An Artifact-based Agent Framework for Adaptive and Reproducible Medical Image Processing

Lianrui Zuo[1][0000−0002−5923−9097], Yihao Liu[1][0000−0003−3187−9903], Gaurav Rudravaram[1][0009−0001−9087−462X], Karthik Ramadass[1,2][0000−0002−4610−3860], Aravind R. Krishnan[1][0009−0000−1829−8245], Michael D. Phillips[1], Yelena G. Bodien[6], Mayur B. Patel[6], Paula Trujillo[7], Yency Forero Martinez[8], Stephen A. Deppen[13,9], Eric L. Grogan[13,9], Fabien Maldonado[9], Kevin McGann[9], Hudson M. Holmes[9,10], Laurie E. Cutting[4,5], Yuankai Huo[2][0000−0002−2096−8065], Bennett A. Landman[1,2,3,11,12][0000−0001−5733−2127]

[1]Department of Electrical and Computer Engineering,
[2]Department of Computer Science, [3]Department of Biomedical Engineering,
[4]Peabody College, [5]Vanderbilt Brain Institute,
Vanderbilt University, Nashville, TN, United States

[6]Department of Surgery, [7]Department of Neurology, [8]Department of Medicine,
[9]Department of Thoracic Surgery, [10]Department of Biomedical Informatics,
[11]Department of Radiology and Radiological Sciences,
Vanderbilt University Medical Center, Nashville, TN, United States

[12]Vanderbilt University Institute of Imaging Science, Nashville, TN, United States
[13]Veteran Affairs Tennessee Valley Healthcare System, Nashville, TN, United States

lianrui.zuo@vanderbilt.edu

Abstract. Medical imaging research is increasingly shifting from controlled benchmark evaluation toward real-world clinical deployment. In such settings, applying analytical methods extends beyond model design and requires dataset-aware workflow configuration and provenance tracking. Two requirements therefore become central: **adaptability**, the ability to configure workflows according to dataset-specific conditions and evolving analytical goals; and **reproducibility**, the guarantee that all transformations and decisions are explicitly recorded and re-executable. Here, we present an artifact-based agent framework that introduces a semantic layer to augment medical image processing. The framework formalizes intermediate and final outputs through an artifact contract, enabling structured interrogation of workflow state and goal-conditioned assembly of configurations from a modular rule library. Execution is delegated to a workflow executor to preserve deterministic computational graph construction and provenance tracking, while the agent operates locally to comply with most privacy constraints. We evaluate the framework on real-world clinical CT and MRI cohorts, demonstrating adaptive configuration synthesis, deterministic reproducibility across repeated executions, and artifact-grounded semantic querying. These results show that adaptive workflow configuration can be achieved without compromising reproducibility in heterogeneous clinical environments. Our code is available at `https://github.com/MASILab/medimg-agent`

Keywords: CT · MRI · Clinical translation · Agent

1 Introduction

Methodological progress in medical imaging has been driven by advances in model architectures and learning paradigms, typically evaluated on curated benchmark datasets or structured research cohorts [3,6,9]. Public initiatives and grand challenges have further accelerated innovation by providing standardized datasets and controlled evaluation protocols [1,8,18,19,20]. These settings enable rigorous comparison of methods under well-defined experimental conditions.

In contrast, when the goal becomes applying existing analytical methods to a new research or clinical dataset, the operational landscape changes substantially. Imaging data originate in hospital and research picture archiving and communication systems (PACS), where heterogeneous file organization, mixed modalities, nested archives, and acquisition planning images are common (Figure 1). For example, a single subject folder very often contains raw DICOM slices mixed with localizer images, reformatted series, and non-imaging documents, without standardized naming or directory structure. Transforming such data into analyzable form requires multi-stage curation, format conversion, quality control, and preprocessing. These steps often demand considerable manual and cohort-specific effort. As a result, workflow assembly becomes an expertise-driven process that is difficult to generalize and reproduce.

These challenges are further compounded in clinical environments governed by privacy regulations and data use agreements. In many settings, raw images or metadata cannot be transmitted to external settings. Ironically, the scenarios that require the greatest interactive effort for dataset adaptation are precisely those in which cloud-based models may be constrained or require additional regulatory and institutional approvals. Intelligent support systems must therefore be designed to operate within such constraints while preserving deterministic execution and traceable provenance.

To this end, two key requirements emerge. First is adaptability: the ability to reason about heterogeneous datasets and evolving analytical goals to configure appropriate processing workflows. Second is reproducibility: the guarantee that all transformations and decisions are explicitly represented and deterministically re-executable [11,14]. Existing workflow engines [4,2] ensure deterministic execution once a workflow is defined, and standardization initiatives [12] improve structural consistency. However, these tools assume pre-specified configurations and do not provide mechanisms for dataset-aware adaptation or semantic inspection of intermediate workflow state.

To bridge this gap, we propose an artifact contract-based agent framework that introduces a semantic layer above deterministic workflow execution. The

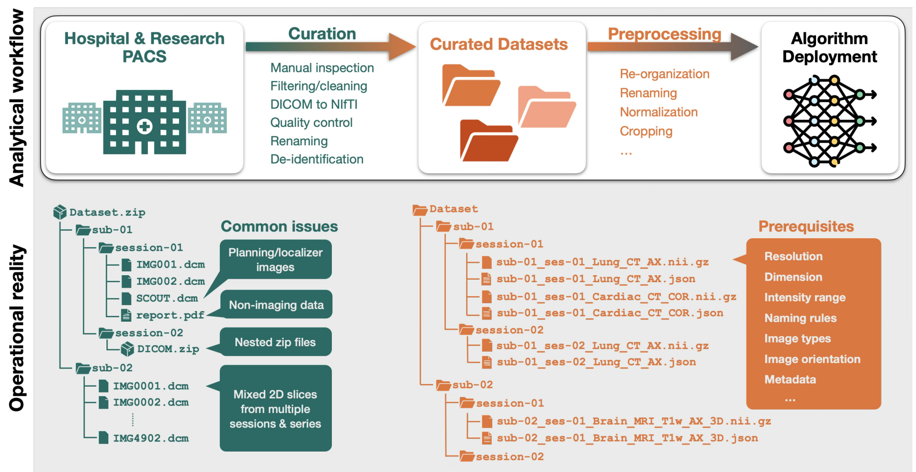

Fig. 1. The typical analytical workflow and operational reality in imaging research. Imaging data originate in hospital and research PACS, where heterogeneous file structures, mixed modalities, and nested archives are common. Extensive curation and processing are often required before most analytical algorithms can be deployed.

framework defines an artifact contract that encodes all workflow outputs and states (intermediate and final) into structured, queryable records with explicit provenance. A semantic query module operates strictly over these contract-compliant artifacts to support dataset-level inspection. For analytical objectives, a workflow assembly module synthesizes goal-conditioned configurations from a modular rule library, enabling dataset-aware adaptation while preserving deterministic execution through an underlying workflow executor. The entire system is deployed fully locally to comply with regulations in many clinical environments. The contributions are threefold:

- An artifact-based contract that formalizes workflow state for structured and auditable reasoning;
- A constrained agent layer enabling goal-conditioned workflow assembly and artifact-grounded semantic querying;
- Empirical validation across clinical CT and MRI cohorts demonstrating adaptive configuration synthesis with deterministic reproducibility.

2 Methods

2.1 Problem Setup

Dataset, Workflow, and Artifact Contract Let $\mathcal{D} = \{d_i\}_{i=1}^{N}$ denote a dataset consisting of N subjects, where each d_i may include multiple longitudinal imaging sessions, file types, and associated metadata. Let $\mathcal{S} = \{s_k\}_{k=1}^{K}$

denote a library of modular processing rules that a user can choose. Each rule s_k specifies a processing step and its dependencies, such as file format conversion or a segmentation algorithm. A workflow configuration is defined as $C = (\pi, \theta)$, where $\pi \subseteq \mathcal{S}$ represents an ordered subset of selected rules together with their dependency graph, and θ denotes rule-level parameters. Given dataset $\mathcal{D}$ and configuration C, execution is performed by a deterministic workflow executor $P(\mathcal{D}, C) = \mathcal{A}$, where $\mathcal{A} = \{a_j\}_{j=1}^{M}$ denotes the set of artifacts produced during execution. Artifacts comprise intermediate and final workflow outputs such as transformed images, derived measurements (e.g., classification labels), quality control indicators, and log records. Conventional workflow engines guarantee reproducibility under fixed configuration C, but assume that configurations are manually specified and that interpretation of workflow state occurs externally. They do not provide structured mechanisms for adaptive configuration to new datasets or semantic interrogation of intermediate state.

Agentic Systems Recent advances in large language models (LLMs) have enabled agentic systems that perform goal-directed reasoning through iterative planning and tool invocation [16,13,10]. In a general agent architecture, a planner selects actions from a tool set conditioned on current state and user objectives, executes those actions, and updates memory based on resulting observations [13]. In the context of medical image processing, however, unrestricted tool invocation is insufficient: execution must remain deterministic and reproducible.

Within our formulation, the dataset $\mathcal{D}$ and artifact set $\mathcal{A}$ define the observable state; the rule library $\mathcal{S}$ defines the available tools [13]; and the configuration $C = (\pi, \theta)$ represents a structured plan over $\mathcal{S}$. We therefore seek a constrained agentic system where planning operates over defined workflow components, memory is grounded in contract-compliant artifacts $\mathcal{A}$ [10], and execution is delegated to the workflow executor $P(\mathcal{D}, C)$. This enables adaptability through configuration synthesis while preserving reproducibility through deterministic execution.

2.2 Proposed Framework

Overview We introduce a constrained artifact-based agent layer (Figure 2) that operates above deterministic workflow execution. Each artifact $a_j \in \mathcal{A}$ is defined under an artifact contract as a tuple $a_j = (t_j, \phi_j, \psi_j)$, where t_j denotes artifact type, ϕ_j denotes structured attributes (e.g., modality, resolution, measurements), and ψ_j encodes provenance information including generating rule and dependency relationships. This contract provides a formally defined, machine-readable state representation over which semantic reasoning is grounded. On top of $\mathcal{A}$, the agent layer implements two constrained functions: **Workflow Assembly:** $\mathit{\Gamma}_{\text{assemble}}(g, \mathcal{A}, \mathcal{S}) = C$, which synthesizes a configuration C by selecting and composing rules from $\mathcal{S}$ conditioned on the current artifact state and analytical goal g; **Semantic Querying:** $\mathit{\Gamma}_{\text{query}}(q, \mathcal{A}) = r$, which returns a response r derived strictly from contract-compliant artifacts to address the user's query q. These operators form a constrained agentic loop where planning and querying are performed over structured state, while execution remains delegated to the deterministic workflow executor $P(\mathcal{D}, C)$ (Figure 2).

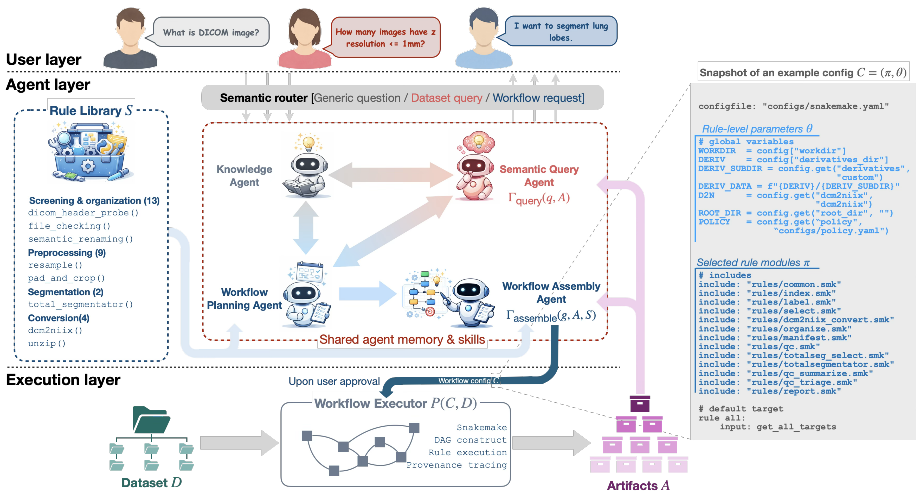

Fig. 2. Overview of the proposed agentic system. User requests are routed to either (1) Knowledge Agent (generic questions), (2) a Semantic Query Agent operating on artifacts $\mathcal{A}$, or (3) a workflow pathway for analytical goals. The Workflow Planning and Assembly Agents construct a configuration C from rule library $\mathcal{S}$ and artifacts $\mathcal{A}$, which is executed by a workflow manager to transform dataset $\mathcal{D}$ into artifacts $\mathcal{A}$.

Adaptation via Workflow Assembly To deploy a workflow on a new dataset $\mathcal{D}_{\text{new}}$, the system first performs an inspection skill to identify available modalities, file organization, and prior processing status. This inspection summarizes observable dataset properties and becomes part of the artifact state.

Given an initial analytical request g_0, the planning component evaluates whether the objective is well-defined with respect to the current artifact state and available rule library $\mathcal{S}$. Because users may not fully know dataset characteristics or available tools, the agent may iteratively refine the request into a structured and executable goal g by clarifying required inputs, outputs, and processing assumptions. Once a feasible goal g is established, $\mathit{\Gamma}_{\text{assemble}}$ selects rules from $\mathcal{S}$ whose declared inputs are satisfied by existing artifacts and whose outputs fulfill the objective. Dependency constraints are enforced during selection, and rule-level parameters are assigned to produce an explicit configuration $C = (\pi, \theta)$, as shown in Figure 2. The assembled workflow and directed acyclic graph (DAG) are then presented to the user for approval before execution. Adaptation therefore occurs through goal planning and configuration synthesis, while computation remains deterministic.

Semantic Query The semantic query function enables inspection of workflow state without re-accessing raw files. Natural-language queries are interpreted as constraints over artifact attributes ϕ_j and provenance ψ_j. For example, a query about available T1-weighted scans with specific resolution is translated into filters over modality and voxel spacing fields stored in artifacts. All responses are generated exclusively from contract-compliant artifacts and reference explicit

ϕ_j and ψ_j. Because the language model operates only over structured artifact fields, responses are grounded in recorded workflow state rather than inferred from unstructured data.

Reproducibility through Workflow Executor Once a configuration C is generated, execution is delegated to a workflow executor (`Snakemake` [4] in our implementation). The executor constructs a DAG from declared rule inputs and outputs, ensuring that dependencies are resolved prior to execution. Given fixed dataset $\mathcal{D}$, configuration C, and computational environment, execution is deterministic. Intermediate and final artifacts are produced only when declared inputs are satisfied, and rule-level provenance is recorded automatically.

Implementation Details The agent operates locally via `Ollama` using the DeepSeek-R1 (14B) model in inference mode. Workflow rules are organized as modular `.smk` components and registered in a rule catalog that explicitly declares inputs, outputs, and parameter interfaces. Artifacts are stored as structured `.json` or `.csv` records consistent with the artifact contract. Workflow execution is performed using Snakemake (v7.x), which constructs the DAG from the generated configuration and executes tasks with file-based dependency tracking. Experiments were conducted on a Linux workstation equipped with an NVIDIA A6000 GPU for downstream model inference.

3 Experiments and Results

We evaluate whether the proposed framework (i) preserves deterministic execution under repeated runs, (ii) assembles goal-conditioned workflows across varying cohorts, and (iii) supports artifact-grounded semantic query without rerunning. Experiments were conducted on three cohorts including both research and clinical settings (Table 1): `NLST` (research CT; moderate heterogeneity), `LungCaTrial` (clinical trial CT; high heterogeneity), and `BrainICU` (research MRI; moderate heterogeneity). Across cohorts, we test our method under multiple analytical goals including data conversion/curation and downstream analysis (i.e., lung lobe and nodule segmentation [15], cancer risk estimation [7], harmonization [21], and brain segmentation [17]), resulting in 9 different workflows.

Reproducibility For each dataset and analytical goal, we invoke the agent twice after clearing memory and intermediate states, then compare the resulting workflow graphs. Reproducibility is measured by DAG equivalence between the two runs. Across all settings, the induced DAGs were identical (Table 1), indicating that, under fixed goals and environment, the planning-to-configuration pathway yields consistent executable workflows and deterministic execution graphs.

Adaptability We measure how well the agent adapts workflows to each cohort and analytical goal using three metrics. First, Initial Rule Matching (IRM) quantifies overlap between the agent's initial proposed rule set and a domain-expert configuration, reflecting the quality of dataset-aware planning immediately after inspection. Second, Planning Iterations (PL) counts the number of user-agent exchanges before finalizing a configuration, capturing interaction cost.

Table 1. Dataset characteristics and quantitative evaluation of reproducibility and adaptability. For each dataset and analytical goal, we report the number of rules in the domain-expert ground-truth configuration, reproducibility via DAG equivalence across repeated executions, and adaptability metrics†.

Dataset Characterization						Assembly		Reprod.	Adaptability		
Dataset	Modality	Cohort	Heteroge.	#subj	#ses	Analytical goal	#Rules	DAG	IRM(%)↑	PL↓	FO(%)↑
`NLST` [5]	CT	Research	Mid.	200	592	convert+curate	11	✓	100	1	98.8
						lung lobe seg	14	✓	85.7	4	97.9
						lung nodule seg	14	✓	100	3	97.4
`LungCaTrial`	CT	Clinical	High	636	2,950	convert+curate	12	✓	100	2	99.0
						lung nodule seg	15	✓	100	4	90.7
						lung cancer risk	17	✓	88.2	6	91.1
`BrainICU`	MRI	Research	Mid.	113	135	convert+curate	17	✓	94.1	5	100
						harmonization	19	✓	100	8	97.7
						brain seg	20	✓	100	8	96.3

Example: Goal Planning (`NLST`, lung nodule seg)

```
User > I want to segment the lung nodules in this dataset.
Agent> Current status: DICOMs have been converted to NIFTI through dcm2niix (registered rule
0.1.3a). Needs confirmation: 1. Kernel. I see multiple kernels of the same session. Do you want
to segment them all? 2. Resolution. Do you want to choose the best resolution? ...
User > 1. Segment them all. 2. Choose the best resolution (in z-dimension) ···
Agent> This is an updated workflow with rules ... Assumptions made: [omit to save space]. If
you approve, I will assemble and execute.
```

†**Abbreviations: IRM:** Initial Rule Matching. Percentage of rules matched between agent's first selection and ground truth; **PL:** Number of planning iterations (chats from user) before a workflow is assembled by the agent; **FO:** Percentage of sessions whose final outputs satisfy the intended goal against manually curated results.

Third, Final Output (FO) reports the percentage of sessions whose outputs satisfy the intended goal against manually curated results.

As shown in Table 1, the agent's initial plans are close to expert-defined pipelines, with IRM> 85% across all cohorts and analytical goals. Additionally, iterative planning resolves ambiguities in user intent (e.g., which series to target), as shown in Table 1 example. This is reflected by modest PL values (a workflow is accurately assembled after 1 to 8 rounds of chats between agent and user) even in the clinical cohorts, while maintaining high session-level success in most tasks. We then evaluated all failure cases: mismatches are concentrated in cases of missing/corrupted data and ambiguous series selection (e.g., multiple candidate series satisfying similar constraints), which typically require manual intervention even in conventional workflows.

Semantic Query We evaluate semantic querying across three categories: status, filtering/counting, and provenance. For each category, 20 representative questions are evaluated using exact-match accuracy against ground truth. We used two local LLMs in this experiment (Table 2). Status queries are generally answered correctly, as they simply require checking whether certain files exist (e.g., "Has lung nodule segmentation been executed for subject X?"). Filtering/counting queries are more demanding because they require structured attribute reasoning, such as checking specific voxel resolution or scanner manufacturer. Provenance queries require tracing rule-level dependencies, for example

Table 2. Semantic query evaluation (`deepseek-r1:14b` and `qwen2.5:14b`) and ablation experiment with/without artifact contract. Accuracy denotes exact-match percentage over 20 queries per category.

Dataset	Model	Contract?	Status	Filter	Provenance
NLST	DeepSeek	✗	95%	20%	80%
	Qwen	✓	100%	90%	100%
	DeepSeek	✓	100%	95%	100%
LungCaTrial	DeepSeek	✗	90%	15%	80%
	Qwen	✓	100%	85%	100%
	DeepSeek	✓	100%	90%	100%
BrainICU	DeepSeek	✗	95%	10%	75%
	Qwen	✓	95%	90%	100%
	DeepSeek	✓	100%	90%	100%

Example (Filter/counting)

```
User > How many NIfTI images were
acquired using a Siemens scanner and
have slice thickness greater than 1 mm?
Agent> 23. File paths can be retrieved
from artifact records in
data_inventory.csv and qa_report.csv.
```

identifying which preprocessing steps generated a particular result. The proposed method accurately answers all status and provenance queries, and only made one to two mistakes in dataset query questions. Upon investigating error cases, we found that most failures were caused by incomplete or corrupted DICOM headers, leading to missing or inaccurate artifact attributes ϕ_j. For example, body part was inferred from DICOM headers, and this process can be unreliable. These errors therefore reflect limitations of upstream metadata rather than ungrounded generation. Incorporating additional skills, such as image-based body-part regression, may further improve robustness.

Ablation Study To isolate the effect of artifact contract, we perform an ablation in which the agent is denied access to the artifact registry and instead receives only directory listings and raw filenames, mimicking a conventional manual scripting workflow. The same query set is evaluated under this condition.

After removing the artifact contract, the largest degradation occurs in filtering/counting queries. For example, when asked to identify sessions with voxel resolution below a specified threshold, filename-based reasoning fails because resolution is not encoded in directory names; it is only available in structured artifact fields extracted during preprocessing. Provenance queries also degrade without artifact access. When asked which rule generated a particular segmentation output, the grounded system retrieves the recorded rule identifier and dependency chain from ψ_j, whereas the filename-only condition lacks explicit traceability and must rely on heuristic inference. These results demonstrate the importance of contract-compliant artifacts. The artifact contract therefore serves not only as a reproducibility mechanism but also as a grounding interface that reduces inference over unstructured file organization.

4 Discussion and Conclusion

We presented an artifact contract-based agent framework for adaptive and reproducible medical image processing across heterogeneous research and clinical datasets. The system separates semantic planning from deterministic execution: workflow configurations are synthesized at the semantic layer, while execution remains reproducible through a workflow manager with explicit provenance tracking. Experiments show that configurations can be generated across diverse CT and MRI cohorts, and that repeated runs produce identical DAGs. In our ablation study, we show that the artifact contract improves the semantic query accuracy. The main limitation arises from incomplete artifact attributes. In our experiments, most query errors were caused by missing or corrupted metadata, leading to inaccurate artifact fields. This highlights the dependence of semantic query on upstream data quality. Similarly, adaptability is bounded by the coverage of the rule catalog; unsupported analytical objectives require additional rules. Additionally, a more detailed analysis on the computational overhead and time cost during user–agent communication is valuable for the clinical application of the proposed framework. Nevertheless, the framework is designed for extension. New rules and skills can be added without altering the overall architecture, enabling progressive enrichment of the artifact schema and planning capabilities. By grounding interaction in structured workflow state and preserving deterministic execution, the proposed system provides a practical and extensible foundation for adaptive medical image analysis in real-world settings.

Acknowledgments. We sincerely thank Adam M. Saunders and Michael E. Kim for sharing their experience with `Snakemake`, which contributed to the development of this work. This research was funded by the National Cancer Institute (NCI) grant R01-CA253923-04 and R01-CA253923-04S1. This work was also supported by the following awards: National Science Foundation CAREER1452485; U01CA196405; UL1-RR024975-01 from the National Center for Research Resources and UL1-TR000445-06 from the National Center for Advancing Translational Sciences; the Martineau Innovation Fund grant through the Vanderbilt-Ingram Cancer Center Thoracic Working Group; and NCI Early Detection Research Network grant 2U01-CA152662. Additional support was provided by U01-CA152662, SNF300329 (Helmsley), NSF NAIRR Pilot 240468, and P50-HD103537. During the development of this work, we used generative AI to create code segments from task descriptions and to assist with debugging, editing, and code autocompletion. In addition, generative AI technologies were used to help structure sentences and perform grammatical checks. We emphasize that the conceptualization, ideation, and all prompts provided to the AI originated entirely from the authors' intellectual and creative contributions. The authors take full responsibility for reviewing and validating all AI-generated content included in this work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, J., Wei, S., Honkamaa, J., Marttinen, P., Zhang, H., Liu, M., Zhou, Y., Tan, Z., Wang, Z., Wang, Y., et al.: Beyond the LUMIR challenge: The pathway to

foundational registration models. arXiv preprint arXiv:2505.24160 (2025)
2. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C.: Nextflow enables reproducible computational workflows. Nature biotechnology **35**(4), 316–319 (2017)
3. Ferrucci, L., Resnick, S.M., Deal, J.A.: Baltimore Longitudinal Study of Aging (BLSA). Hearing Loss Rehabilitation and Higher-Order Auditory and Cognitive Processing p. 116 (2023)
4. Köster, J., Rahmann, S.: Snakemake—a scalable bioinformatics workflow engine. Bioinformatics **28**(19), 2520–2522 (2012)
5. Kramer, B.S., Berg, C.D., Aberle, D.R., Prorok, P.C.: Lung cancer screening with low-dose helical CT: results from the National Lung Screening Trial (NLST) (2011)
6. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. medrxiv pp. 2019–12 (2019)
7. Li, T.Z., Xu, K., Gao, R., Tang, Y., Lasko, T.A., Maldonado, F., Sandler, K.L., Landman, B.A.: Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. In: Medical Imaging 2023: Image Processing. vol. 12464, pp. 229–238. SPIE (2023)
8. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
9. Naseer, I., Akram, S., Masood, T., Jaffar, A., Khan, M.A., Mosavi, A.: Performance analysis of state-of-the-art CNN architectures for LUNA16. Sensors **22**(12), 4426 (2022)
10. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. pp. 1–22 (2023)
11. Peng, R.D.: Reproducible research in computational science. Science **334**(6060), 1226–1227 (2011)
12. Poldrack, R.A., Markiewicz, C.J., Appelhoff, S., Ashar, Y.K., Auer, T., Baillet, S., Bansal, S., Beltrachini, L., Benar, C.G., Bertazzoli, G., et al.: The past, present, and future of the brain imaging data structure (BIDS). Imaging Neuroscience **2**, imag–2 (2024)
13. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023)
14. Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P., Taufer, M.: Enhancing reproducibility for computational methods. Science **354**(6317), 1240–1241 (2016)
15. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
16. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)

17. Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., Lasko, T.A., et al.: Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. Medical Image Analysis **90**, 102939 (2023)
18. Zhang, J., Zuo, L., Dewey, B.E., Remedios, S.W., Hays, S.P., Pham, D.L., Prince, J.L., Carass, A.: Harmonization-Enriched Domain Adaptation with Light Fine-tuning for Multiple Sclerosis Lesion Segmentation. In: Medical Imaging 2024: Clinical and Biomedical Imaging. vol. 12930, pp. 633–639. SPIE (2024)
19. Zhang, J., Zuo, L., Dewey, B.E., Remedios, S.W., Liu, Y., Hays, S.P., Pham, D.L., Mowry, E.M., Newsome, S.D., Calabresi, P.A., et al.: Uniself: A unified network with instance normalization and self-ensembled lesion fusion for multiple sclerosis lesion segmentation. Medical Image Analysis p. 103954 (2026)
20. Zuo, L., Dewey, B.E., Carass, A., He, Y., Shao, M., Reinhold, J.C., Prince, J.L.: Synthesizing Realistic Brain MR Images with Noise Control. In: International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI). pp. 21–31. Springer (2020)
21. Zuo, L., Liu, Y., Xue, Y., Dewey, B.E., Remedios, S.W., Hays, S.P., Bilgel, M., Mowry, E.M., Newsome, S.D., Calabresi, P.A., et al.: HACA3: A Unified Approach for Multi-site MR Image Harmonization. Computerized Medical Imaging and Graphics **109**, 102285 (2023)