SHIFT: BENCHMARKING CLINICAL COMPUTER-USE AGENTS OVER FULL PATIENT TRAJECTORIES

Anonymous authors

ABSTRACT

Computer-use agents operate software through its graphical interface and could work inside the electronic health record (EHR) systems clinicians use. Yet most medical AI benchmarks still rely on pre-curated text, images, or structured data, and evaluate isolated tasks rather than a continuous course of care. It therefore remains unclear whether agents can navigate clinical interfaces, maintain a patient's evolving state, and carry care from initial assessment through follow-up. Here we introduce SHIFT (Stateful Hospital Interaction for Full Trajectories), a benchmark for clinical computer-use agents over complete patient trajectories. SHIFT builds on a real EHR (OpenMRS) and image viewer (OHIF), connecting interviews, charting, orders, results, image interpretation, diagnosis, treatment, and follow-up to a shared patient state. We score agents along seven clinical capabilities, using case-specific dependency graphs to check the evidence supporting each scoring item while accepting clinically valid alternatives. Across nine models, GPT-6-Astra achieves the highest mean score of 53.82 out of 100. Five models score below 30. Common errors include incorrect patient information, misinterpreted test results, and neglected earlier findings. Scores are lower in completed six-patient sessions. By exposing where agents lose track of patient state and evidence, SHIFT quantifies the gap that must be closed before computer-use agents can safely operate real clinical software.

1 INTRODUCTION

Computer-use agents operate software through graphical interfaces and could work inside the electronic health record (EHR), order-entry systems, and imaging workstations used by clinicians. OSWorld and WebArena have moved evaluation into real computer and web environments, while medical benchmarks increasingly involve active interviews, tool use, and clinical decisions (Xie et al., 2024; Zhou et al., 2024; Schmidgall et al., 2026). Yet many medical benchmarks use pre-curated data for isolated tasks, such as static medical question answering in MedQA (Jin et al., 2021), rather than evaluating a continuous course of care. It remains unclear whether agents can use these interfaces, track patient state, and manage care from initial assessment through follow-up.

Clinical care requires judgments and software actions to remain connected. An agent may know that a test is needed but fail to locate or submit the order, or access a report without using its findings in later diagnosis and treatment. As new results appear and the treatment stage changes, it must also revisit earlier judgments and determine whether the existing plan still applies. Clinicians spend substantial time interacting with EHRs, and order-entry design can introduce medication-error risks (Sinsky et al., 2016; Koppel et al., 2005). These findings make the clinical interface both the setting in which care is delivered and part of the performance being evaluated.

Existing work follows two main directions. Interactive clinical benchmarks organize care around interviews and decisions, but tests and treatments are often represented as language instructions or structured actions. Software-oriented benchmarks evaluate reading, writing, and operating clinical systems, but commonly use individual tasks or recorded workflows as the evaluation unit. These approaches advance clinical reasoning and software interaction separately, but they do not fully test whether an agent can acquire evidence through an interface and use it across a continuous episode for the same patient.

We introduce SHIFT (Stateful Hospital Interaction for Full Trajectories), which uses a complete patient trajectory as the evaluation unit. SHIFT builds on a real EHR, OpenMRS, and integrates

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the OHIF image viewer, connecting interviews, charting, orders, results, image interpretation, diagnosis, treatment, and follow-up to a shared patient state. Agents operate the workstation through browser screenshots, actively collect information, submit orders, track results, and carry evidence into later decisions. Persistent system records and clinical content together establish whether an action occurred, whether evidence was interpreted correctly, and whether a later plan is supported.

Oncology provides a useful setting because staging, treatment selection, and response assessment depend on integrating imaging, pathology, and other clinical evidence across multiple stages. Reasonable care paths are also not unique: clinicians may change the order of tests or choose equivalent evidence. SHIFT therefore evaluates seven shared clinical capabilities while constructing a case-specific dependency graph for each patient. Each item is scored against its evidence requirements, while clinically valid alternative tests, evidence combinations, and action sequences remain acceptable. The graph specifies evidence requirements without prescribing a fixed interface sequence.

Our contributions are threefold. First, we build an executable environment based on real clinical software in which agents complete care from initial assessment through follow-up while maintaining a changing patient state. Second, we develop a scoring method that combines seven shared capabilities with case-specific evidence dependencies. Third, we compare nine models across six oncology pathways and 600 cases, and examine how information errors affect later decisions and how concurrent management changes clinical performance.

2 Related Work

Interactive clinical care simulation. Prior work has extended the evaluation of medical agents from single-answer assessment to interactive clinical care. AgentClinic and AI Hospital model clinical encounters using simulated physicians, patients, and examinations. MedChain organizes cases into interdependent clinical stages, while CP-Env provides a controllable environment for clinical-pathway simulation (Schmidgall et al., 2026; Fan et al., 2024; Liu et al., 2025; Zhu et al., 2025). These benchmarks examine how agents develop care plans through successive interviews, tests, and clinical decisions. However, they advance care through simulated dialogue or tool calls and do not require agents to interact directly with the graphical user interface (GUI) of clinical software. SHIFT requires agents to place orders, write clinical notes, and review results through a GUI (Table 1). Successful task completion therefore depends on both correct clinical judgment and effective software use.

Table 1: Scope of representative clinical-agent benchmarks.

Benchmark	GUI interaction	Full patient trajectory	Persistent patient state	Multimodal evidence	Multi-patient management
AgentClinic	N/A	N/A	✓	✓	N/A
AI Hospital	N/A	N/A	✓	N/A	N/A
MedChain	N/A	*Partial*	✓	✓	N/A
CP-Env	N/A	*Partial*	✓	N/A	N/A
MedAgentBench	N/A	N/A	N/A	✓	N/A
PhysicianBench	N/A	N/A	N/A	✓	N/A
MedAgentGym	N/A	N/A	N/A	✓	N/A
CareFlow	✓	N/A	N/A	✓	N/A
SHIFT (ours)	✓	✓	✓	✓	✓

✓ indicates coverage. *Partial* indicates limited coverage. N/A marks a dimension outside the task setting considered here. Entries describe the scope of this comparison, not each benchmark's full capabilities.

Clinical tool use and software interaction. MedAgentBench and PhysicianBench evaluate agents in electronic health record (EHR) environments. PhysAssistBench combines patient interviews with EHR assistance, while MedAgentGym focuses on code-based biomedical data analysis (Jiang et al., 2025; Liu et al., 2026; Du et al., 2026; Xu et al., 2025). CarePilot's CareFlow benchmark examines screen-based interaction and clinical software workflows (Ghosh et al., 2026). However, these benchmarks use individual clinical requests, visits, or software workflows as their unit of evalua-

tion, rather than assessing care for the same patient continuously from initial presentation through follow-up. It remains unclear whether agents can manage care across successive stages or how early omissions affect subsequent treatment and follow-up. SHIFT evaluates the full patient episode as a single continuous task, with a shared patient state updated across interviews, tests, diagnosis, treatment, and follow-up. The agent must integrate prior findings with new results and adjust the care plan as needed until the full episode of care is complete.

Process evaluation of clinical care. A correct final diagnosis or treatment recommendation does not, by itself, establish that an agent obtained and correctly used the necessary clinical evidence. At the same time, clinically appropriate action sequences should not be penalized solely for differing from a reference workflow. PaperBench decomposes complex tasks into rubric items to assess task completion (Starace et al., 2025). For clinical care, evaluation must assess not only whether individual actions are completed but also whether subsequent decisions are supported by appropriate evidence. SHIFT specifies the evidence required to support each clinical judgment in a case. Scoring considers both the correctness of the judgment and whether the agent obtained and correctly used the required evidence. Agents may choose different tests or testing sequences, provided that these choices are clinically appropriate and yield the evidence required to support the judgment.

3 The SHIFT Environment

3.1 Clinical Environment and Interaction

SHIFT builds on OpenMRS 3 (Asaolu, 2024). We retain its chart, order, and result management functions and add patient interviews and oncology-specific documentation. Orthanc stores DICOM data, while the OHIF viewer supports series selection, slice navigation, window adjustment, zooming, and lesion measurement (Jodogne, 2018; Ziegler et al., 2020).

Browser screenshots are the agent's only source of clinical observations. The agent uses mouse and keyboard actions to operate the workstation, types interview questions into the interface, and reads patient replies in later screenshots. A Playwright executor carries out the low-level actions. The agent has no direct access to the DOM, backend database, or private scoring interfaces. It must obtain evidence through the corresponding interface actions. Writing that an order was submitted or a result reviewed does not prove that the action occurred.

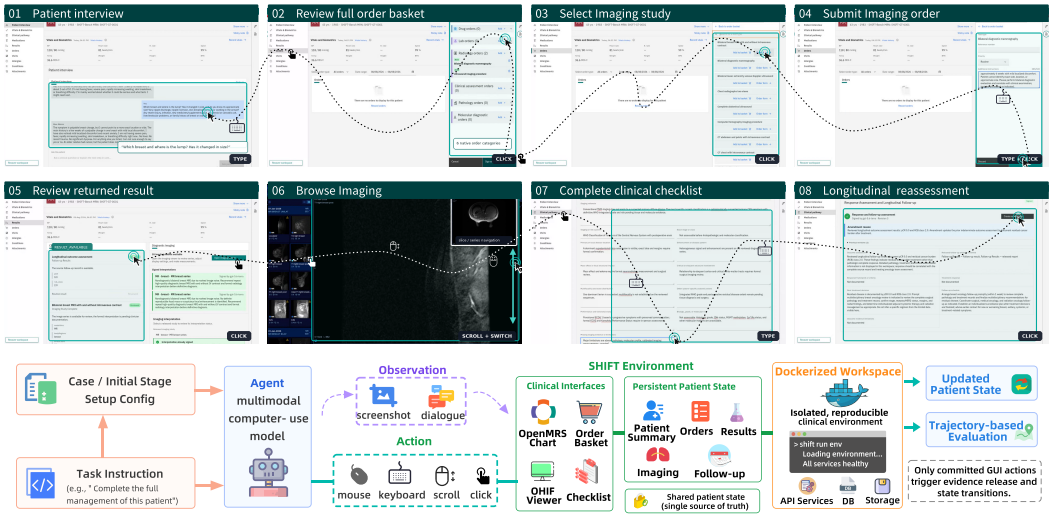

Figure 1: SHIFT's clinical environment and patient workflow. OpenMRS provides the clinical workstation and OHIF the image viewer. Patient dialogue appears in the interface and is read through screenshots, not through a separate input channel. The screens show the available interface functions rather than a required action sequence.

3.2 Patient State and Clinical Workflow

Figure 1 shows the workflow. For a patient requiring cancer assessment, the agent reviews the initial chart, asks about symptoms and treatment history, and orders appropriate tests. Once results are available, it reviews reports, opens images, measures lesions, and uses the findings to form a diagnosis, assess disease extent and risk, and plan treatment. At follow-up, it must compare new findings with baseline evidence to assess response and revise the plan. The patient's history, tests, and care records remain in the system throughout the episode, so the agent must continue to use earlier findings in later decisions.

3.3 Cases and Reference Standards

SHIFT contains 600 cases across six oncology pathways, with 100 cases per pathway: breast cancer (BC), colorectal liver metastases (CRLM), glioblastoma (GBM), hepatocellular carcinoma (HCC), head and neck squamous cell carcinoma (HNSCC), and non-small cell lung cancer (NSCLC). Each case includes an initial chart, interview information, tests and their results, linked images, clinical objectives, and conditions for ending the task.

Three clinicians develop and review the reference standards in separate roles. The first completes the task in the evaluation interface and records actions, clinical judgments, and supporting evidence. The second checks these against the case materials to confirm that they are clinically appropriate and can be carried out in the system. The third reviews disagreements and makes the final decision. The clinicians also record acceptable alternative tests, treatment plans, and action sequences, so that clinically reasonable differences are not marked as errors. Information unavailable in a case, such as pathology results, molecular test results, or follow-up records, is marked separately and is not counted as an agent omission.

4 Evaluation Method

SHIFT evaluates each patient episode across seven clinical capabilities. Each capability contains case-specific scoring items with predefined point values. Each item receives a single score based on action completion, clinical correctness, and whether the agent obtained and correctly used the required evidence. Weighted item scores are summed to obtain capability scores and a case total.

4.1 Scoring Criteria

All cases are evaluated on the same seven clinical capabilities. Each capability has a fixed maximum score, and these maximum scores sum to 100 points (Table 2). The scoring items within each capability depend on the case. For example, image interpretation in a breast cancer (BC) case may focus on tumor size and axillary lymph nodes, whereas a glioblastoma (GBM) case may involve enhancement, mass effect, and postoperative changes. Every case includes scoring items for all seven capabilities.

The items and their point values are specified in the case rubric before evaluation. All models are evaluated using the same rubric for a given case. Items are not removed, and points are not redistributed, in response to model performance.

Information that is absent from the case and cannot be obtained through permitted actions is identified when the rubric is constructed. The agent is not required to provide this information. Instead, the relevant items assess whether the agent recognizes the information gap and proposes an appropriate follow-up plan. This rule does not apply to information that the agent could have obtained but failed to retrieve or use correctly.

4.2 Item Scoring

Scoring checks both what the agent did and whether its clinical choices and judgments were correct. Saved orders, reports, and action records are used to check whether the agent submitted an order, opened a test result or image, or saved a report. The selected tests, image interpretations, and care

Table 2: Clinical capabilities, maximum points, and assessment focus.

Capability	Points	Assessment focus
History and initial assessment	15	History, risk factors, and clinical summary
Evidence strategy and use	13	Test selection, order submission, and result review
Imaging and core evidence interpretation	17	Findings, measurements, interpretation, and limitations
Diagnostic synthesis	15	Diagnosis, differential diagnosis, and confirmation plan
Disease extent, severity, and risk	10	Disease extent, staging, and risk assessment
Treatment and shared decision-making	20	Treatment rationale, alternatives, and patient preferences
Longitudinal reassessment	10	Baseline comparison, follow-up findings, and plan revision
Total	**100**	

plans are compared with the case materials, scoring rubric, and clinician reference trajectory. Each scoring decision includes an explanation and references to the supporting records.

Evidence requirements are checked as part of item scoring, as described in Section 4.3. These checks help determine the item's score. They do not introduce a separate deduction.

Let s_{kit} denote the score for item t under capability i in case k. Each item receives one of five scores. A score of 0 indicates that none of the required work was completed or that a predefined zero-credit rule applies. A score of 0.25 indicates some valid work but a major error in a key part of the item. A score of 0.50 indicates partial completion, with the completed parts largely correct. A score of 0.75 indicates that the main content is correct, with minor omissions. A score of 1.00 indicates full and correct completion, including all evidence requirements. The case rubric defines the criteria for each score before evaluation.

4.3 Evidence Requirements

We construct a directed acyclic graph (DAG), $G_k = (V_k, E_k)$, from the case materials and the clinician reference trajectory. Each node represents a scoring item. A directed edge links an item that provides clinical information to another item that requires that information. The graph records these evidence dependencies, while the case rubric specifies the evidence needed for each judgment and the acceptable alternatives.

For each clinical judgment, scoring checks whether the agent had obtained and correctly used the required evidence when making that judgment. An earlier item's score does not automatically determine a later item's score. Different tests or combinations of evidence are acceptable when they meet the same requirement, so the agent does not need to follow a single sequence of interface actions.

For example, consider two lesions, A and B, with baseline measurements of 5 cm and 3 cm. Both lesions remain unchanged during follow-up. The agent compares the current 3 cm measurement of lesion B with the baseline 5 cm measurement of lesion A and reports a 40% reduction. The agent has opened images, measured a lesion, calculated a percentage, and recorded a conclusion, but the comparison uses measurements from different lesions. The conclusion is therefore incorrect. Scoring must check that the measurements refer to the same lesion, rather than only checking whether the interface actions were completed.

The case rubric specifies which evidence omissions result in zero credit and which allow partial credit. For example, a treatment may be appropriate for a patient, but the agent may place the order without checking the allergy history or test results required by the case. In a rubric that scores order entry and pre-treatment verification as separate items, the order-entry item may receive credit while the verification item receives zero. Other items are scored according to their own requirements.

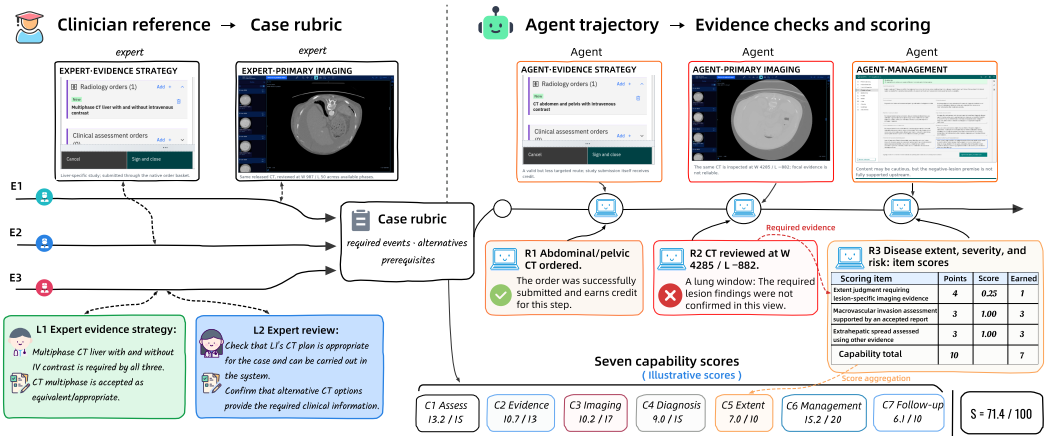

Figure 2: Evidence checks and scoring for one patient episode. Case-specific evidence requirements guide item scoring. Item scores are combined into seven capability scores and a case total. All scores are illustrative.

4.4 Score Aggregation

Each item's score is multiplied by its maximum point value. These points are added within each capability, and the seven capability scores are then added to obtain the case total. Let T_{ki} be the scoring items for capability i in case k, and let w_{kit} be the maximum point value of item t. The capability score S_{ki} and case total S_k are

$$S_{ki} = \sum_{t \in T_{ki}} w_{kit} s_{kit}, \qquad S_k = \sum_{i=1}^{7} S_{ki}. \tag{1}$$

Item point values are nonnegative and sum to the fixed maximum for their capability in Table 2. The seven capability maxima total 100 points, so the case score ranges from 0 to 100. Each item is counted once, with no additional deductions during aggregation.

The model score gives equal weight to each oncology pathway. We first average case scores within each pathway and then average the six pathway means. Let K_d be the set of cases in pathway d, with $N_d = |K_d|$. The model score is

$$M = \frac{1}{6} \sum_{d=1}^{6} \frac{1}{N_d} \sum_{k \in K_d} S_k. \tag{2}$$

Figure 2 illustrates the calculation for one patient. In the 10-point "Extent, Severity, and Risk" capability, a 4-point lesion-extent item receives 0.25 for some valid work but a major error under the case rubric, earning $4 \times 0.25 = 1$ point. Two other items, worth 3 points each, are fully completed with the required evidence. The capability score is therefore $1 + 3 + 3 = 7$ out of 10. Adding all seven capability scores gives the illustrated case total of 71.4 out of 100.

Safety events, including actions on the wrong patient, dangerous clinical decisions, and missed critical results, are recorded separately and reported alongside the clinical scores. The total score measures completion of the tasks defined in the case rubric. It does not, on its own, show that the episode was managed safely.

5 Experiments and Results

5.1 Experimental Setup

We evaluate nine models on 600 cases across six oncology pathways. The experiments examine single-patient performance, clinical capabilities, and multi-patient performance.

We repeat the single-patient evaluation independently three times. Overall scores are reported as means and standard deviations across the three runs. Pathway and capability scores are reported as means across runs. Tool-call counts are collected from one run and summarized as means and standard deviations across cases.

For six-patient experiments, clinical scores are calculated only from sessions that complete the full six-patient workflow. GPT-5.6 Terra is also evaluated in two- and three-patient sessions to examine performance as the number of patients increases. All comparisons are descriptive.

5.2 Performance in Single-Patient Care

GPT-6-Astra achieves the highest overall score of 53.82, exceeding GPT-5.6 Terra by 12.46 points. Five of the nine models score below 30 (Table 3). Performance also varies across oncology pathways. GPT-6-Astra scores 68.47 on GBM and 43.29 on HCC, a difference of 25.18 points. HCC is the lowest-scoring pathway for eight of the nine models.

The average number of tool calls ranges from 196.5 to 472.6 per case across models (Table 4). Gemini 2.5 Flash makes the fewest calls, but some of its cases terminate early. Lower call counts therefore do not necessarily indicate more efficient task completion.

Table 3: Overall and pathway-specific scores. All scores are out of 100.

		Oncology pathway					
Model	**Overall**	**BC**	**CRLM**	**GBM**	**HCC**	**HNSCC**	**NSCLC**
GPT-6-Astra	**53.82 (2.31)**	**57.16**	**50.83**	**68.47**	**43.29**	**52.08**	**51.09**
GPT-5.6 Terra	41.36 (1.68)	45.80	39.60	51.84	33.92	36.78	40.22
Claude Opus 5	36.57 (2.09)	42.73	34.68	44.92	29.53	33.16	34.40
GPT-5.6 Luna	32.84 (1.27)	37.95	30.42	40.36	27.18	29.07	32.06
Grok 4.5	29.61 (1.84)	34.28	27.53	36.41	24.02	27.16	28.26
Claude Haiku 4.5	26.43 (1.06)	31.57	25.04	31.28	20.16	25.42	25.11
Gemini 3.5 Flash	25.31 (1.73)	28.74	24.19	30.06	21.37	23.88	23.62
Gemini 2.5 Flash	23.87 (1.12)	27.36	22.58	29.14	19.42	22.61	22.11
Qwen3.8-Max	22.92 (1.49)	21.85	23.82	31.20	17.96	25.13	17.56

Overall scores are reported as mean (SD) across three independent runs ($n = 3$). Pathway scores are three-run means. Bold values indicate the highest mean in each column.

Table 4: Tool calls per case (mean $\pm$ SD across cases).

Model	**Tool calls**	**Model**	**Tool calls**
GPT-6-Astra	214.7 ± 16.2	Claude Haiku 4.5	269.7 ± 44.3
GPT-5.6 Terra	226.3 ± 31.8	Gemini 3.5 Flash	331.8 ± 28.5
Claude Opus 5	472.6 ± 38.4	Gemini 2.5 Flash	196.5 ± 33.2
GPT-5.6 Luna	396.2 ± 55.7	Qwen3.8-Max	448.9 ± 52.6
Grok 4.5	308.4 ± 22.1		

Statistics are from a single run. Each tool invocation counts once. Text-only responses are excluded.

5.3 Performance Across Clinical Capabilities

Eight of the nine models earn more than 50% of the available points in initial assessment (Table 5). In contrast, all nine models score below 40% in both image interpretation and the "Extent, Severity, and Risk" capability. GPT-6-Astra achieves the highest scores in these two capabilities, at 6.58/17 (38.7%) and 3.74/10 (37.4%), respectively.

Three cases from individual runs illustrate errors in the clinical workflow. In BC case `ISPY1_1206`, Gemini 3.5 Flash records a 43-year-old patient's age as 20. It then cites the patient being "under 30" as the reason for selecting ultrasound as the initial test and receives 3.75/15 for initial assessment.

In BC case `ISPY1_1064`, Claude Haiku 4.5 interprets `pCR = 0` as indicating that the patient did not receive neoadjuvant treatment. The indicator actually means that pathological complete response was not achieved. The model orders initial diagnostic tests again and receives 1.50/10 for follow-up.

In NSCLC case `LUNG1-272`, GPT-5.6 Terra reads and acknowledges a record of the patient's death but still schedules specialist assessment, tissue sampling, and further staging and treatment planning within one week. It receives 2.50/10 for follow-up.

Table 5: Scores for the seven clinical capabilities.

Model	Initial (15)	Evidence (13)	Imaging (17)	Diagnosis (15)	Extent/risk (10)	Treatment (20)	Follow-up (10)
GPT-6-Astra	10.93	**7.41**	**6.58**	**7.82**	**3.74**	**11.17**	**6.17**
GPT-5.6 Terra	**11.86**	7.29	3.67	3.08	1.64	8.92	4.90
Claude Opus 5	10.72	6.64	3.41	2.87	1.53	7.68	3.72
GPT-5.6 Luna	10.18	6.07	2.63	2.24	1.21	6.71	3.80
Grok 4.5	9.67	5.58	2.49	2.03	1.06	5.76	3.02
Claude Haiku 4.5	9.03	5.21	2.28	1.81	0.94	4.54	2.62
Gemini 3.5 Flash	8.76	5.47	1.36	1.22	0.73	5.13	2.64
Gemini 2.5 Flash	8.41	4.69	2.01	1.58	0.88	4.27	2.03
Qwen3.8-Max	6.82	4.52	2.43	2.21	1.28	3.92	1.74

Scores are means across three independent runs. Maximum points are shown in parentheses. Bold values indicate the highest mean in each column. Capability definitions appear in Table 2.

5.4 Performance in Multi-Patient Care

The six-patient evaluation includes same-disease and cross-disease sessions. Same-disease results are averaged equally across the six oncology pathways. Each cross-disease session includes one patient from each pathway.

Among sessions that complete the full six-patient workflow, all nine models score below their single-patient baselines (Figure 3a). Scores decrease by 5.85–8.87 points in the same-disease condition and by 2.25–9.58 points in the cross-disease condition. GPT-6-Astra achieves the highest scores in both conditions, at 47.97 and 51.57, respectively. Seven models show smaller decreases in the cross-disease condition than in the same-disease condition. Claude Haiku 4.5 and Gemini 2.5 Flash show the opposite pattern.

We further examine GPT-5.6 Terra with two, three, and six patients (Figure 3b). With the pathway fixed to BC, its scores decrease by 2.70, 4.80, and 9.31 points, respectively, from its single-patient BC baseline of 45.80.

In the cross-disease condition, the corresponding decreases are 1.67, 2.55, and 2.85 points relative to single-patient baselines matched to the disease composition of each group. Because both patient number and disease composition change across these groups, the results do not isolate the effect of patient number. The two- and three-patient experiments each include only four sessions, so these results are exploratory. GPT-5.6 Terra completes all nine sessions in the six-patient BC condition.

No wrong-patient actions are observed for GPT-6-Astra or GPT-5.6 Terra. Their patient-switch success rates are 94.60% and 89.40%, respectively. The remaining models have wrong-patient action rates of 0.40%–2.40% and switch success rates of 65.80%–85.20%.

6 Discussion and Limitations

The results suggest that agents are better at organizing initial patient information than at interpreting evidence and using it to update care plans. Case analyses show that incorrect information can affect later decisions, while important information already obtained may not be used to revise the plan. A coherent care plan alone does not establish that the underlying clinical process is reliable. Evaluation must therefore examine the evidence supporting each judgment. SHIFT uses case-specific

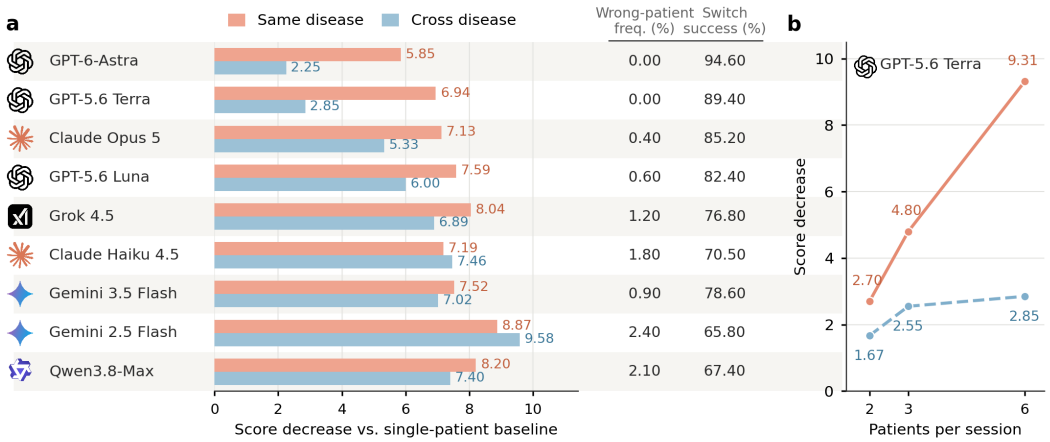

Figure 3: Model performance in multi-patient care. **(a)** Clinical score decreases relative to single-patient baselines, wrong-patient action rates, and patient-switch success rates for nine models managing six patients. Clinical scores include only sessions that complete the full six-patient workflow. Same-disease results are averaged equally across the six pathways. Cross-disease sessions include one patient from each pathway. Wrong-patient action rates divide patient-specific actions on the wrong patient's record by all tool actions. Switch success rates divide successful switches by initiated switches. Failed switches are not counted as wrong-patient actions. **(b)** GPT-5.6 Terra with two, three, and six patients. Same-disease groups use BC cases involving neoadjuvant chemotherapy (NACT). Cross-disease groups contain BC NACT and NSCLC in the two-patient condition, add GBM at three patients, and include all six pathways at six patients. Score decreases are measured against single-patient baselines matched to each group's disease composition.

dependency graphs to check individual scoring items and identify cases where software actions are completed but the associated clinical judgments lack adequate support.

In completed six-patient sessions, all nine models score below their single-patient baselines, while patient-switch success rates and wrong-patient action rates vary across models. Managing several patients increases the workload and requires agents to keep patient information and care plans separate when switching records. Clinical scores should therefore be reported alongside these operational metrics to assess both task completion and correct patient selection.

This study covers only six oncology pathways, and the simulated environment cannot fully reproduce clinical communication, collaboration, and workload. Case rubrics and dependency graphs cannot cover every reasonable care option, and scoring consistency requires further validation. Multi-patient clinical scores include only sessions that complete the full workflow. These scores describe performance among completed sessions and do not measure how often a model finishes the full task. Patient-load experiments have small samples. Cross-disease conditions also change both patient number and disease composition, so their results do not isolate the effect of patient number and remain exploratory.

7 Conclusion

We present SHIFT, a benchmark for evaluating clinical agents on complete patient episodes, from initial assessment to follow-up, through the GUI of clinical software. SHIFT includes 600 cases across six oncology pathways and uses case-specific dependency graphs to check the evidence supporting clinical judgments. Evaluation of nine models shows that initial assessment is generally stronger than image interpretation and assessment of disease extent and risk. All nine models also score below their single-patient baselines in completed six-patient sessions. These results highlight the need to evaluate both final clinical decisions and the actions and evidence used to reach them.

AI USE STATEMENT

We used large language models solely to improve the grammar, wording, and readability of the manuscript. The authors reviewed all revisions and take full responsibility for the final content.

REFERENCES

David Asaolu. O3: The new OpenMRS explained and the investments that made it possible. https://openmrs.org/o3-the-new-openmrs-explained-and-the-investments-that-made-it-possible/, June 2024. Accessed: 2026-09-26.

Tianming Du, Peijie Yu, Sihan Shang, Danli Shi, My Linh Nguyen, Shengbo Gao, Guangyuan Li, Yinghong Yu, Yan Jiang, Qianlong Zhao, Behzad Bozorgtabar, Shaoxiong Ji, Jiazhen Pan, Daniel Rueckert, and Jiancheng Yang. Are LLMs ready to assist physicians? PhysAssistBench for interactive doctor–patient–EHR assistance. *arXiv preprint arXiv:2606.18613*, 2026.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. AI Hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*, 2024.

Akash Ghosh, Tajamul Ashraf, Rishu Kumar Singh, Numan Saeed, Sriparna Saha, Xiuying Chen, and Salman Khan. CarePilot: A multi-agent framework for long-horizon computer task automation in healthcare. *arXiv preprint arXiv:2603.24157*, 2026. Introduces the CareFlow benchmark.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9), 2025. doi: 10.1056/AIdbp2500144.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Sébastien Jodogne. The Orthanc ecosystem for medical imaging. *Journal of Digital Imaging*, 31 (3):341–352, 2018. doi: 10.1007/s10278-018-0082-y.

Ross Koppel, Joshua P. Metlay, Abigail Cohen, Brian Abaluck, A. Russell Localio, Stephen E. Kimmel, and Brian L. Strom. Role of computerized physician order entry systems in facilitating medication errors. *JAMA*, 293(10):1197–1203, 2005. doi: 10.1001/jama.293.10.1197.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. MedChain: Bridging the gap between LLM agents and clinical practice with interactive sequence. In *Advances in Neural Information Processing Systems*, 2025.

Ruoqi Liu, Imran Q. Mohiuddin, Austin J. Schoeffler, Kavita Renduchintala, Ashwin Nayak, Prasantha L. Vemu, Shivam C. Vedak, Kameron C. Black, John L. Havlik, Isaac Ogunmola, Stephen P. Ma, Roopa Dhatt, and Jonathan H. Chen. PhysicianBench: Evaluating LLM agents in real-world EHR environments. *arXiv preprint arXiv:2605.02240*, 2026.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9(1), 2026. doi: 10.1038/s41746-026-02674-7.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016. doi: 10.7326/M16-0961.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May D. Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Xin Liu, Carl Yang, Yang Xie, and Wenqi Shi. MedAgentGym: A scalable agentic training environment for code-centric reasoning in biomedical data science. *arXiv preprint arXiv:2506.04405*, 2025.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

Yakun Zhu, Zhongzhen Huang, Qianhan Feng, Linjie Mu, Yannian Gu, Shaoting Zhang, Qi Dou, and Xiaofan Zhang. CP-Env: Evaluating large language models on clinical pathways in a controllable hospital environment. *arXiv preprint arXiv:2512.10206*, 2025.

Erik Ziegler, Trinity Urban, Danny Brown, James Petts, Steve D. Pieper, Rob Lewis, Chris Hafey, and Gordon J. Harris. Open health imaging foundation viewer: An extensible open-source framework for building web-based imaging applications to support cancer research. *JCO Clinical Cancer Informatics*, 4:336–345, 2020. doi: 10.1200/CCI.19.00131.