

Landmark event The landmark event is defined as the patient's first systemic line of cancer therapy, as indicated in the source data.

Population and timing All breast and lung cancer patients alive at the landmark are included.

Permitted evidence All declared EHR, pathology, and radiology evidence available up to and including the landmark event is permitted. Eligibility depends on when information became available, not only the date of the underlying clinical event. Later pathology sign-out, report addenda, and retrospectively populated fields must not enter an earlier prediction.

Target and reference label Let T_D be time from the landmark to death and C time to last verified alive follow-up or the administrative cutoff, whichever applies under the frozen rule. The observed label is $Y = \min(T_D, C)$ with event indicator $\delta = 1$ if death is observed by C, and 0 otherwise. A censored observation establishes survival through Y; it is not a death at Y.

A death on day 80 before cutoff day 100 gives $Y = 80$, $\delta = 1$; last verified alive on day 100 with no observed death gives $Y = 100$, $\delta = 0$.

Required output Each model returns a single positive, finite numerical prediction of the number of days from the observation cutoff to death. Larger values indicate longer predicted survival, and smaller values indicate earlier death. Outputs need not be interpreted as calibrated estimates of remaining lifespan, as the scoring specification is comparative, not absolute.

Scoring specification OS is scored using Uno's concordance index (C-index), which measures how well predictions order patients by survival time. For patient pairs whose survival order can be determined from the observed outcomes, a prediction is concordant when the patient who dies earlier receives a shorter predicted survival time. Censored observations contribute only when the ordering can be established. Higher scores indicate better ordering, with 0.5 corresponding to chance and 1.0 to perfect concordance.

Meaning and evidence status OS results measure how well each system ranks patients by subsequent survival using information available at the landmark event. They do not establish the accuracy of predicted survival times in days or of mortality probabilities at specific horizons.

7 Task 2: Time to Next Treatment (TTNT)

Question Given information available at the prediction landmark, how well does the system forecast the time to the next line of therapy or death?

Landmark event The landmark event is defined as the patient's first systemic line of cancer therapy, as indicated in the source data.

Population and timing All breast and lung cancer patients alive at the landmark are included.

Permitted evidence All declared EHR, pathology, and radiology evidence available up to and including the landmark event is permitted. Eligibility depends on when information became available, not only the date of the underlying clinical event. Subsequent treatment decisions, later progression documentation, and retrospectively populated fields must not enter an earlier prediction.

Target and reference label Let T_N be time from the landmark to the start of a subsequent systemic line of cancer therapy. A change in line of therapy is determined by a change in the line-of-therapy numeral recorded in the source documents by the providing healthcare institution. Let T_D be time from the landmark to death. The target is time to either event, $T = \min(T_N, T_D)$. Death before the next line of therapy counts as an event. Let C be time to the last follow-up confirming that neither event has occurred or the administrative cutoff, whichever applies under the frozen rule. The observed label is $Y = \min(T, C)$ with event indicator $\delta = 1$ if either event is observed by C, and 0 otherwise. A censored observation establishes that neither event occurred through Y; it is not an event at Y.

Required output Each model returns a single positive, finite numerical prediction of the number of days from the landmark event to the next line of therapy or death, whichever occurs first. Larger values indicate a longer predicted time before either event, and smaller values indicate an earlier event. Outputs need not be interpreted as calibrated estimates of time remaining before either event, as the scoring specification is comparative, not absolute.

Scoring specification TTNT is scored using Uno's concordance index (C-index), which measures how well predictions order patients by time to the next line of therapy or death. For patient pairs whose event order can be determined from the observed outcomes, a prediction is concordant when the patient who experiences either event earlier receives a shorter predicted time. Censored observations contribute only when the ordering can be established. Higher scores indicate better ordering, with 0.5 corresponding to chance and 1.0 to perfect concordance.

Meaning and evidence status TTNT results measure how well each system ranks patients by subsequent time to the next line of therapy or death using information available at the landmark event. They do not establish the accuracy of predicted times in days or of event probabilities at specific horizons. Treatment timing also reflects clinical practice and access to care, so these results do not directly measure treatment benefit.

8 Scoring and analysis

8.1 Endpoint scores and summary results

OS and TTNT are scored separately within the breast and lung cancer cohorts. The headline score for each cancer type is the arithmetic mean of its OS and TTNT C-indices, with equal weight assigned to the two tasks. Individual endpoint scores are reported alongside the mean because the average can conceal differences in performance between tasks.

8.2 Repeated predictions

A repetition is a separate prediction request for the same patient, task, and input condition. Stochastic models can return different predictions when given the same information. Combining repeated predictions can reduce variability and improve ranking performance by limiting the influence of unusually high or low outputs, although improvement is not guaranteed.

For OS and TTNT, the scorer takes the median of the accepted predictions across repetitions to produce one prediction per evaluation case. One accepted prediction is sufficient. Failed repetitions are excluded from the median; if none yields an accepted prediction, the case is handled under the missing-prediction policies described below. Predictions from different input conditions or workflow configurations are not combined.

Repetitions are distinct from retries following an error. Repetitions provide multiple predictions for aggregation, while retries attempt to obtain a valid response for an individual repetition. Repeated inference increases computation and API usage, so its cost must be considered alongside any improvement in prediction performance.

8.3 Input ablations

An ablation evaluates the same prediction task with a specified subset of the available patient information. These comparisons help assess whether adding pathology or radiology improves performance beyond the EHR alone. The benchmark supports EHR alone, EHR with pathology, EHR with radiology, and EHR with both imaging modalities. The prediction landmark and reference outcome remain unchanged across input conditions for the same patient.

The all-available-evidence condition includes every patient in the task cohort and all permitted evidence available for that patient. Each specified modality combination includes only patients who have every required modality. Consequently, patient populations can differ across ablations unless otherwise indicated. A difference between their overall scores can reflect both the information supplied and the patients included. Comparisons intended to isolate the contribution of a modality therefore require filtering for the same patients in both conditions.

The scorer aggregates repetitions and calculates endpoint scores separately for each ablation, including the accepted-only, failure-penalized, and clinical-fallback views. It does not average predictions across ablations or automatically select the highest-scoring condition.

8.4 Concordance calculation

The scorer calculates Uno's C-index using scikit-survival. Predicted durations are negated before scoring so that shorter predicted times correspond to higher risk. The scoring configuration specifies the evaluation horizon and the reference survival observations used to estimate censoring weights.

Each comparable pair consists of a patient with an observed event before the evaluation horizon and another patient whose observed time is later. An observed event and a censored observation at the same time are also comparable; two observed events at the same time are not. Concordant predictions receive full credit,

discordant predictions receive zero credit, and tied predictions receive half credit. These contributions are weighted to account for censoring.

The scorer reports the number of eligible and evaluated cases and the number of comparable pairs for each score. If a metric cannot be estimated, for example because there are no comparable pairs, it returns no numerical score and records the reason.

8.5 Confidence intervals and paired comparisons

We calculated 95% confidence intervals using a patient-level bootstrap with 1,000 resamples. Within each scored group, breast cancer, lung cancer, or all 412 patients, we sampled patients with replacement. Each resample contained the same number of patients as the original scored group.

We recalculated scores using scikit-survival's implementation of Uno's C-index, with the same predictions, reference labels, and evaluation horizons used for the headline scores. For paired model comparisons, the same sampled patients were used for both models, and their score difference was calculated within each resample. The confidence interval was defined by the 2.5th and 97.5th percentiles of the 1,000 bootstrap estimates.

8.6 Failed predictions and clinical fallback

If no repetition yields an accepted prediction, the evaluation case is recorded as missing. An API error, timeout, or invalid response therefore leaves the case without an aggregate prediction only when no accepted response is available from another repetition. The scorer reports three versions of each endpoint score:

- **Accepted predictions only.** Uno's C-index is calculated using cases with an accepted aggregate prediction. This measures ranking performance when the system returns a result.
- **Failure penalized.** All otherwise comparable patient pairs remain in the calculation. A pair receives zero concordance credit if either patient lacks a prediction, while its censoring weight remains in the denominator. This incorporates failure to return a result into the performance assessment.
- **Clinical fallback.** Each missing prediction is replaced with the fixed, task-specific number of days supplied in the scoring configuration. This value is specified independently of the evaluation labels. Uno's C-index is then calculated with these substituted predictions included.

In the results below, each model returned a valid response for every patient and task, so no fallback or penalties were required.

The clinical fallback score represents a workflow in which an alternative estimate remains available when the model fails. This addresses the practical requirement that clinical decisions continue even when a prediction service is unavailable. The implemented fallback in this case was the median value observed for the task in the test set. Comparing the three scores distinguishes performance on successful responses from performance when failures must also be handled.

9 Results Across Models

We report a headline score that averages the OS and TTNT C-indices within each cancer type. Protege selected the top-performing frontier results after comparing harness configurations and ablations. We recorded the report time and cost across 2,254 predictions from the 412 patients.

System	Breast mean	Lung mean	Median seconds per prediction	USD per prediction
Standard Model Bio	0.747	0.561	2	0.009
Gemini 3.7 Flash	0.803	0.536	108	0.32
GPT-5.6-sol	0.755	0.532	108	1.53
Claude Opus 5	0.800	0.550	166	1.50

Table 2. The headline score averages the OS and TTNT C-indices within each cancer type. Cohort totals are breast $n = 271$ and lung $n = 141$.

9.1 Endpoint specific estimates

Endpoint	SMB	GPT-5.6-Sol-max	Gemini-3.7-Flash	Claude-Opus-5.5-max
Breast overall survival	0.786 [0.693, 0.869]	0.780 [0.592, 0.925]	0.841 [0.737, 0.923]	0.816 [0.680, 0.928]
Lung overall survival	0.591 [0.488, 0.695]	0.524 [0.431, 0.618]	0.541 [0.443, 0.632]	0.570 [0.482, 0.658]
Breast next treatment	0.707 [0.625, 0.782]	0.729 [0.661, 0.800]	0.764 [0.711, 0.821]	0.783 [0.727, 0.840]
Lung next treatment	0.530 [0.452, 0.606]	0.539 [0.464, 0.606]	0.531 [0.451, 0.599]	0.530 [0.457, 0.598]

Table 3. Endpoint C-index point estimates with 95% bootstrap confidence intervals in brackets.

Breast cancer SMB's mean score is lower than GPT's by 0.008, Gemini's by 0.056, and Claude's by 0.053 C-index units. The endpoint estimates show that SMB is slightly above GPT for OS and below it for TTNT. A single mean obscures that distinction.

Lung cancer SMB's mean exceeds GPT's by 0.029, Gemini's by 0.025, and Claude's by 0.011. The lung means are relatively close to the chance-ordering reference. These rounded point estimates do not establish statistical superiority.

Reported uncertainty This report evaluates a family of 36 paired comparisons evaluating performance differences ($\Delta = C_{\text{comparator}} - C_{\text{SMB}}$) across 3 comparator models (GPT-5.6-sol, Gemini-3.7-Flash, Claude-Opus-5), 2 indications (breast, lung), 2 endpoints (OS, TTNT), and 3 evaluation arms (headline, EHR-only, EHR+DICOM). All reported 95% bootstrap confidence intervals are unadjusted for multiple comparisons. Across these 36 comparisons, 34 include zero and 2 exclude zero in favor of Claude for breast TTNT. Additionally, the breast OS analysis includes only 20 observed deaths in the headline/EHR-only cohorts ($n = 271$) and 15 in the EHR+DICOM cohort ($n = 232$), yielding wide confidence intervals. The complete 36-row table of paired differences, confidence bounds, and reproduction script references are provided in Table A.[1]

Implications for model selection SMB's main advantage in this comparison is its reported execution cost and speed. The performance case is more mixed: frontier systems lead on the breast-cancer mean, while SMB leads on the lung-cancer mean. The Pareto charts show these tradeoffs using the reported point estimates.

9.2 Pareto comparisons

The preferred direction is toward the upper left: higher outcome-ranking performance with less time or cost. A system is Pareto-dominated when another is at least as good on both plotted measures and better on one. Among the reported breast-cancer point estimates, SMB and Gemini are non-dominated in both charts. For lung cancer, SMB dominates on these two measures. This describes the reported estimates, without accounting for uncertainty or differences in workload.

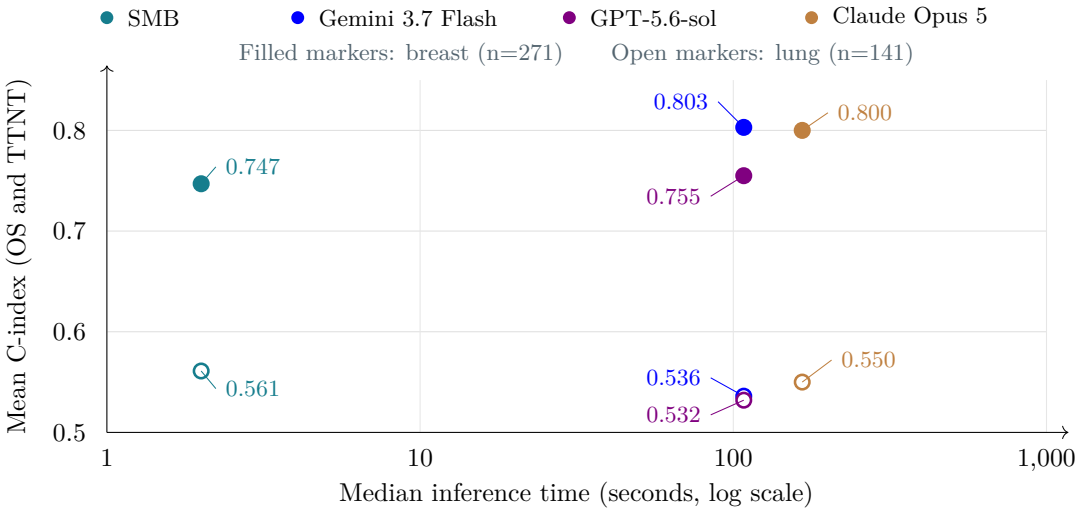

Figure 3. Prediction performance and inference time. The horizontal axis shows median inference time in seconds on a logarithmic scale. Performance is the mean of OS and TTNT C-indices.

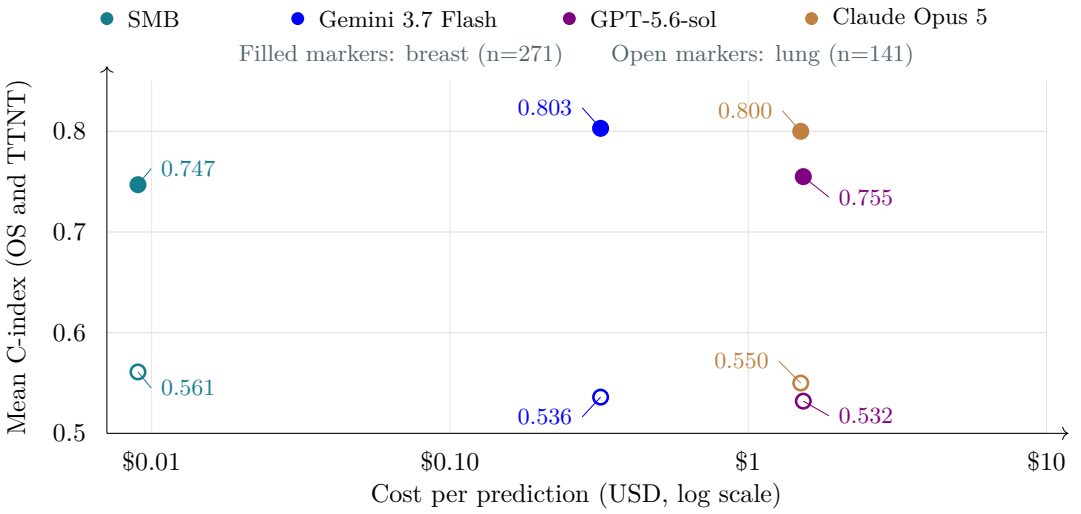

Figure 4. Prediction performance and inference cost. The horizontal axis shows reported USD per prediction on a logarithmic scale. Performance is the mean of OS and TTNT C-indices.

10 Cost and turnaround

10.1 Reported execution measures

SMB's reported median inference time is 2 seconds, compared with 108 seconds for GPT and Gemini and 166 seconds for Claude. The comparator-to-SMB median ratios (54x and 83x) describe the relative central tendency of single-run model execution speed under benchmark conditions. Because they are ratios of aggregate medians rather than paired patient-level speedups, they do not capture tail latency, throughput, or the total time from a clinical request to an actionable result.

Reported cost per prediction is $0.009 for SMB and $0.32–$1.53 for the comparators, corresponding to approximately 36–170 times SMB's reported cost. Training, implementation, operations, and deployment costs are not accounted for in these figures.

10.2 Data Steps for Each Prediction

One prediction represents one end-to-end patient inference for a single patient, task, and system condition. All intermediate review calls, synthesis steps, retries, and repair prompts belong strictly to that single end-to-end inference unit.

The completed results support a comparison of reported execution figures; they do not yet establish matched end-to-end workload economics.

Future runs should report median and 95th-percentile completion time, throughput at fixed concurrency, and total execution spend. Separate model-only time from the full workflow so that preprocessing and evidence coverage can be compared.

11 Assessment

11.1 Implications for adoption

SMB merits further evaluation where high prediction volumes make execution cost important. Its lung-cancer point estimates are encouraging, although the current evidence is insufficient to establish superiority. For breast cancer, the performance gains of frontier models offer potential utility, but must be balanced against a 36- to 170-fold cost increase and significantly longer response times.

Organizations should validate the shortlisted system on their own patients and intended task. That review should include absolute risk accuracy, failed predictions, integration effort, and full workflow cost. The present study supplies an initial comparison, rather than a deployment decision.

11.2 Limits of the findings

The study evaluates retrospective prediction. It does not show whether use of a model changes treatment decisions or improves patient outcomes. Faster model responses may be useful, but their contribution to clinical turnaround depends on the surrounding workflow.

Performance at hospitals outside the evaluated cohort remains untested in the reported results.

A Separate demonstration and protocol versions

For each task, we give below an example of one patient's evidence cutoff and ground truth reference.

Demonstration task	Evidence cutoff	Example of Reference Described in the Narrative
Time to death	Diagnosis	Death 87 days after diagnosis; event observed
Treatment or death	First-line treatment start	Death at 67 days; no intervening qualifying treatment change

Table 4. Definitions and references for each task.

For this example patient, their record has the same 83 images across all assessed models: a pathology thumbnail, 32 tissue tiles, and 50 radiology images.

Indication	Endpoint	Arm	GPT-5.6-sol	Gemini-3.7-Flash	Claude-Opus-5
Breast	OS	headline	-0.006 [-0.227, +0.172]	+0.055 [-0.077, +0.183]	+0.030 [-0.123, +0.179]
Lung	OS	headline	-0.068 [-0.176, +0.049]	-0.051 [-0.148, +0.047]	-0.021 [-0.114, +0.070]
Breast	OS	ehr_only	+0.006 [-0.160, +0.157]	+0.052 [-0.079, +0.170]	+0.077 [-0.055, +0.194]
Lung	OS	ehr_only	-0.058 [-0.165, +0.066]	-0.071 [-0.181, +0.046]	-0.001 [-0.098, +0.098]
Breast	OS	ehr_dicom	-0.029 [-0.234, +0.182]	+0.052 [-0.140, +0.247]	+0.059 [-0.112, +0.236]
Lung	OS	ehr_dicom	+0.022 [-0.161, +0.195]	+0.059 [-0.089, +0.196]	+0.101 [-0.023, +0.242]
Breast	TTNT	headline	+0.021 [-0.056, +0.099]	+0.057 [-0.006, +0.126]	**+0.075 [+0.021, +0.133]**
Lung	TTNT	headline	+0.009 [-0.064, +0.092]	+0.001 [-0.084, +0.078]	+0.000 [-0.077, +0.075]
Breast	TTNT	ehr_only	+0.041 [-0.025, +0.114]	+0.058 [-0.004, +0.123]	**+0.068 [+0.015, +0.127]**
Lung	TTNT	ehr_only	-0.006 [-0.087, +0.078]	-0.018 [-0.107, +0.068]	+0.001 [-0.071, +0.078]
Breast	TTNT	ehr_dicom	+0.007 [-0.081, +0.094]	+0.038 [-0.024, +0.098]	+0.043 [-0.016, +0.105]
Lung	TTNT	ehr_dicom	-0.043 [-0.147, +0.061]	-0.030 [-0.159, +0.092]	-0.011 [-0.121, +0.103]

Table 5. Paired C-index differences ($\Delta = C_{\text{comparator}} - C_{\text{SMB}}$) with unadjusted 95% bootstrap confidence intervals across all 36 evaluation conditions.

B Sources

[1]

[2] Brown K. *Introducing Standard Model Biomedicine*. September 30, 2025. Developer description of the model and intended uses. Official Standard Model Biomedicine article. Accessed October 20, 2026.

[3] Standard Model Bio. *End-to-end example*. Public documentation for `smb-v1-1.7b`, patient embeddings and fitted task heads. Official implementation example. Accessed October 20, 2026.

[4] US Centers for Disease Control and Prevention. *Deaths and Mortality*. Final 2024 data. CDC FastStats. Accessed October 20, 2026.

[5] Hanna TP et al. *Mortality due to cancer treatment delay: systematic review and meta-analysis*. BMJ 2020;371:m4087. doi:10.1136/bmj.m4087.

[6] Gitlin M, McGarvey N, Shivaprakash N, Cong Z. Time duration and health care resource use during cancer diagnoses in the United States: A large claims database analysis. *Journal of Managed Care & Specialty Pharmacy*. 2023;29(6):659–670. doi:10.18553/jmcp.2023.29.6.659.

[7] International Conference on Harmonisation. *E9: Statistical Principles for Clinical Trials*. U.S. Food and Drug Administration; September 1998. `https://www.fda.gov/media/71336/download`