Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Schema-enforced large language model framework for sarcoma tumor board simulation: a methodology development and reproducibility pilot study

Tekoshin Ammo

tekoshin.ammo@uksh.de

University Hospital Schleswig-Holstein

Moritz Englich

University Hospital Schleswig-Holstein

Fabio Dos Santos Adrego

University Hospital Schleswig-Holstein

Maximilian Jacobi

University Hospital Schleswig-Holstein

Alida Wilckens

University Hospital Schleswig-Holstein

Philipp Kruppa

University Hospital Schleswig-Holstein

Bastian Bonaventura

University Hospital Schleswig-Holstein

Stefan Niemuth

University Hospital Schleswig-Holstein

Selina M. Weiler

University Hospital Schleswig-Holstein

Victoria Wachenfeld-Teschner

University Hospital Schleswig-Holstein

Anja M. Boos

University Hospital Schleswig-Holstein

Gerrit Freund

University Hospital Schleswig-Holstein

Article

Keywords: large language models, tumor board, sarcoma, decision support, hallucination, reproducibility, schema validation, clinical artificial intelligence

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10035573/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 11th, 2026. See the published version at https://doi.org/10.1038/s41598-026-74868-8.

Abstract

Background and objectives.

Large language models (LLMs) may support decision-making in multidisciplinary tumor boards (MDTs), but rigorous evaluation is hindered by hallucinations and poor reproducibility, where identical inputs yield divergent recommendations. Sarcomas, with heterogeneous biology, rare subtypes, and reliance on multidisciplinary consensus, represent a demanding test case. We developed and validated a schema-enforced LLM framework producing reproducible, hallucination-resistant sarcoma tumor board recommendations as a basis for subsequent concordance studies against historical MDT decisions.

Methods.

We built an LLM-based tumor board simulator using Anthropic Claude Opus 4.5 (temperature 0.0) producing structured JSON output via tool-use enforcement against a Pydantic schema covering nine clinical domains and 21 controlled-vocabulary decision codes. The methodology was iteratively refined in three phases (free-text Markdown; schema-enforced JSON; semantic clarifications with conditional logic). The final framework (schema v4) was tested on 51 fully de-identified sarcoma cases from a single center (September 2020 – September 2025), each processed three times. The pre-specified primary endpoint was per-case decision-code reproducibility (proportion of codes identical across all three runs). Hallucinations were classified into five pre-specified categories.

Results.

Overall reproducibility was 93.9% (1,006 / 1,071 decision-code instances stable). Seventeen cases (33.3%) achieved 100% reproducibility across all 21 codes. Per-domain reproducibility ranged from 88.9% (radiotherapy) to 98.0% (imaging). Six of 21 codes (29%) reached 100% stability. The reconstruction decision was least stable (74.5%), reflecting clinically interchangeable alternatives. No structurally fabricated content occurred in any of the 154 runs under the final schema.

Conclusions.

Schema-enforced LLM output with explicit semantic and conditional rules achieved high reproducibility and prevented structurally fabricated hallucinations in sarcoma tumor board simulation. Residual variability clustered in clinically equivocal domains, supporting subsequent concordance studies against historical MDT decisions.

Introduction

Multidisciplinary tumor boards (MDTs) are regarded as a central component of contemporary oncological decision-making, integrating the expertise of multiple surgical specialties — including plastic surgery — radiation oncology, medical oncology, radiology, and pathology to formulate consensus treatment recommendations (1, 2). Sarcomas are particularly demanding for MDT decision-making: they comprise more than 100 histological subtypes with diverse biological behavior, are individually rare, frequently require multimodal therapy, and benefit substantially from treatment at specialized centers, because early diagnostic and therapeutic decisions — including biopsy route, subtype classification, margin planning, and multimodal sequencing — directly influence resectability, local control, functional outcome, and the risk of avoidable reoperations (3, 4). Sarcoma MDTs therefore represent a demanding test case for any computational decision-support tool.

Large language models (LLMs) have rapidly emerged as possible tools for MDT decision support, with reported agreement between LLM recommendations and human MDT decisions ranging from approximately 70% to 90% across cancer types (5). In breast cancer, GPT-4 achieved 70.6% concordance with MDT recommendations across complex case profiles, with substantial differences between models and model versions (6). In renal cell carcinoma, a recent pilot study reported 62.1% agreement between LLM and MDT across 103 cases (7). In gynecologic oncology, LLM-enabled digital-twin frameworks have been proposed to support precision-medicine decision-making in complex rare gynecological tumors (8). In otorhinolaryngological oncology, ChatGPT-4o reached 84% concordance and Llama 3 reached 92% concordance with MDT decisions for the binary curative-versus-palliative distinction across 25 cases (9).

In contrast, evidence specifically addressing sarcoma has been more limited and less encouraging. In a recent benchmark against 21 European sarcoma centers using five anonymized soft-tissue sarcoma cases from a published ring trial, LLMs (including GPT-4 (2023), Claude 3.5 Sonnet (2024), and Gemini Pro (2023)) showed inter- and intra-model consistency in only 20% of cases, with concordance to the most common MDT recommendation reaching just 20–60% across models (10). The best-performing model, Claude 3.5 Sonnet, aligned with human consensus in only 60% of cases. Critically, the authors documented that LLM recommendations occasionally contained potentially harmful content and that explicit derivation from established sarcoma guidelines was identifiable in only 24.8–55.2% of responses (10). In a complementary evaluation, our group assessed ChatGPT-4o across all major MDT specialties using systematically prompt-engineered techniques and reported markedly heterogeneous performance: mean scores were highest for surgical specialties (4.48 of 5) and considerably lower for radiation oncology, with an overall mean of 3.76 indicating moderate effectiveness (11). Notably, MDT recommendations were themselves heterogeneous across the 21 sarcoma centers in Li et al. (10), contextualizing but not explaining LLM variability. Both prior studies relied on free-text outputs without schema enforcement, leaving methodological reproducibility largely uncharacterized.

Two interrelated methodological problems emerge from this literature. First, **hallucinations** — the generation of fabricated guideline citations, invented prognostic statistics, or unauthorized output sections — compromise both clinical safety and scientific interpretability (11, 12). Asgari et al. demonstrated that combinations of structured prompting, function calls, and JavaScript Object Notation (JSON)-based outputs systematically reduced hallucination rates in medical text summarization (12). Here, a prompt refers to the natural-language input provided to the LLM that specifies the task, clinical context, expected response format, and any content constraints. These findings suggest that architectural constraints on LLM output may provide stronger safeguards than prompt-level instructions alone. Second, **reproducibility** — the property that identical inputs yield identical outputs — is rarely formally evaluated in oncology LLM studies. The 20% intra-model consistency reported by Li et al. (10) illustrates that without explicit reproducibility constraints, single-shot LLM outputs cannot constitute a stable basis for retrospective concordance studies: comparing a single model run to a fixed historical baseline is misleading because that single run contains both the valid signal (the underlying logic) and stochastic noise (random fluctuation). Any observed difference mainly reflects stochastic fluctuation rather than true model improvement or decline.

Recent advances in LLM tooling have made architectural solutions to both problems available. Tool-use APIs and grammar-constrained decoding now allow strict enforcement of predefined output schemas, eliminating structurally implausible outputs (such as invented sections or non-conforming JSON) at the inference layer rather than via prompt-level instructions (13, 14). Pydantic-based schema definitions enforce controlled vocabularies and field constraints — predefined rules for whether specific fields are required, optional, restricted to certain data types, or allowed only under defined clinical conditions — with run-time validation. Combined with deterministic sampling parameters (temperature 0), these techniques offer a path to reproducible structured output. To our knowledge, schema-enforced output has not been systematically applied to sarcoma MDT simulation and its impact on LLM reproducibility in this domain has not yet been explored.

We therefore set out to develop a schema-enforced LLM framework explicitly designed for reproducible sarcoma MDT simulation (Fig. 1). The framework combines (I) structured tool-use output against a Pydantic-defined schema covering 21 decision codes across nine clinical domains; (II) controlled vocabularies with strict enumerations for all categorical decisions; (III) seven coordinated prompting techniques including role prompting, rule-based prompting, negative prompting, and conditional-logic constraints; and (IV) iterative refinement of semantic specifications across three development phases. We hypothesized that this architecture would yield decision-code reproducibility substantially higher than the 20% reported for unconstrained LLM outputs (10) and would prevent structural hallucinations.

The aims of the present pilot study were (1) to develop and document a schema-enforced LLM framework for sarcoma MDT simulation; (2) to quantify decision-code reproducibility across a diverse pilot dataset of 51 sarcoma cases; (3) to characterize the distribution of remaining variability across clinical domains and individual decision codes; and (4) to provide the methodological foundation for subsequent concordance studies against historical MDT decisions, which are the explicit subject of follow-up work currently in preparation. This pilot study deliberately separates methodology development from clinical validation: we contend that quantitatively reproducible model output is a precondition for any subsequent clinical validation, and that bundling the two would obscure the source of disagreement between LLM and MDT (model variability versus genuine clinical difference). Reporting follows the TRIPOD-LLM guideline for studies of large language models in clinical research (17).

Materials and methods

Study design and ethics

This was a retrospective methodology development and validation study conducted at a single academic center treating sarcoma patients. All methods were carried out in accordance with relevant guidelines and regulations, including the principles of the Declaration of Helsinki. The study, which used fully de-identified clinical case texts, was reviewed by the Ethics Committee of the University of Lübeck (Ethikkommission der Universität zu Lübeck), University Hospital Schleswig-Holstein (UKSH), Campus Lübeck, Germany (reference number / Aktenzeichen 2025 – 520). The need to obtain ethical approval was waived by the Ethics Committee of the University of Lübeck, given the retrospective design and the fully de-identified nature of the data. The requirement for informed consent was also waived by the Ethics Committee of the University of Lübeck for the same reason.

Pilot dataset

A consecutive series of 51 anonymized sarcoma cases was prepared from institutional records covering the period September 2020 – September 2025. The case-selection criterion was prior presentation in the institutional sarcoma multidisciplinary tumor board. The cases show a span of diverse histologies, anatomical sites, grades, ages, and clinical scenarios encountered in routine sarcoma practice. Each case comprised an unstructured clinical narrative covering patient demographics, comorbidities, imaging findings, pathology report, prior treatments where applicable, and the current clinical question.

All cases were fully de-identified prior to LLM submission via an automated anonymization pipeline. Patient names and exact dates of birth were removed (only age and sex retained); all specific calendar dates were replaced with [DATE] placeholders; standalone years were replaced with [YEAR] placeholders; hospital names, ward identifiers, and named clinical personnel were replaced with generic descriptors (e.g., “external reference pathology institute” in place of named pathologists). The anonymization script wrote a per-case audit log documenting all substitutions, which was retained for verification. After automated processing, every case file was manually reviewed for residual identifiers prior to inclusion.

Source case data remain at the institution; aggregated results are deposited in the study code repository (see Data availability).

LLM platform and inference parameters

All inference was performed using Anthropic Claude Opus 4.5, accessed via the official Anthropic Python SDK (version 0.97.0; Python 3.13). Inference was performed at temperature 0.0 (deterministic sampling preference), with a maximum output of 4,000 tokens, and structured-output enforcement via Anthropic's tool-use API. The exact model version, system prompt, and inference parameters are deposited in the code repository and identified by SHA-256 hashes in each run log.

Schema definition

We defined a Pydantic schema (Pydantic version 2.13.3) covering nine clinical domains corresponding to the standard scope of a sarcoma multidisciplinary tumor board: imaging recommendations, staging recommendations, surgical strategy, surgical margin goals, reconstructive strategy, functional reconstruction, radiotherapy strategy, radiotherapy dose, and systemic therapy. Across these domains the schema enforced 21 individual decision codes with strict type constraints (**Table 1**). For each treatment-strategy domain, controlled vocabularies were defined as Python Literal enumerations (**Table 2**). Boolean fields covered imaging modality recommendations (MRI, CT, XR, US, PETCT), specific staging recommendations (CT_CHEST, CT_ABD_PEL), and functional reconstruction sub-components (NERVE_TRANSFER, TENDON_RECON, VASCULAR_RECON). Numeric fields covered radiotherapy total dose in Gray and number of fractions, accepting integers or null (where no radiotherapy was indicated). Free-text fields were limited to two recommendation slots (staging_recommended.OTHER, chemo_1l/chemo_2l) for clinical specifics not expressible in the controlled vocabulary; these fields were retained for content but excluded from strict-equality reproducibility analysis a priori. To reduce semantic ambiguity, all schema fields were supplemented with operational output definitions, explicitly specifying whether positive Boolean values referred to future recommended actions rather than already-completed components of the diagnostic workup.

System prompt design

The system prompt combined seven established prompting techniques in a coordinated architecture (**Table 3**; the full system prompt is provided in Supplementary File 1). In addition, two semantic clarifications were added during development: (i) a "recommended-semantics" rule for imaging and staging Boolean fields, where True indicates that the modality is part of the recommended workup regardless of timing or whether it has already been performed; and (ii) a "definitive-plan semantics" rule for functional reconstruction Booleans, where True is set only if the procedure is definitively part of the planned reconstruction.

Iterative methodology development

Three development phases were conducted on a 10-case pilot subset before scaling to the full case set (**Table 4**). This subset was purposively selected to cover heterogeneous sarcoma presentations, including different histological subtypes, anatomical sites, tumor grades, treatment intents, and levels of clinical complexity, thereby exposing the framework to a broad range of decision scenarios during schema development. The final framework reported below uses schema v4, which additionally added watch_and_wait as a surgical decision option to accommodate desmoid fibromatosis and indolent presentations where active surveillance is the international first-line strategy.

Run protocol and logging

Run protocol and logging
Each of the 51 cases was processed three times under the final framework. For each run, a standardized metadata set was logged (**Table 5**). Runs that failed due to transient API errors were retried automatically, without direct flagging; persistent failures were logged and flagged for inspection. A resume script with consecutive-failure detection enabled robust completion of the full run set across multiple sessions.

Endpoints

The pre-specified **primary endpoint** was per-case decision-code reproducibility, defined as the proportion of the 21 decision codes per case that returned identical values across all three runs of the same case. Secondary reproducibility endpoints were per-code stability across cases and per-domain reproducibility (**Table 6**). The hallucination rate, stratified into five pre-specified categories (H1–H5), was reported as a structural-integrity endpoint.

Reproducibility metric

For each decision code in each case, all three runs were compared for strict equality of returned values. A code was considered “stable” within a case if and only if all three runs returned identical values; otherwise it was counted as “unstable” for that case. Reproducibility was reported at three levels of aggregation, each answering a different question: case-level for clinical interpretability, code-level to identify systematically problematic decisions, and domain-level to align with the schema architecture (**Table 6**).

Hallucination classification

Outputs were screened for content not derivable from the input case text. Five categories were defined a priori before manuscript drafting (**Table 7**). Hallucination screening for the present pilot was performed by the first author; multi-rater adjudication is a planned extension.

Statistical analysis

Reproducibility metrics are reported as proportions and percentages. Given the descriptive and methodological nature of this pilot, no inferential statistical tests were applied; confidence intervals are not reported, as the unit of variability of interest (intra-model run variation) is itself the studied quantity rather than a sample-to-population inference target.

Results

Dataset characteristics

The pilot dataset comprised 51 anonymized sarcoma cases (**Table 8**). Median patient age was 73 years (range 28–93; age extracted in 48/51 cases [94.1%]). The age distribution was skewed toward older patients, with 26 cases (51.0%) in the ≥ 70-years age group and 19 (37.3%) in the 50–69-years age group, consistent with the demographic profile of soft-tissue sarcoma in routine clinical practice. Sex distribution was 25 male (49.0%) and 23 female (45.1%) cases (sex extracted in 48/51 cases [94.1%]).

Histological distribution was diverse, comprising 17 distinct subtypes (Fig. 2). The most frequent subtypes were atypical lipomatous tumor / well-differentiated liposarcoma (n = 11, 21.6%), synovial sarcoma (n = 6, 11.8%), leiomyosarcoma (n = 6, 11.8%), dedifferentiated liposarcoma (n = 4, 7.8%), and myxofibrosarcoma (n = 4, 7.8%). Together, the liposarcoma family (atypical lipomatous tumor, dedifferentiated, myxoid, and not-otherwise-specified liposarcomas) accounted for 18

cases (35.3%). Less frequent histologies included dermatofibrosarcoma protuberans (n = 3, 5.9%), undifferentiated pleomorphic sarcoma (n = 3, 5.9%), rhabdomyosarcoma (n = 3, 5.9%), and nine additional individual subtypes each represented by one or two cases.

Anatomical sites spanned multiple body regions (Fig. 3). Tumor grade was G3/high-grade in 27 cases (52.9%), G2/intermediate in 5 (9.8%), G1/low-grade in 14 (27.5%), and not specified in 5 (9.8%).

Iterative development outcomes

In Phase 1 (free-text Markdown output), exploratory inspection of pilot-case outputs revealed multiple hallucination categories. The model generated invented guideline citations (e.g., references to European Society for Medical Oncology (ESMO) and National Comprehensive Cancer Network (NCCN) sections without textual basis in the input), fabricated prognostic statistics (e.g., specific 5-year survival percentages), and unauthorized output sections beyond the requested scope (e.g., “Surveillance Plan”, “Clinical Trial Recommendations”). Decision-code identification across re-runs of the same case yielded inconsistent values, with structural variability dominating run-to-run differences.

In Phase 2, schema-enforced JSON output via tool use eliminated structural hallucinations: non-conforming outputs were rejected at the API level, and unauthorized sections were architecturally impossible. Nevertheless, exploratory qualitative review of pilot cases suggested residual run-to-run instability, mainly in Boolean decision fields. The instability appeared to arise less from JSON-structure errors than from underspecified semantics — for example, whether True for “MRI recommended” referred to an MRI that should still be performed, or to MRI as part of the overall workup, including studies already completed. In Phase 3 (semantic clarification and conditional-logic constraints), the addition of explicit semantic rules and coupled-decision constraints resolved most of the remaining Boolean ambiguity. Schema v4 — which additionally added watch_and_wait as a surgical option — was used for the full validation run reported below.

Reproducibility — primary endpoint

The full validation run comprised 154 LLM runs across 51 cases (50 cases × 3 runs + 1 case × 4 runs due to a re-run during pilot validation). Across 1,071 decision-code instances (51 cases × 21 codes, with one additional code-instance set from the re-run), 1,006 instances were stable across all runs of their respective case, corresponding to an overall reproducibility of **93.9%** (**Table 9a**; **Fig. 4**).

Per-case reproducibility (Fig. 4) ranged from 61.9% to 100% across the 51 cases. Seventeen cases (33.3%) achieved 100% reproducibility, with all 21 decision codes returning identical values across all three runs. A further 26 cases (51.0%) achieved ≥ 95% reproducibility (20 of 21 codes stable). All but one case achieved ≥ 80% reproducibility. The single case below the 80% threshold (case 44, 61.9% reproducibility) corresponded to a clinically ambiguous elderly patient in whom multiple decision codes (surgical strategy, margins goal, reconstruction, radiotherapy decision and dose) showed concurrent run-to-run variability — consistent with the case representing a genuine clinical decision-making boundary rather than isolated model instability.

Per-domain reproducibility (**Table 9b**; **Fig. 5**) followed clinically interpretable patterns. Imaging modality recommendations achieved the highest stability at 98.0% (250 of 255 code-instances stable). Systemic therapy decisions reached 97.1% (198/204). Surgery and reconstruction domains both reached 93.1% (95/102 and 190/204, respectively). Staging recommendations reached 89.5% (137/153), and radiotherapy reached 88.9% (136/153). The lower stability in radiotherapy was driven primarily by variation in dose specifications (50/25, 60/30, and 66/33 fractionation schemes).

Per-code stability across cases (**Table 9c**; **Fig. 6**) revealed that 6 of 21 individual decision codes (29%) achieved 100% stability. These were: imaging recommendations for ultrasound and PET-CT, the staging recommendation for CT chest, functional reconstruction Booleans for nerve transfer and tendon reconstruction, and the second-line chemotherapy and

immunotherapy/targeted-therapy fields. The lowest-stability code was the reconstruction decision (74.5%, 38/51), where the model occasionally oscillated between adjacent options such as none, local_flap, and pedicled_flap — reflecting genuine clinical equipoise in soft-tissue reconstruction. The free-text staging field (staging_recommended.OTHER) reached 76.5%; manual inspection indicated that several apparent disagreements reflected semantically equivalent rephrasings rather than substantive differences in recommended workup.

Distribution of reproducibility (Fig. 7) was concentrated in the high range, with median per-case reproducibility 95.2% and mean 93.8%. The distribution was strongly left-skewed, with one outlier at 61.9% (case 44).

Use of the watch_and_wait surgical option

The watch_and_wait surgical option, added in schema v4, was selected by the model in five cases (cases 10, 16, 36, 39, and 44). All five cases involved older patients (≥ 70 years), indolent histologies (e.g., desmoid fibromatosis context), or clinically borderline indications for surgical intervention. In each of these five cases, watch_and_wait was selected in some but not all runs — indicating that the model used the option specifically in scenarios where multiple management strategies are clinically defensible, rather than defaulting to it.

Hallucination assessment

Across all 154 runs of the final framework, no instances of H1, H2, or H3 hallucinations were observed.

H1 (numerical prognostic claims): no fabricated survival statistics, recurrence percentages, or other specific numerical prognostic figures appeared in any output.

H2 (named guidelines, trials, or studies): no invented guideline references (ESMO, NCCN, EORTC, etc.) or named trial citations appeared in any output.

H3 (anatomical or imaging measurements): no measurements were introduced beyond those reported in the input.

H4-class general clinical-knowledge claims occurred occasionally (e.g., statements such as “this histology is generally chemosensitive”) but were not counted as hallucinations per the a-priori definition; manual inspection did not identify factually incorrect H4 statements in the screened sample.

H5-class patient-specific reasoning extensions were rare and, where observed, were clinically reasonable.

Discussion

In this methodological pilot study, a schema-enforced LLM framework for sarcoma multidisciplinary tumor board simulation achieved 93.9% overall reproducibility across 21 decision codes in 51 cases (1,006 of 1,071 stable code-instances over three runs per case). Seventeen of 51 cases (33.3%) showed full reproducibility across all decision codes, and structurally fabricated content — invented guideline citations, fabricated prognostic statistics, and unauthorized output sections — was absent across all 154 runs under the final schema. The remaining variability concentrated in clinical domains with genuine equipoise: reconstruction decisions (74.5% stability), free-text staging recommendations (76.5% stability), and radiotherapy dose specifications (86.3% stability), where alternative options correspond to clinically accepted standards rather than obvious model errors. To our knowledge, this is the first systematic evaluation of schema-enforced LLM output in sarcoma tumor board simulation and provides a methodological foundation for subsequent concordance studies against historical multidisciplinary decisions.

The reproducibility achieved here contrasts sharply with prior reports of unconstrained LLM outputs in sarcoma. Li and colleagues, benchmarking GPT-4, Claude 3.5 Sonnet, and Gemini Pro against 21 European sarcoma centers, documented intra- and inter-model consistency in only 20% of five soft-tissue sarcoma cases (10). Their best-performing

model (Claude 3.5 Sonnet) aligned with human consensus in 60% of cases, and explicit derivation from established guidelines was identifiable in only 24.8–55.2% of responses. While direct comparison is limited by different metrics, case sets, and prompting strategies, the intra-model consistency showed an approximately five-fold difference (20% vs. 93.9%), which is consistent with the architectural prediction that schema enforcement should substantially stabilize categorical decision outputs. The present framework was designed in part to address the previously shown domain heterogeneity (11) by encoding the seven prompting techniques recommended in that work into a coordinated schema-constrained architecture.

In other oncology contexts, reported LLM–MDT concordance has ranged from approximately 50% to 90%, with values above 80% achieved primarily in domains with structured decision spaces (e.g., binary curative-versus-palliative classification in head-and-neck cancer) (8). In renal cell carcinoma, schema-free LLM output achieved 62.1% agreement with the multidisciplinary team across 103 cases (7). In breast cancer, GPT-4 reached 70.6% concordance, with substantial variation between model versions (6). None of these studies systematically quantified intra-model reproducibility, leaving open the question of whether observed concordance values reflect stable model behavior or single realizations of variable outputs. The present pilot supports the interpretation that, under appropriate architectural constraints, this question can be made tractable.

Beyond the ring-trial benchmark by Li and colleagues (10), a growing body of work has tested LLMs against sarcoma-specific guideline corpora and individual sarcoma cases. Matsuoka and colleagues evaluated ChatGPT-4 alignment with the Japanese Orthopaedic Association guidelines for soft-tissue sarcoma management and reported an overall alignment rate of 86% (two complete and 17 partial alignments out of 19 cases), but flagged 14% of responses as non-aligned with current practice, underscoring the need for continued validation (18). In a separate study against the German evidence-based S3 guideline for adult soft-tissue sarcoma, Li and colleagues compared ChatGPT-3.5 and ChatGPT-4 on 80 clinical questions; GPT-4 outperformed GPT-3.5 in accuracy and adequacy, particularly for general STS treatment and extremity/trunk sarcoma questions, but both models occasionally produced misleading or potentially dangerous output, again highlighting the need for human oversight (19). Valentini and colleagues assessed ChatGPT on 25 sarcoma-specific questions evaluated by three sarcoma experts across five criteria (completeness, misleadingness, accuracy, timeliness, appropriateness), reporting a median score of 18.3 of 25; performance was strong for general questions and definitions but markedly weaker for treatment-related queries, with only 45% of responses rated good or very good, and the authors cautioned sarcoma patients about the risk of LLM-derived misinformation (20).

The cross-tumor literature paints a similarly mixed picture. In breast cancer, Sorin and colleagues tested ChatGPT-3.5 against the recommendations of a multidisciplinary breast tumor board for 10 patients and reported alignment in 70% of cases (21); Lukac and colleagues, evaluating ChatGPT-3.5 as an adjunct to the multidisciplinary tumor board in primary breast cancer, similarly described moderate, case-dependent performance and emphasized that LLMs are not yet reliable enough for autonomous tumor-board use (22). In gastric cancer, Park and Chae retrospectively compared AI-generated treatment recommendations to MDT decisions and reported moderate concordance, limited by the absence of structured output and verifiable guideline derivation (23). In gastrointestinal oncology, Aghamaliyev and colleagues described ChatGPT's gastrointestinal tumor-board performance as inconsistent — captured in their "limping dance partner" framing of the LLM as a junior collaborator that requires constant supervision (24). In musculoskeletal oncology, Truhn and colleagues showed that GPT-4 could derive plausible orthopaedic treatment recommendations from MRI reports but without any structured-output guarantee or formal hallucination accounting (25). In head-and-neck oncology, Schmidl and colleagues directly compared Claude 3 Opus and ChatGPT-4 on 50 head-and-neck cancer cases; Claude 3 Opus produced more accurate diagnostic workups, but neither model relied on prompt-engineering techniques designed to reduce hallucination, and the authors concluded that both should be regarded merely as adjuncts to the MDT (26).

Across these heterogeneous efforts three problems recur: free-text outputs that defy systematic comparison, hallucinated guideline references or fabricated statistics, and the complete absence of intra-model reproducibility metrics. Most prior studies further relied on schema-free prompting without enforcing decision-code grammars, which makes their concordance estimates inherently confounded with run-to-run model variability. The architectural approach taken in the present work — schema-enforced tool-use output combined with controlled vocabularies and conditional-logic constraints — addresses these problems at the inference layer rather than via post-hoc filtering, and supplies the reproducibility baseline that a meaningful clinical concordance study has so far been lacking.

Three architectural choices appear to drive the observed reproducibility. First, controlled vocabularies with strict enumerations eliminate the stylistic variability that dominates free-text outputs: the model can no longer choose between "*wide local excision*", "*wide resection with R0 margins*", or "*definitive surgical excision*" — only wide_excision is grammatically valid. Second, tool-use enforcement at the inference layer rejects schema-non-conforming outputs structurally, rather than relying on prompt-level instruction adherence; this provides a stronger backstop than prompt engineering alone, consistent with the framework proposed by Asgari and colleagues for medical text summarization (12). Third, conditional-logic constraints (e.g., systemic_decision = "no" forces chemo_1l = "none") prevent semantically inconsistent code combinations that can otherwise emerge from independent token-level sampling decisions across distant fields. Together, these mechanisms shift the locus of variability away from output structure and toward genuine clinical decision points, where some variability is desirable and reflective of real practice. By standardizing fields, labels, and response format, the schema reduces avoidable differences caused by inconsistent wording or missing information; as a result, variation between outputs is more likely to reflect meaningful clinical uncertainty rather than technical noise.

This is particularly important in sarcoma care, where treatment decisions often depend on factors such as tumor subtype, anatomical site, resectability, patient fitness, morbidity, and treatment intent. In these situations, more than one management strategy may be clinically reasonable. The schema therefore supports comparability without forcing artificial uniformity, allowing reviewers to distinguish between errors in structure and legitimate differences in clinical judgment.

The remaining 6.1% of decision-code instability did not occur randomly across the schema. It concentrated in domains where multiple options are clinically defensible. In reconstruction (74.5% stability), the model occasionally oscillated between adjacent options such as no flap, local flap, and pedicled flap — choices that depend on intra-operative findings and surgeon preference rather than on a single guideline-mandated answer. In radiotherapy dose (86.3% stability for both Gy total and fractions), variation occurred between 50/25, 60/30, and 66/33 fractionation schemes, all of which are within accepted standard ranges for soft-tissue sarcoma. In free-text staging recommendations (76.5% stability), instability frequently reflected semantically equivalent rephrasings ("*Liver MRI may be considered for full characterization*" versus "*MRI liver may be considered for characterization*") rather than substantive differences.

The lowest-reproducibility case (case 44, 61.9%) was a clinically ambiguous elderly patient in whom watch-and-wait, wide excision, and various reconstructive choices all had legitimate justifications — a scenario in which an MDT itself would likely show internal disagreement. Variability of this kind reflects clinical reality rather than model instability and may, in fact, be a desirable property of any decision-support tool.

Several schema-design implications emerged during development. The watch_and_wait surgical option was added in schema v4 specifically because the original schema implicitly assumed a surgical intervention for every case — an inappropriate constraint for desmoid fibromatosis, indolent lipomatous tumors, and selected cases with a high patient age. Its selective use in five cases (10, 16, 36, 39, 44), all involving elderly patients or borderline indications, suggests the model uses the option appropriately when available.

These refinements are important because they connect the schema more closely to real-world sarcoma care, where treatment decisions are rarely binary and often depend on tumor biology, anatomical site, patient fitness, expected morbidity, and patient preference. For clinicians, a schema that better reflects these realities may improve interpretability and trust: the output is easier to review when it mirrors the options and trade-offs discussed in multidisciplinary tumor boards. This was a key reason for developing the schema iteratively rather than treating it as a fixed technical framework. Each revision helped identify where the structure was too rigid, where clinically meaningful options were missing, and where model outputs risked appearing complete while failing to represent the actual decision space.

The schema is primarily intended for clinicians, researchers, and multidisciplinary sarcoma teams who need structured, comparable outputs from complex case descriptions. Its benefit for patients is indirect but clinically relevant: by allowing options such as observation, palliative-intent management, or procedure-specific surgery types to be represented explicitly, the schema may reduce overtreatment-oriented framing and better preserve patient-centered considerations such as frailty, quality of life, and treatment goals. From a user-friendliness perspective, a more clinically aligned schema also lowers the cognitive burden for reviewers, because fewer outputs require manual reinterpretation or correction before they can be used for audit, research, or decision-support evaluation.

Several limitations should be acknowledged. First, this is a single-center pilot of 51 cases; while case selection spanned diverse histologies, anatomical sites, grades, and treatment intents, generalizability to other centers and populations is untested. Second, the framework was evaluated on a single LLM (Claude Opus 4.5); whether comparable reproducibility would be achieved with other contemporary models — and whether schema enforcement reduces inter-model differences — is the subject of separate ongoing work in our group. Third, and most importantly, this study deliberately addresses methodological reproducibility only and does not assess clinical plausibility or accuracy of LLM recommendations. We elected to separate these aims because reproducible output is a logical prerequisite for any meaningful clinical concordance study: comparing a single realization of a non-deterministic model to historical MDT decisions confounds genuine clinical disagreement with stochastic model variability. Formal multi-rater clinical plausibility assessment and concordance analysis against historical MDT decisions are the explicit subject of follow-up work currently in preparation. Fourth, the schema is sarcoma-specific; extension to other cancer types would require de-novo schema redevelopment, although the methodological architecture should transfer. Fifth, the framework was developed and tested in English; multilingual extension is untested. Sixth, the anonymization process replaced specific dates with placeholders, which may have removed clinically relevant temporal context (e.g., interval to recurrence, time from diagnosis to initial treatment, and delays between successive treatment modalities) in some cases. Finally, although hallucination categories were defined a priori and outputs were screened systematically, formal blinded multi-rater hallucination adjudication was not performed; the absence of structural hallucinations under the final schema is a strong observation but should be interpreted in this methodological context.

This pilot enables several lines of follow-up. First, a clinical concordance study comparing schema-enforced LLM recommendations to documented historical MDT decisions in the same case set is currently in preparation and will provide the clinical validation that the present study deliberately excludes. Second, a multi-model comparison applying the identical schema and prompting architecture to GPT-class and Gemini-class models will quantify how much model choice contributes to recommendation variability when output structure is held constant. Third, schema extension to other sarcoma-relevant clinical scenarios — neoadjuvant response assessment, metastatic decision-making, palliative pathway selection — will test the framework's flexibility. Fourth, the methodological architecture is in principle transferable to other rare cancers with structured MDT decision spaces, and schema templates for these contexts could be developed. Beyond these planned studies, future work could explore additional methodological extensions such as semantic and network-based representations of MDT and LLM decision-making, qualitative analyses of reasoning processes, comparisons between human and model reproducibility, uncertainty-aware recommendation frameworks, and systematic ablation experiments to identify the architectural drivers of reproducibility and robustness.

Regarding follow-up studies, future schema iterations may also benefit from additional refinements, such as a separate code for Mohs micrographic surgery (relevant for dermatofibrosarcoma protuberans), tolerance-aware comparison logic for free-text fields to account for semantic equivalence, and explicit handling of palliative-intent scenarios where the schema's curative-intent default may be inappropriate.

The clinical implications of this pilot should be framed cautiously. Reproducibility is a necessary but insufficient property of a clinical AI tool: a perfectly reproducible model can still be reproducibly wrong. The present study establishes that schema enforcement makes LLM outputs *evaluable* — that is, capable of supporting the kind of systematic concordance and accuracy studies that have until now been confounded by model variability and structural hallucinations. It does not, by itself, establish that schema-enforced LLMs are ready for use in tumor board decision support. Translation to clinical use will require, at minimum, the planned concordance study, multi-rater plausibility assessment, prospective evaluation, and integration with existing clinical workflows under appropriate human oversight as recommended by the TRIPOD-LLM reporting framework (17).

Conclusions

Schema enforcement combined with explicit semantic rules and conditional-logic constraints provides a methodological foundation for reproducible LLM-based sarcoma tumor board simulation. In a 51-case pilot with 154 runs, overall decision-code reproducibility reached 93.9%, structural hallucinations were eliminated, and remaining variability concentrated in clinical domains of genuine equipoise. This framework enables the systematic clinical concordance studies that until now have been precluded by model variability, and it provides a transferable architecture for analogous applications in other cancer types. Clinical translation requires the formal validation studies that are the explicit subject of follow-up work currently in preparation.

Glossary

JSON (JavaScript Object Notation): a lightweight, human-readable data-interchange format used here for the structured tumor-board output.

Pydantic / Pydantic validation: a Python data-validation library (version 2.13.3 in this study) that enforces type constraints, allowed values, and conditional rules on each schema field at run time.

Tool-use API: an interface provided by the LLM vendor that allows the caller to specify a strict output schema; non-conforming outputs are rejected at the inference layer rather than via prompt-level instructions.

Zero-shot prompting: a prompting strategy in which the model is asked to perform a task without being shown completed input/output examples in the prompt.

Few-shot semantic anchoring: a prompting strategy in which a small number of brief examples are provided to clarify the intended semantics of specific fields, without giving full output exemplars.

Reproducibility (in this study): the property that identical inputs yield identical outputs; quantified here as the proportion of decision codes whose values are identical across three runs of the same case.

Hallucination (LLM): the generation by an LLM of content not derivable from the input — for example, fabricated guideline citations, invented prognostic statistics, or unauthorized output sections (12).

Declarations

Data availability

The code, schema definitions, the system prompt text (with its SHA-256 hash), and the aggregated reproducibility results generated and analysed during the current study are available in the sarcoma-tumorboard-llm repository on GitHub: https://github.com/tekoshinammo-dotcom/sarcoma-tumorboard-llm. The source case data are not publicly available because they contain potentially identifying clinical information and are subject to data-protection requirements under the EU General Data Protection Regulation (GDPR) and institutional governance, but de-identified data are available from the corresponding author (T.A.) on reasonable request and subject to a data-use agreement and institutional approval.

Ethics statement

This retrospective methodological pilot study used fully de-identified case texts and was carried out in accordance with relevant guidelines and regulations, including the principles of the Declaration of Helsinki. The study was reviewed by the Ethics Committee of the University of Lübeck (Ethikkommission der Universität zu Lübeck), University Hospital Schleswig-Holstein (UKSH), Campus Lübeck, Germany (reference number / Aktenzeichen 2025-520). The need to obtain ethical approval was waived by the Ethics Committee of the University of Lübeck, given the retrospective design and the fully de-identified nature of the data. The requirement for informed consent was also waived by the Ethics Committee of the University of Lübeck for the same reason. No patients were treated, contacted, or recruited as part of this methodological study; all analyses were performed on de-identified textual records of cases that had previously been presented to the institutional sarcoma multidisciplinary tumor board as part of routine clinical care.

Author contributions

TA: conceptualization, methodology, software, investigation, formal analysis, data curation, writing — original draft. **ME:** methodology, validation, writing — review and editing. **MJ:** methodology, validation, writing — review and editing. **AW:** methodology, validation, writing — review and editing. **FDSA:** investigation, writing — review and editing. **PK:** investigation, writing — review and editing. **BB:** investigation, writing — review and editing. **SN:** investigation, writing — review and editing. **SMW:** investigation, writing — review and editing. **VW-T:** investigation, writing — review and editing. **AMB:** supervision, writing — review and editing. **GF:** supervision, writing — review and editing. All authors approved the final manuscript.

Funding

The authors declare that no external funding was received for the research, authorship, or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that generative AI tools were used during the preparation of this manuscript for language polishing and to assist with drafting the schema and analysis code described in the Methods. All scientific content, study design, data analysis, interpretation, and conclusions were generated and verified by the authors.

Acknowledgements

The authors thank the institutional sarcoma multidisciplinary tumor board for prior case discussion underpinning the case set used in this methodology study.

References

1. Pillay, B. et al. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: a systematic review of the literature. *Cancer Treat. Rev.* **42**, 56–72. 10.1016/j.ctrv.2015.11.007 (2016).
2. Specchia, M. L. et al. The impact of tumor board on cancer care: evidence from an umbrella review. *BMC Health Serv. Res.* **20** (1), 73. 10.1186/s12913-020-4930-3 (2020).
3. Gronchi, A. et al. Soft tissue and visceral sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **32** (11), 1348–1365. 10.1016/j.annonc.2021.07.006 (2021).
4. Strauss, S. J. et al. Bone sarcomas: ESMO–EURACAN–GENTURIS–ERN PaedCan Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **32** (12), 1520–1536. 10.1016/j.annonc.2021.08.1995 (2021).
5. Wang, X. et al. Artificial intelligence in multidisciplinary tumor boards enhancing decision making and clinical outcomes in oncology. *iScience* **28** (12), 114082. 10.1016/j.isci.2025.114082 (2025).
6. Griewing, S. et al. Evolution of publicly available large language models for complex decision-making in breast cancer care. *Arch. Gynecol. Obstet.* **310**, 537–550. 10.1007/s00404-024-07565-4 (2024).
7. Bertolo, R. et al. Role of large language models in the multidisciplinary decision-making process for patients with renal cell carcinoma: a pilot experience. *NPJ Precis Oncol.* **9**, 257. 10.1038/s41698-025-01014-4 (2025).
8. Lammert, J. et al. Large language models-enabled digital twins for precision medicine in rare gynecological tumors. *NPJ Digit. Med.* **8**, 420. 10.1038/s41746-025-01810-z (2025).
9. Buhr, C. R. et al. Assessment of decision-making with locally run and web-based large language models versus human board recommendations in otorhinolaryngology, head and neck surgery. *Eur. Arch. Otorhinolaryngol.* **282**, 1593–1607. 10.1007/s00405-024-09153-3 (2025).
10. Li, C. P. et al. The imitation game: large language models versus multidisciplinary tumor boards: benchmarking AI against 21 sarcoma centers from the ring trial. *J. Cancer Res. Clin. Oncol.* **151**, 248. 10.1007/s00432-025-06304-9 (2025).
11. Ammo, T. et al. Evaluating ChatGPT-4o as a decision support tool in multidisciplinary sarcoma tumor boards: heterogeneous performance across various specialties. *Front. Oncol.* **14**, 1526288. 10.3389/fonc.2024.1526288 (2025).
12. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit. Med.* **8** (1), 274. 10.1038/s41746-025-01670-7 (2025).
13. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55** (12), 1–38. 10.1145/3571730 (2023).
14. Achiam, J. et al. GPT-4 technical report. arXiv preprint. (2023). 10.48550/arXiv.2303.08774
15. Bai, Y. et al. Constitutional AI: harmlessness from AI feedback. *arXiv preprint*. 10.48550/arXiv.2212.08073 (2022).
16. Anthropic Tool use with Claude [Internet]. Anthropic Documentation; [accessed 2026 May 4]. (2024). Available from: https://docs.anthropic.com/claude/docs/tool-use
17. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31** (1), 60–69. 10.1038/s41591-024-03425-5 (2025).

18. Matsuoka, M. et al. Evaluating the alignment of artificial intelligence-generated recommendations with clinical guidelines focused on soft tissue tumors. *J. Surg. Oncol.* 10.1002/jso.27874 (2024).
19. Li, C. P. et al. Comparing ChatGPT-3.5 and ChatGPT-4's alignments with the German evidence-based S3 guideline for adult soft tissue sarcoma. *iScience* **27**, 111493. 10.1016/j.isci.2024.111493 (2024).
20. Valentini, M. et al. Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients? *Front. Public. Health*. **12**, 1303319. 10.3389/fpubh.2024.1303319 (2024).
21. Sorin, V. et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. **9**, 44. 10.1038/s41523-023-00557-8 (2023).
22. Lukac, S. et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch. Gynecol. Obstet.* **308**, 1831–1844. 10.1007/s00404-023-07130-5 (2023).
23. Park, Y. E. & Chae, H. The fidelity of artificial intelligence to multidisciplinary tumor board recommendations for patients with gastric cancer: a retrospective study. *J. Gastrointest. Cancer*. **55**, 365–372. 10.1007/s12029-023-00967-8 (2024).
24. Aghamaliyev, U. et al. ChatGPT's Gastrointestinal Tumor Board Tango: a limping dance partner? *Eur. J. Cancer*. **205**, 114100. 10.1016/j.ejca.2024.114100 (2024).
25. Truhn, D. et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci. Rep.* **13**, 20159. 10.1038/s41598-023-47500-2 (2023).
26. Schmidl, B., Hütten, T., Pigorsch, S. & Stögbauer, F. Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. *Eur. Arch. Otorhinolaryngol.* **281**, 6099–6109. 10.1007/s00405-024-08828-1 (2024).

Tables

Table 1. Schema definition — decision codes by clinical domain.

Domain	Decision code	Type	Allowed values / range	Output semantics
Imaging recommendations	imaging_recommended.MRI	Boolean	True / False	True if MRI is recommended as a future diagnostic step; False if not recommended or already completed without further need.
	imaging_recommended.CT	Boolean	True / False	True if CT is recommended as a future diagnostic step.
	imaging_recommended.XR	Boolean	True / False	True if X-ray is recommended as a future or pending imaging modality.
	imaging_recommended.US	Boolean	True / False	True if ultrasound is recommended as a future or pending imaging modality.
	imaging_recommended.PETCT	Boolean	True / False	True if PET-CT is recommended as a future or pending imaging modality.
Staging recommendations	staging_recommended.CT_CHEST	Boolean	True / False	True if CT chest is recommended for staging and remains to be performed.
	staging_recommended.CT_ABD_PEL	Boolean	True / False	True if CT abdomen/pelvis is recommended for staging and remains to be performed.
	staging_recommended.OTHER	Free text	≤ 300 characters or null	Additional staging modality or procedure recommended beyond predefined fields; null if none.

Surgical strategy	surgery_decision	Enum	none / biopsy / wide_excision / compartment_resection / amputation / watch_and_wait	Selected value represents the next intended surgical management step.
	margins_goal	Enum	R0_intended / R1_acceptable / unclear	Selected value represents the intended margin strategy for the planned operation.
Reconstruction	recon_decision	Enum	none / skin_graft / local_flap / pedicled_flap / free_flap	Selected value represents the planned reconstructive approach.
	functional_recon.NERVE_TRANSFER	Boolean	True / False	True if nerve transfer is explicitly recommended or planned.
	functional_recon.TENDON_RECON	Boolean	True / False	True if tendon reconstruction or transfer is recommended.
	functional_recon.VASCULAR_RECON	Boolean	True / False	True if vascular reconstruction is recommended.
Radiotherapy	rt_decision	Enum	no / neoadjuvant / adjuvant / intraoperative	Selected value represents the intended radiotherapy strategy.
	rt_dose.GY_TOTAL	Integer	0–80 or null	Numeric total planned radiotherapy dose in Gy; null if not specified.
	rt_dose.FRACTIONS	Integer	0–40 or null	Number of radiotherapy fractions if explicitly specified; null if not specified.
Systemic therapy	systemic_decision	Enum	no / neoadjuvant / adjuvant / palliative	Recommended systemic-therapy strategy and treatment intent / timing.
	chemo_1l	Free text	Controlled regimen string or “none”	First-line regimen only; use “none” if no systemic regimen is planned.

	chemo_2l	Free text	Controlled regimen string or “none”	Second-line or subsequent systemic regimen recommended; “none” if not recommended.
	io_target	Enum	yes / no / unclear	“yes” if immunotherapy or targeted therapy is recommended; “unclear” if discussed but not resolved.

Schema version 4; 21 decision codes across 6 treatment domains. Field types: Boolean (True/False flag), Enum (a single value from a fixed controlled vocabulary), Integer (integer or null), Free text (length-limited string or null). Free-text fields (staging_recommended.OTHER, chemo_1l, chemo_2l) are constrained by length and by conditional-logic rules (e.g., chemo_1l must be “none” if systemic_decision = "no"). All field constraints are enforced via Pydantic 2.13.3 with strict Literal types and validated at the inference layer via Anthropic’s tool-use API.

Table 2. Controlled vocabularies for categorical decisions (Pydantic Literal enumerations).

Decision code	Allowed values
surgery_decision	none / biopsy / wide_excision / compartment_resection / amputation / watch_and_wait
margins_goal	R0_intended / R1_acceptable / unclear
recon_decision	none / skin_graft / local_flap / pedicled_flap / free_flap
rt_decision	no / neoadjuvant / adjuvant / intraoperative
systemic_decision	no / neoadjuvant / adjuvant / palliative
io_target (immunotherapy / targeted)	yes / no / unclear

Each categorical decision is restricted to one value from a fixed enumeration enforced by Pydantic Literal types and validated at the inference layer through the tool-use API.

Table 3. The seven coordinated prompting techniques combined in the system prompt (schema v4).

#	Technique	Implementation in this study
1	Role prompting	Explicit domain identity (“Sarcoma-Tumorboard-LLM-v4”).
2	Rule-based prompting	Six numbered absolute rules, including “use only information present in the case”, “never fabricate guideline references”, “never produce numerical prognostic statistics”, and “never introduce sections beyond the schema” (analogous to Constitutional AI (15)).
3	Negative prompting	Explicit prohibition of fabricated trial names, named guidelines (ESMO, NCCN, EORTC) unless quoted from the input, and specific survival or recurrence percentages.
4	Schema-constrained output	Tool-use API (16) validates each output against the Pydantic schema at the inference layer; non-conforming responses are rejected by the API rather than relying on prompt-level instruction adherence.
5	Few-shot semantic anchoring	Brief domain-specific examples for Boolean field interpretation without providing complete output exemplars.
6	Conditional-logic encoding	Coupled-decision constraints (e.g., systemic_decision = "no" forces chemo_1l = "none"; rt_decision = "no" forces rt_dose.GY_TOTAL and rt_dose.FRACTIONS to null).
7	Controlled vocabulary	Strict enumerations enforced by Pydantic Literal types (see Table 2).

Table 4. The three iterative methodology development phases.

Phase	Technical change	Key observation / problem addressed
Phase 1 — Free-text Markdown output	Baseline system prompt, unstructured Markdown recommendations.	Free-text outputs exhibited multiple categories of hallucination (fabricated guideline citations, invented prognostic statistics, unauthorized output sections such as “Surveillance Plan” and “Clinical Trial Recommendations”).
Phase 2 — Schema-enforced JSON output	Pydantic schema + tool-use API enforcement; structured, machine-validated output.	Structural hallucinations (invented sections, non-conforming output) were eliminated by API-level validation; semantic ambiguities in Boolean fields produced run-to-run instability in 4–5 decision codes per pilot case.
Phase 3 — Semantic clarification and conditional logic (schema v4)	Explicit Boolean-field semantics, conditional-logic constraints; addition of watch_and_wait as a surgical option.	Remaining Boolean ambiguity resolved; final framework used for full validation run.

Table 5. Per-run metadata logged for each of the 154 inference runs.

Metadata field	Description
Prompt hash	SHA-256 hash of the system prompt (truncated to 12 hexadecimal characters).
Case hash	SHA-256 hash of the case text.
Model identifier	claude-opus-4-5.
Inference parameters	Temperature, max tokens.
Wall-clock timestamp	Date and time of inference.
Token usage	Input and output tokens.
Schema version	v4.
Run index	1, 2, or 3.

Table 6. Three reproducibility aggregation levels.

Aggregation level	Definition	Question answered
Per-case reproducibility (primary endpoint)	Number of stable decision codes divided by 21 codes per case.	Which cases yield the most stable recommendations?
Per-code stability across cases	Number of cases in which the code was stable divided by total number of cases (n = 51).	Which decision codes are systematically problematic?
Per-domain reproducibility	Number of stable code-instances within a clinical domain divided by total code-instances in that domain across all cases.	Which clinical domains contribute most to remaining variability?

Table 7. Pre-specified hallucination categories (H1–H5).

Category	Definition	Counting rule
H1 — Numerical prognostic claims	Specific numerical prognostic figures (e.g., “5-year survival 65 %”, “local recurrence rate 18 %”).	Always counted as hallucination regardless of clinical plausibility (the model cannot verify such figures from the input alone).
H2 — Named studies / guidelines	Named studies, trials, or guidelines not present in the input (e.g., “per ESMO 2021 guidelines”, “based on STRASS trial”).	Always counted as hallucination.
H3 — Anatomical / imaging measurements	Specific anatomical or imaging measurements not reported in the input.	Always counted as hallucination.
H4 — General clinical-knowledge claims	General clinical claims (e.g., “myxofibrosarcoma is generally radiosensitive”).	Tracked descriptively; counted as hallucination only if factually incorrect.
H5 — Patient-specific reasoning extensions	Patient-specific reasoning beyond the input (e.g., assumed comorbidity status not in the case).	Tracked descriptively; counted as hallucination only if factually incorrect.

Table 8. Pilot dataset characteristics (n = 51).

Characteristic	Value	Note
Patient demographics		
Median age, years	73 (range 28–93)	Extracted from 48/51 cases; 3 cases described age qualitatively.
Sex: female / male	23 (45.1 %) / 25 (49.0 %)	Extracted from 48/51 cases.
Age ≥ 70 years	26 (51.0 %)	
Histological subtype (top 5)		
Atypical lipomatous tumor / WD liposarcoma	11 (21.6 %)	
Synovial sarcoma	6 (11.8 %)	
Leiomyosarcoma	6 (11.8 %)	
Dedifferentiated liposarcoma	4 (7.8 %)	
Myxofibrosarcoma	4 (7.8 %)	
Other histologies (≤ 3 cases each)	20 (39.2 %)	12 distinct subtypes.
Total distinct histologies	17	
Anatomical location		
Thigh / knee	20 (39.2 %)	Lower extremity
Shoulder / axilla	7 (13.7 %)	Upper extremity
Chest wall / thoracic	4 (7.8 %)	Trunk
Forearm / elbow	4 (7.8 %)	Upper extremity
Lower leg	4 (7.8 %)	Lower extremity
Foot / ankle	3 (5.9 %)	Lower extremity
Other locations	9 (17.6 %)	5 distinct sites
Tumor grade		
G3 / high-grade	27 (52.9 %)	
G2 / intermediate	5 (9.8 %)	
G1 / low-grade	14 (27.5 %)	
Not specified	5 (9.8 %)	

Demographic and clinicopathological characteristics extracted from de-identified case texts via automated regex-based parsing with manual verification.

Table 9. Reproducibility results.

9a. Overall reproducibility.

Metric	Value
Cases analyzed	51
Total LLM runs	154
Decision codes per case	21
Total decision-code instances	1,071
Stable code-instances	1,006
Overall reproducibility	**93.9 %**
Cases with 100 % reproducibility	17 / 51 (33.3 %)
Cases with ≥ 95 % reproducibility	43 / 51 (84.3 %)
Cases with ≥ 80 % reproducibility	50 / 51 (98.0 %)

9b. Per-domain reproducibility.

Clinical domain	Stable / total	Reproducibility	Interpretation
Imaging recommendations	250 / 255	98.0 %	Highest stability; Boolean modality flags well defined.
Systemic therapy	198 / 204	97.1 %	High stability; clear treatment-intent decisions.
Surgery	95 / 102	93.1 %	High stability; minor variation in watch_and_wait borderline cases.
Reconstruction	190 / 204	93.1 %	Driven mainly by clinically equivalent flap-choice variation.
Staging recommendations	137 / 153	89.5 %	Lower stability driven by free-text OTHER field.
Radiotherapy	136 / 153	88.9 %	Variation between accepted dose schemes (50/25, 60/30, 66/33).

9c. Per-decision-code stability across 51 cases.

Decision code	Stable / total	Stability	Domain
imaging_recommended.US	51 / 51	100.0 %	Imaging
imaging_recommended.PETCT	51 / 51	100.0 %	Imaging
staging_recommended.CT_CHEST	51 / 51	100.0 %	Staging
functional_recon.NERVE_TRANSFER	51 / 51	100.0 %	Reconstruction
functional_recon.TENDON_RECON	51 / 51	100.0 %	Reconstruction
chemo_2l	51 / 51	100.0 %	Systemic
io_target	51 / 51	100.0 %	Systemic
imaging_recommended.MRI	50 / 51	98.0 %	Imaging
imaging_recommended.XR	50 / 51	98.0 %	Imaging
functional_recon.VASCULAR_RECON	50 / 51	98.0 %	Reconstruction
margins_goal	49 / 51	96.1 %	Surgery
imaging_recommended.CT	48 / 51	94.1 %	Imaging
rt_decision	48 / 51	94.1 %	Radiotherapy
systemic_decision	48 / 51	94.1 %	Systemic
chemo_1l	48 / 51	94.1 %	Systemic
staging_recommended.CT_ABD_PEL	47 / 51	92.2 %	Staging
surgery_decision	46 / 51	90.2 %	Surgery
rt_dose.GY_TOTAL	44 / 51	86.3 %	Radiotherapy
rt_dose.FRACTIONS	44 / 51	86.3 %	Radiotherapy
staging_recommended.OTHER	39 / 51	76.5 %	Staging
recon_decision	38 / 51	74.5 %	Reconstruction

Codes sorted by stability in descending order. Six codes (29 %) reached 100 % stability across all 51 cases. The lowest-stability code (recon_decision) reflects clinical equipoise between adjacent reconstruction options (e.g., local_flap vs. pedicled_flap) rather than stochastic model instability.

Figures

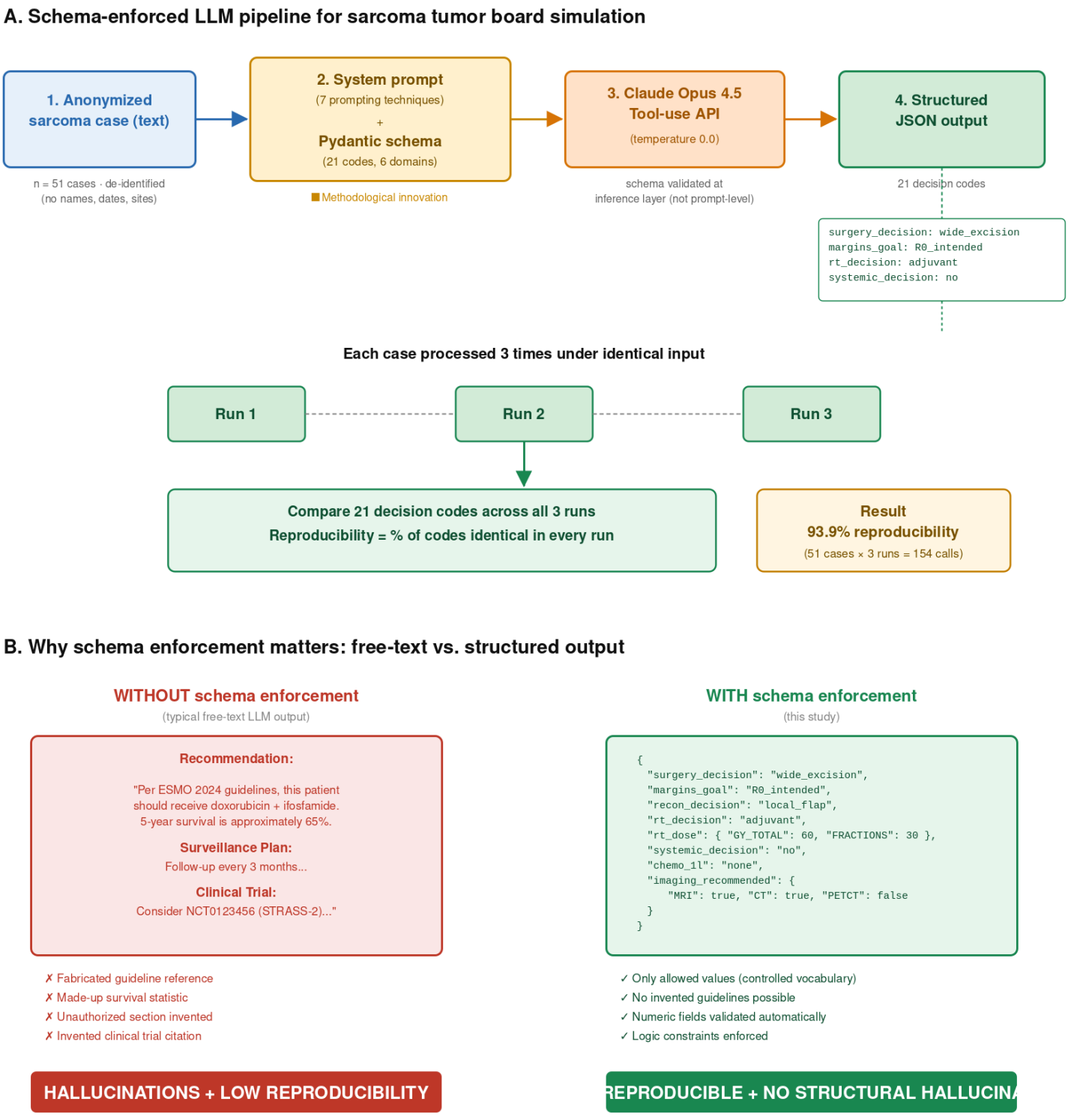


Figure 1

Overview of the schema-enforced LLM pipeline for sarcoma tumor board simulation. **(A)**Pipeline: each anonymized sarcoma case (n = 51 de-identified cases) is processed by the system prompt (seven prompting techniques) together with the Pydantic schema (21 decision codes across nine clinical domains, validated at the inference layer of Anthropic Claude Opus 4.5 via the tool-use API at temperature 0.0). Each case is processed three times under identical input; reproducibility is computed as the percentage of the 21 decision codes that are identical across all three runs (overall 93.9 % across 51 cases × 3 runs = 154 calls). **(B)** Effect of schema enforcement: free-text LLM output (left, red)

frequently produces fabricated guideline references, made-up survival statistics, unauthorized sections, and invented trial citations, with low reproducibility. Schema-enforced output (right, green) is restricted to controlled vocabularies, numeric fields are validated automatically, and conditional-logic constraints prevent inconsistent combinations, yielding 93.9 % reproducibility and no structural hallucinations.

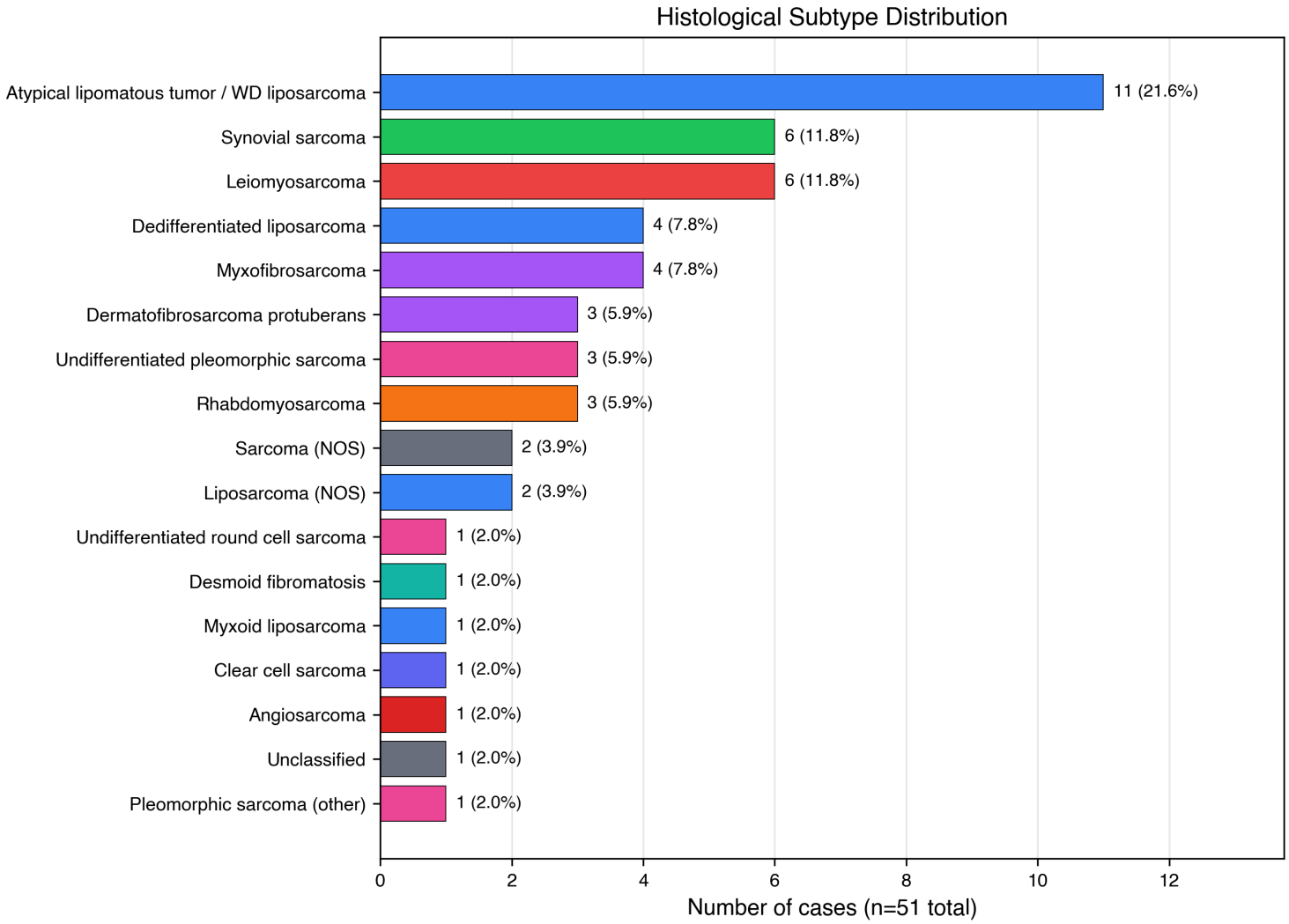


Figure 2

Histological subtype distribution across the 51-case pilot dataset, color-coded by sarcoma family.

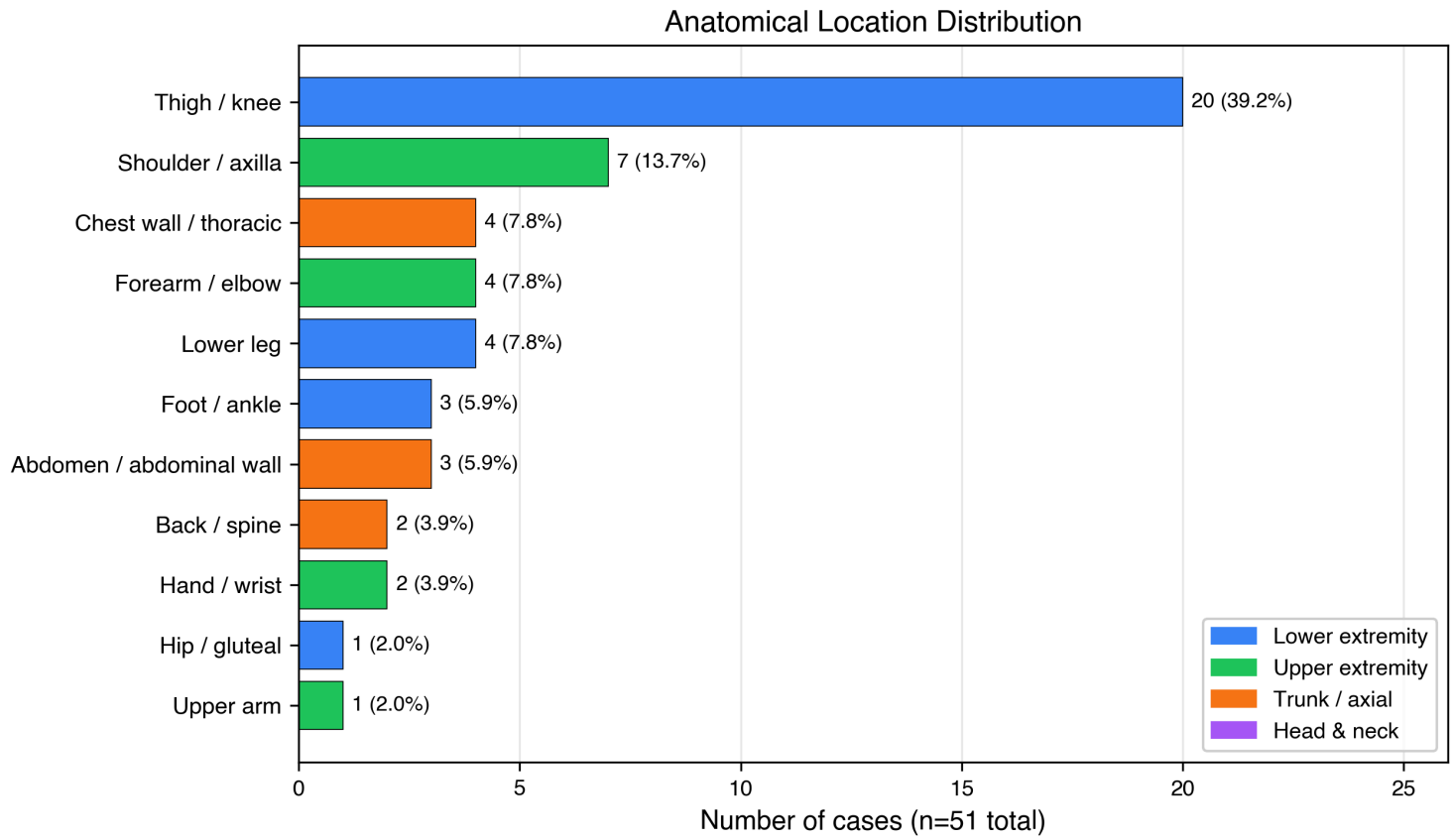


Figure 3

Anatomical location distribution, color-coded by body region (blue: lower extremity; green: upper extremity; orange: trunk / axial; purple: head and neck).

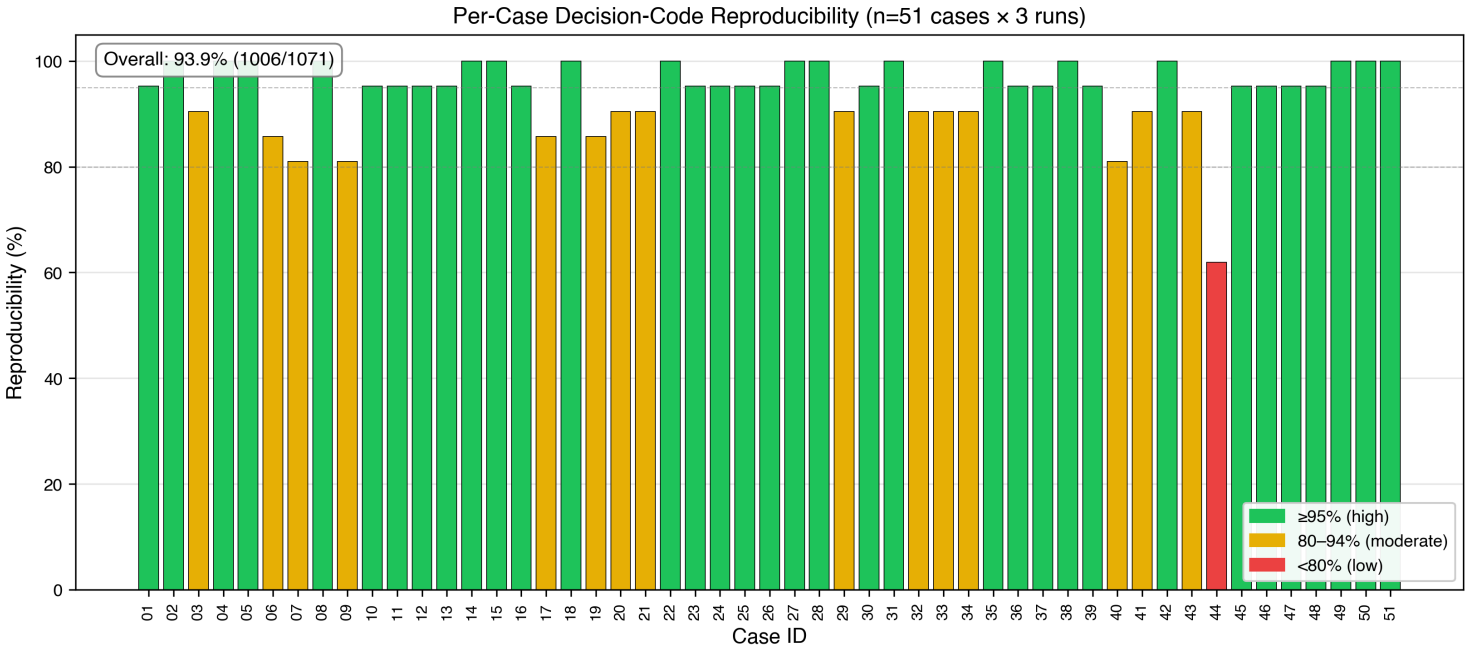


Figure 4

Per-case decision-code reproducibility (51 cases, three runs each). Color coding: green ≥ 95 %; yellow 80–94 %; red < 80 %. Yellow and red cases concentrate in elderly patients with multiple ambiguous decision points. Overall reproducibility 93.9 % (1,006 / 1,071).

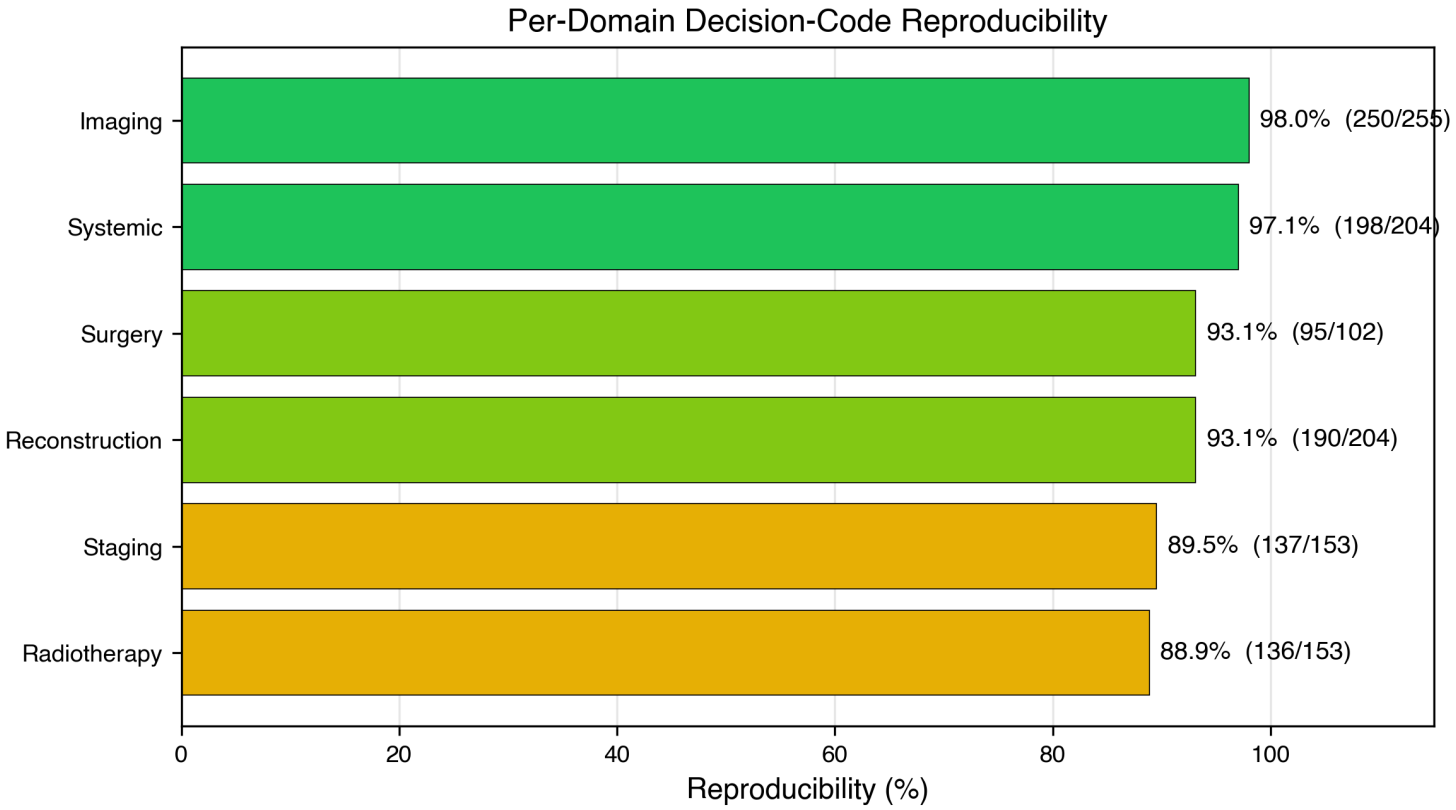


Figure 5

Per-domain decision-code reproducibility, sorted by stability. Imaging and systemic-therapy domains showed near-perfect stability; radiotherapy and staging domains showed lower stability driven primarily by clinically equipotent variations.

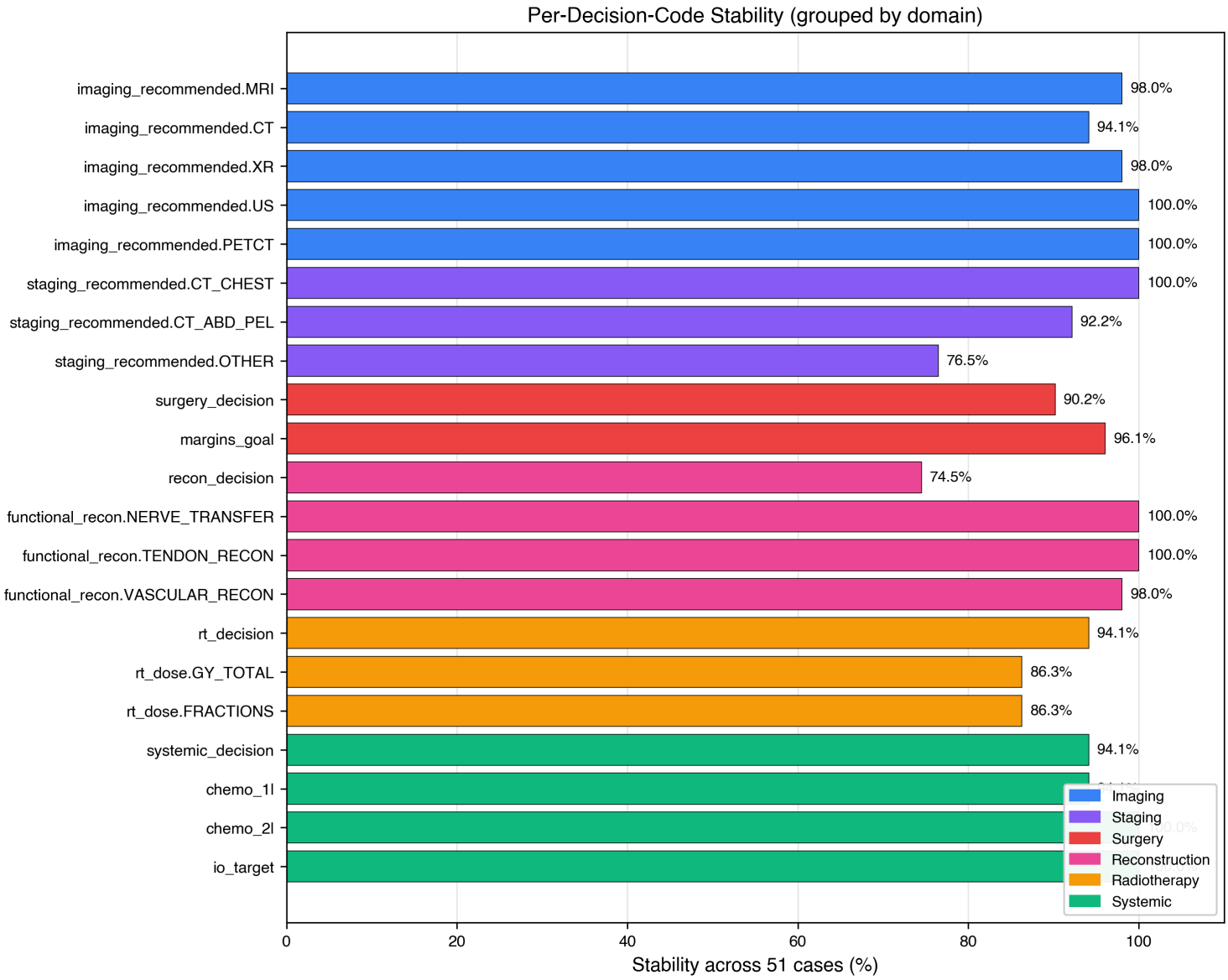


Figure 6

Per-decision-code stability across 51 cases. Six of 21 codes achieved 100 % stability; reconstruction decision was the lowest-stability code at 74.5 %.

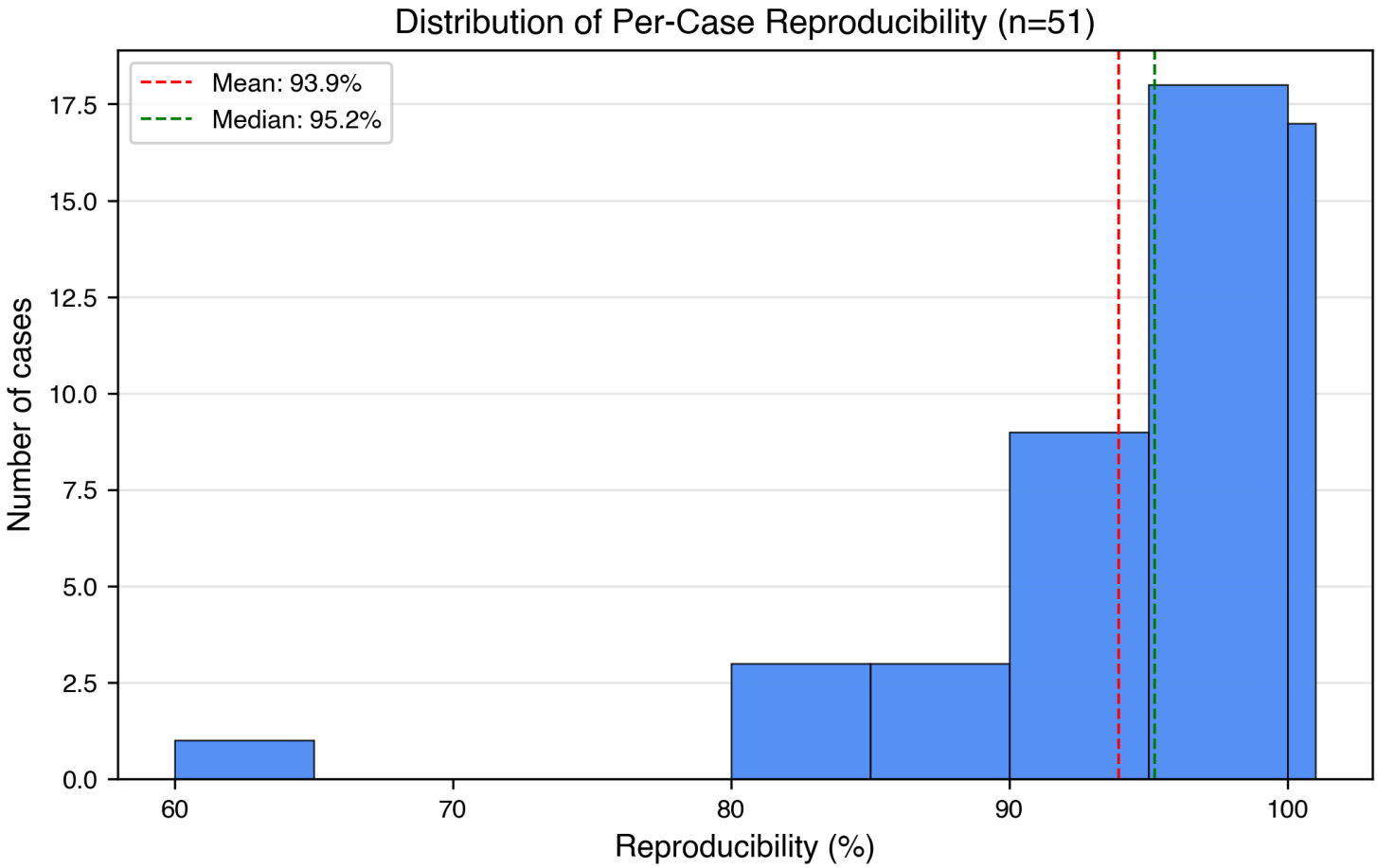


Figure 7

Distribution of per-case reproducibility values. Median 95.2 %, mean 93.8 %. One outlier at 61.9 % (case 44, complex elderly patient with multiple ambiguous decision points).

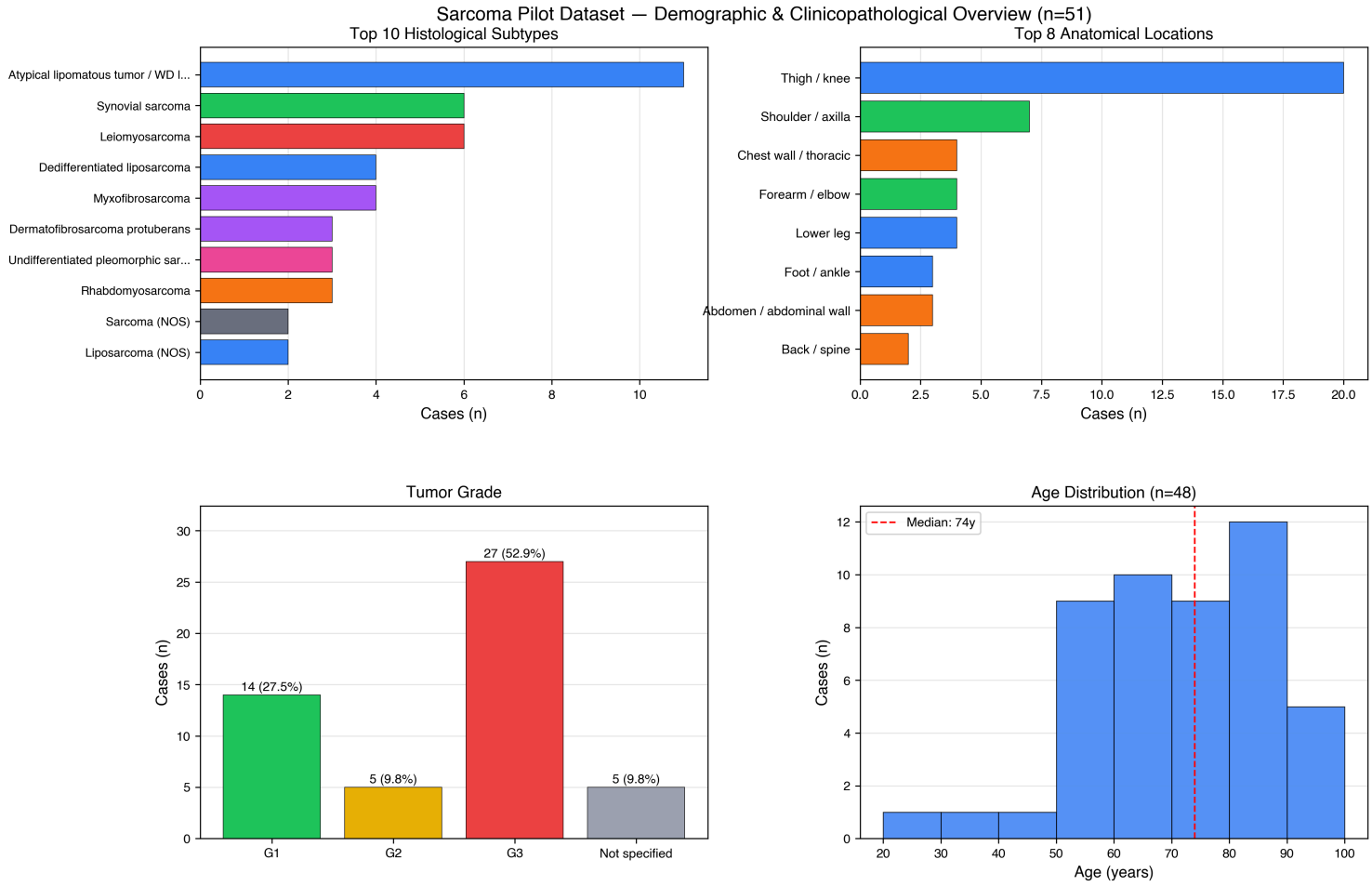


Figure 8

(graphical-abstract candidate). Combined dataset overview dashboard showing top histologies, anatomical locations, tumor grade, and age distribution at a glance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx