Multi-Agent Reasoning for Cardiovascular Imaging Phenotype Analysis

Weitong Zhang[1*], Mengyun Qiao[2,3,4*], Chengqi Zang[5], Steven Niederer[6], Paul M Matthews[3,8,9], Wenjia Bai[1,3,4], Bernhard Kainz[1,7]

[1] Department of Computing, Imperial College London, London, UK
[2] Department of Mechanical Engineering, University College London, London, UK
[3] Department of Brain Sciences, Imperial College London, London, UK
[4] Data Science Institute, Imperial College London, London, UK
[5] University of Tokyo, Tokyo, JP
[6] National Heart and Lung Institute, Imperial College London, London, UK
[7] FAU Erlangen-Nürnberg, Nürnberg, DE
[8] UK Dementia Research Institute, Imperial College London, London, UK
[9] Rosalind Franklin Institute, Harwell Science and Innovation Campus, Didcot, UK
wz1820,mq21@ic.ac.uk

Abstract. Identifying the associations between imaging phenotypes and disease risk factors and outcomes is essential for understanding disease mechanisms and improving diagnosis and prognosis models. However, traditional approaches rely on human-driven hypothesis testing and selection of association factors, often overlooking complex, non-linear dependencies among imaging phenotypes and other multi-modal data. To address this, we introduce a **M**ulti-agent **E**xploratory **S**ynergy for the **H**eart (MESHAgents) framework that leverages large language models as agents to dynamically elicit, surface, and decide confounders and phenotypes in association studies, using cardiovascular imaging as a proof of concept. Specifically, we orchestrate a multi-disciplinary team of AI agents, which spontaneously generate and converge on insights through iterative, self-organizing reasoning. The framework dynamically synthesizes statistical correlations with multi-expert consensus, providing an automated pipeline for phenome-wide association studies (PheWAS). We demonstrate the system's capabilities through a population-based study of imaging phenotypes of the heart and aorta. MESHAgents autonomously uncovered correlations between imaging phenotypes and a wide range of non-imaging factors, identifying additional confounder variables beyond standard demographic factors. Validation on diagnosis tasks reveals that MESHAgents-discovered phenotypes achieve performance comparable to expert-selected phenotypes, with mean AUC differences as small as $-0.004_{\pm 0.010}$ on disease classification tasks. Notably, the recall score improves for 6 out of 9 disease types. Our framework provides clinically relevant imaging phenotypes with transparent reasoning, offering a scalable alternative to expert-driven methods.

[*] Equal contribution

Keywords: Multi-agent system · Phenome-wide association studies · Cardiovascular imaging · Integration of imaging and non-imaging data.

1 Introduction

Medical imaging is essential for deriving quantitative phenotypes of anatomical structure and function and understanding their associations with various non-imaging factors, such as genetic predispositions [1], environmental conditions [2] and physiological characteristics [3] etc. How to better integrate imaging with non-imaging data to perform multi-modal data analysis has received increasing attention in the imaging community. To ensure reliable and unbiased associations between imaging and non-imaging data, it is crucial to identify associations affecting both risk factors and imaging phenotypes [4]. A confounder is an association variable that influences both the independent variable and the dependent variable, creating a spurious association between them. Ignoring confounders can lead to misinterpretations in phenome-wide association studies (PheWAS) and genome-wide association studies (GWAS), impacting clinical analyses. Confounding factors also need to be considered when evaluating bias in machine learning algorithms [5], particularly in medical imaging research [6, 7]. Conventional methods rely on established empirical criteria or expert selection, not accounting for complicated non-linear relationships between imaging phenotypes and health outcomes. Performing automated reasoning on large-scale, multi-modal datasets and integrating the multi-disciplinary expertise is not well investigated in the past.

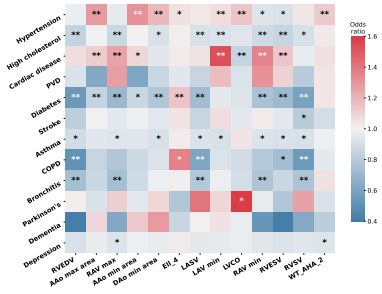
Recent advances in large language models (LLMs) [8, 9] have exhibited notable reasoning abilities across a wide range of tasks and applications [10, 11], with these capabilities stemming from extensive training on vast comprehensive corpora covering diverse topics. In real-world scenarios, LLMs tend to encounter domain-specific tasks (*e.g.*, strategy [12], safety [13]) that necessitate a combination of domain expertise and complex reasoning abilities [14, 15]. This backdrop makes the adoption of LLMs in the medical field a noteworthy research topic, which has recently gained growing attention [16, 17].

Challenges. Two major challenges prevent LLMs from handling tasks in multi-modal medical imaging data analysis: (i) Imaging phenotype analysis demands multi-disciplinary domain expertise [19, 20] spanning medicine, biomechanics, and statistics, yet LLMs suffer from catastrophic forgetting [21, 22] in maintaining comprehensive knowledge across disciplines. (ii) The analysis requires phenotype-aware reasoning to identify robust correlations between imaging phenotypes and a wide range of non-imaging factors [4, 18], whereas LLMs tend to produce *hallucination* [23, 24] that undermines phenotype discovery.

Contributions. (i) We introduce the first multi-agent system for medical imaging PheWAS where domain-specific LLM agents achieve association consensus through structured interactions, mitigating catastrophic forgetting through a multi-disciplinary team. (ii) We demonstrate that role-based collaborative reasoning enhances imaging phenotype discovery without retrieval augmenta-

Expert-Selected[18, 7]	MESHAgents-Discovered
Imaging phenotypes	
RVEDV (mL)	RVEDV (mL)
LVEDV (mL)	RVESV (mL)
LVEF (%)	RVSV (mL)
LVM (g)	LVCO (L/min)
RVEF (%)	LAV min (mL)
LAV max (mL)	LASV (mL)
LAEF (%)	RAV max (mL)
RAV max (mL)	RAV min (mL)
RAEF (%)	AAo max area (mm2)
AAo max area (mm2)	AAo min area (mm2)
AAo distensibility	DAo min area (mm2)
DAo max area (mm2)	WT_AHA_2 (mm)
DAo distensibility	Ell_4 (%)
Confounders	
Sex	Sex
Age	Age
Weight	Weight
Height	Height
	Alcohol intake
	Systolic BP
	Diastolic BP

(a) Parameter comparison



(b) Disease association heatmaps

Fig. 1: **MESHAgents-discovered cardiovascular parameters and their disease associations.** (a) Comparison between expert-selected in literature [18, 7] and MESHAgents-discovered parameters, with overlapping parameters highlighted in green. MESHAgents autonomously identified clinically relevant phenotypes and additional confounders without prior domain knowledge. (b) Heatmaps showing significant associations between discovered imaging phenotypes and disease outcomes on 38,309 participants (**p $<$ 0.01, *p $<$ 0.05).

tion, maintaining interpretable decision trails while reducing hallucination. (iii) Through multi-faceted evaluation, MESHAgents shows superior phenotype discovery over existing single-agent and multi-agent approaches on population imaging data, while achieving diagnostic performance comparable to expert-selected parameters across three classification models (AUC difference $-0.004_{\pm 0.010}$)

Related Work. Recently, medical AI agents have shown promising progress, from single-agent systems like Med-PaLM [25] to collaborative frameworks such as MedAgents [26] and RareAgents [27] for medical question answering and diagnosis. These systems have demonstrated the potential of LLM-based agents in clinical scenarios, including triaging [28] and medical image analysis [29]. Meanwhile, multi-agent collaboration has emerged as a powerful paradigm in LLM applications [30], with approaches like role playing [31] and structured communication [32] enabling more sophisticated reasoning. Recent works have explored adversarial collaboration through debates [33] and negotiation [34] to enhance decision quality. However, these approaches lack consensus mechanisms for orchestrating distributed medical expertise and validating phenotype associations, making them insufficient for faithful medical imaging PheWAS.

2 Multi-agent Exploratory Synergy for the Heart

Problem Formulation. (i) For phenotype association analysis, given a large-scale imaging dataset containing N participants, let $\mathcal{X} = \{x_1, ..., x_N\}$ represent the set of imaging-derived phenotypes where each $x_i \in \mathbb{R}^d$ consists of d structural and functional phenotypes. For each participant, we also observe a set of

We follow the evaluation setting from literature [18].

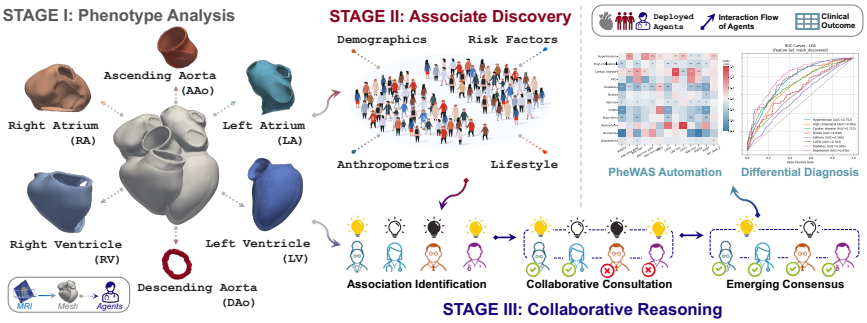


Fig. 2: `MESHAgents` **framework for cardiac imaging phenotype analysis through multi-agent reasoning.** Operating and validating in three stages: (i) specialist agents analyze cardiac structures from imaging measurements; (ii) agents discover confounders via distributed analysis; (iii) collaborative reasoning generates phenome-wide associations, validated through clinical metrics.

non-imaging factors $\mathcal{C} = \{c_1, c_2, ..., c_M\}$ spanning demographics, anthropometrics, lifestyle, and risk factors. Our objective is to discover a set of phenotypes $\mathcal{P} = \{p_1, p_2, ..., p_K\}$ where each phenotype p_k exhibits a strong and meaningful association with certain factors. MESHAgents identifies phenome-wide associations $\mathcal{A} = \{(p_k, c_m, a_{km})\}$ between image phenotype p_k and factor c_m. (ii) For diagnosis, we apply MESHAgents-discovered features to a disease classification model. Each patient profile is represented using MESHAgents-identified parameters, including both imaging phenotypes and non-imaging factors: $\mathcal{X}_i^{MESH} = \{p_k(x_i), c_m\}$ where $p_k \in \mathcal{P}$ and $c_m \in \mathcal{C}$. Standard classifiers then classify these features to a disease type: $f_{classifier}(\mathcal{X}_i^{MESH}) \rightarrow \mathcal{D} = \{d_1, d_2, ..., d_9\}$. This provides a more stringent evaluation than query-based tasks [27], though MESHAgents could support direct diagnostic querying with information retrieval.

MESHAgents Overview. MESHAgents orchestrates an expert team of agents α_i to analyze imaging phenotypes and discover associative factors through three progressive stages: $\mathcal{R}$ is medical imaging data containing structural and functional measurements, and $\mathcal{A}$ represent the identified phenotype-factor relationships contributing to diagnosis.

Dynamic Memory. To enable collective analysis, we design memory-augmented agents that integrates domain expertise with historical experience across local analysis and group reasoning. Each specialist agent α_i maintains a structured, long-term memory bank $\mathcal{M}_i$ that encodes imaging phenotype patterns and associative factors from past analyses. For a given phenotype set $\mathcal{X}$, the framework employs a similarity-based mechanism: $\mathcal{R}_i = \arg\max_{h \in \mathcal{H}_i} \text{Sim}\left(Emb(\mathcal{X}), Emb(h)\right)$, where $\mathcal{H}_i$ represents agent α_i's historical cases and $\mathcal{R}_i$ is the retrieved relevant experience. $Emb(\cdot)$ maps imaging phenotypes to an embedding space. This memory mechanism enables agents to leverage past experiences and statistical evidence when evaluating phenotypes $\mathcal{P}$ and factors $\mathcal{C}$.

Consensus Mechanism. To foster independent reasoning while preventing premature convergence, MESHAgents employ a sequential discussion protocol. In each round t, expert agent α_i formulates its opinion independently based on three essential components: $O_i^t = g_i(\mathcal{V}_i, \mathcal{H}_{t-1}, \mathcal{M}_i)$, where $g_i(\cdot)$ is the opinion generation of agent α_i that synthesizes domain-specific insights, discussion history, and group context. Domain analysis $\mathcal{V}_i$ ensures specialty-specific insights, discussion history $\mathcal{H}_{t-1}$ enables progressive refinement, and memory bank $\mathcal{M}_i$ provides group context. This sequential consensus building offers advantages over simultaneous discussion: (i) it reduces mutual influence during initial assessment, allowing each expert to form an unbiased domain-specific opinion [35], (ii) it enables integration and update of previous insights through $\mathcal{H}_{t-1}$ while avoiding false consensus [13], and (iii) it provides transparent decision trails instead of black-box models. The coordinator agent guides this process until reaching agreement (typically within 10 rounds), ensuring both independent expertise contribution and coherent consensus formation. Critically, this sequential mechanism creates traceable emergence patterns and transparent validation, whether phenotype analysis, associative factor discovery, or collective reasoning process.

Tools. Each agent use a suite of tools $\mathcal{T} = \{T_1, ..., T_J\}$ that process imaging phenotypes $\mathcal{R}$. Each tool T_j implements specific analytical functions: $\mathcal{TR} = \bigcup_{j=1}^{J} T_j(\mathcal{R})$, where this represents the collective evidence generated by applying agent-selected tools to the data. These tools perform statistical significance tests, effect size calculation, and population distribution analysis. The final decision $\mathcal{A}$ emerges through synthesis of agent opinions $\mathcal{O}^{(t)}$ at discussion round t, memory retrievals $\mathcal{R}_i$, and tool-derived evidence $\mathcal{TR}$: $\mathcal{A} = f_{AP}(\{\mathcal{O}^{(t)}\}_{t=1}^{T}, \{\mathcal{R}_i\}_{i=1}^{N}, \mathcal{TR})$, where f_{AP} represents an aggregation function that weights different information sources based on their statistical confidence (< 0.05) and on-topic relevance (> 0.3). This integrated framework provides the foundation for the three-stage analysis pipeline, enabling phenotype valuation $\mathcal{V}_i$, factor effect size assessment $\mathcal{E}_i$, and collaborative reasoning across three stages:

`Stage I: Phenotype Analysis.` In the initial stage, each agent α_i analyses the imaging phenotypes within a specific domain: {left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), ascending aorta (AAo) and descending aorta (DAo)}. Here, we define the specialty by different anatomical structures, similar to specialty training in medical education. Given a phenotype set $\mathcal{P} = p_1, p_2, ..., p_K$, each agent evaluates: $\mathcal{V}_i = \{\phi_i(p_k, \mathcal{R}_i) | p_k \in \mathcal{P}\}$ where $\phi_i(\cdot)$ represents agent α_i's phenotype valuation function incorporating statistical significance assessment, clinical relevance evaluation, population distribution analysis, and stability among the data. Each agent is able to use tools to extract features from statistical pools within the scope of prior knowledge for passing to the discussion. The collective phenotype valuations $\mathcal{V}_i$ form the fidelity of representation before the discussion.

`Stage II: Associative Factor Discovery.` Then, agents examine a wide range of non-imaging factors $\mathcal{C} = c_1, c_2, ..., c_M$ through a similar two-level analysis: For local analysis, each agent α_i evaluates factor effects within its domain: $\mathcal{E}_i = \{(p_n, c_m, \psi_i(p_n, c_m, \mathcal{V}_i)) | p_n \in \mathcal{P}, c_m \in \mathcal{C}\}$ where $\psi_i(\cdot, \cdot)$ quantifies

the phenotype-factor association strength. For global analysis, the coordinator agent aggregates local assessments $\{\mathcal{E}_i\}_{i=1}^{N}$ to establish cross-domain factor effects $\mathcal{E}_G = h(\{\mathcal{E}_i\}_{i=1}^{N})$, where $h(\cdot)$ is a consensus aggregation function. The agents are set to reveal highly significant associations with a wide range of non-imaging factors, such as early-life factors, lifestyle and cognitive functions etc.

`Stage III: Collaborative Reasoning.` The final stage facilitates collaborative reasoning among agents to reach final consensus and report generation. This is inspired by multidisciplinary team (MDT) panel meetings in clinical decision making. We design a dynamic long-term memory mechanism for the agents in MESHAgents, enabling them to store, retrieve, and update memories like human experts. Agents can facilitate personalized analysis and diagnosis based on historical interactions. Meanwhile, agents interact through structured dialogues following protocols of evidence-based argumentation. The consensus formation follows an iterative process until convergence across all decision-making nodes (*i.e.,* recommended phenotype, top associative factor, top diagnosis).

3 Experiments and Results

Dataset and Setting. The UK Biobank data is used, which are available from UK Biobank via a standard application procedure. Cardiac imaging phenotypes are derived from cardiac MR images using a deep learning-based pipeline [18] and released by the UK Biobank . For generalizability and cross-validation, `MESHAgents` mine phenotypes and factors on a training set of 26,893 participants, and tested on the expanded dataset of 38,309 participants, which includes the initial cohort plus additional participants. Meanwhile, we perform hypothesis testing with existing expert knowledge in Fig. 1, and performance comparison with automatic systems in Table 1 and Fig. 3. All experiments are performed on two Nvidia RTX 6000 GPUs (48 GB).

Benchmarks. (i) The LLMs we evaluate are the latest version of GPT-4o and GPT3.5 [36]. All models are set in a zero-shot Chain-of Thought (CoT) [38] setting. (ii) For medical multi-agent frameworks: we select MedAgents [26] and RareAgents [27] as the latest medical frameworks. For those LLMs that originally cannot complete the PheWAS task, we introduce the task definition into their agent initialization process for fair comparison.

Evaluation Metrics. (i) In auto PHeWAS: We evaluate automatically discovered phenotypes through two complementary metrics: *Dependency:* Measures both phenotypic redundancy and hallucination tendencies. For phenotype set $\mathcal{P}$, calculated as $Q(\mathcal{P}) = [1 - \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} |corr(p_i, p_j)|] \times \frac{|\mathcal{P}_{valid}|}{|\mathcal{P}|}$, where $|\mathcal{P}_{valid}|$ is the number of valid phenotypes in set $\mathcal{P}$. Lower values indicate greater phenotypic independence and fewer hallucinated features. *Coverage:* Assesses anatomical comprehensiveness across six anatomical structures (LV, RV, LA, RA, AAo, DAo) in [39, 18] . We define coverage as $C(\mathcal{P}) =$

http://www.ukbiobank.ac.uk/register-apply

https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=157

Model	Auto PheWAS	
	Dependency↓	Coverage↑
Independent LLMs (zero-shot CoT)		
GPT-3.5 [36]	0.642	0.524
GPT-4omini [36]	<u>0.356</u>	0.767
Claude-3.5 [37]	0.520	0.841
Medical Multi-agent Frameworks		
MedAgents [26]	0.509	0.829
RareAgents [27]	0.363	0.629
w/o Mem. & Tool	0.492	<u>0.847</u>
w/o Stage III	0.385	**0.871**
`MESHAgents`	**0.350**	**0.871**

Table 1: Performance comparison of different Agent-based models on two metrics for Auto PheWAS.

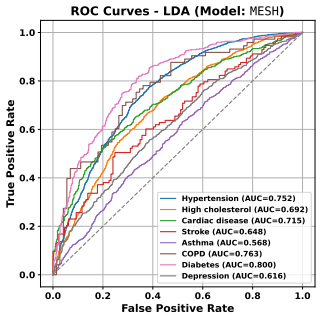


Fig. 3: ROC curves (LDA) for classification across cardiovascular conditions using discovered features.

Table 2: Performance comparison between experts-selected and MESHAgents-discoved on disease classification tasks. (Green and bold values indicate where MESHAgents outperformed analyses using human-selected phenotypes. Summary: For SVM, MESHAgents showed $\overline{AUC} = -0.013_{\pm 0.076}$ and $\overline{Recall} = -0.005_{\pm 0.042}$; for LDA, $\overline{AUC} = -0.004_{\pm 0.010}$ and $\overline{Recall} = +\mathbf{0.020}_{\pm \mathbf{0.033}}$; for AdaBoost, $\overline{AUC} = -0.014_{\pm 0.036}$ and $\overline{Recall} = -0.009_{\pm 0.040}$.)

	Hypertension		High cholesterol		Cardiac disease		PVD		Stroke		Asthma		COPD		Diabetes		Depression	
	AUC	Recall	AUC	Recall	AUC	Recall	AUC	Recall	AUC	Recall	AUC	Recall	AUC	Recall	AUC	Recall	AUC	Recall
SVM																		
Experts	0.739	0.708	0.645	0.672	0.671	0.548	0.316	0.462	0.398	0.805	0.549	0.691	0.561	0.521	0.730	0.670	0.573	0.667
`MESHAgents`	0.726	0.706	0.635	**0.716**	0.645	0.513	0.265	0.462	**0.619**	0.779	0.549	0.676	0.549	0.438	0.678	**0.678**	0.568	**0.679**
LDA																		
Experts	0.753	0.687	0.692	0.655	0.722	0.614	0.479	0.308	0.644	0.628	0.570	0.567	0.760	0.630	0.801	0.732	0.619	0.625
`MESHAgents`	0.752	**0.691**	0.692	**0.660**	0.715	**0.624**	**0.521**	**0.462**	**0.648**	0.619	0.568	**0.568**	**0.763**	**0.685**	0.800	0.732	0.616	0.616
AdaBoost																		
Experts	0.747	0.705	0.676	0.670	0.701	0.609	0.487	0.538	0.587	0.602	0.548	0.542	0.697	0.644	0.775	0.716	0.595	0.611
`MESHAgents`	0.743	0.700	0.673	0.658	0.689	0.601	**0.538**	0.462	**0.619**	0.566	**0.554**	**0.552**	0.649	0.644	**0.777**	**0.724**	0.580	0.598

$w_s \times \frac{|\Omega_{covered}|}{|\Omega|} + w_f \times \frac{|F_{covered}|}{|F_{total}|}$, where $\Omega_{covered}$ represents covered anatomical structures and $F_{covered}$ represents covered structure-function combinations. Higher coverage indicates global representation. (ii) In disease diagnosis: We assess the clinical utility of the MESHAgents-discovered phenotype-factor associations $\mathcal{A}$, using these features to train classification models for nine disease types. We implement three classification algorithms (AdaBoost [40], LDA [41], and SVM [42]) and report two evaluation metrics for classification performance: AUC, which measures overall discriminative ability, and recall, which quantifies the model's capacity to correctly identify positive disease cases. All results are obtained through five-fold cross-validation to ensure statistical reliability, and the practical value of automatically discovered phenotype-factor set.

Results. (i) Auto PheWAS: Independent LLMs generated unreliable phenotype sets with high dependency (GPT-3.5: 0.642), trade-off (GPT-4omini: 0.356 but 0.767), and limited coverage (GPT-3.5: 0.524), reflecting hallucination tendencies and catastrophic forgetting. Existing multi-agent frameworks showed

improvement but exhibited polarization biases (RareAgents: 0.629), either overfocusing on specific imaging phenotypes or generating redundant phenotypes (MedAgents: 0.509). MESHAgents overcame these limitations through three key components: Firstly, evidence-augmented analysis with tools and memory significantly reduced hallucination; Second, the consensus mechanism promoted cross-checking while maintaining diversity; and the distributed expertise analysis prevented knowledge distortion. These components combined to achieve the best dependency score (0.350) and strong coverage (0.871) among all automated methods for PheWAS.

(ii) Disease Association Study: The comparison with imaging phenotypes selected by human in literature [18, 7]. While MESHAgents do not fully match human patterns, it independently discovered significant relevant phenotypes without domain-specific knowledge. Fig. 1(a) shows explainable overlap between MESHAgents-discovered and expert-selected parameters, with MESHAgents identifying alternative markers with diagnostic value. Importantly, MESHAgents achieves this automation while maintaining validity, as evidenced by its disease identification in Fig. 1(b). This shows MESHAgents' ability to effectively balance automated exploration with clinical relevance, offering complementary advantages to traditional expert-driven phenotype association studies.

(iii) Diagnosis: Table 2 shows a comparison between MESHAgent-discovered and expert-selected imaging phenotypes in disease classification across three classifiers. Notably, MESHAgents shows comparable overall performance with mean AUC differences of only $-0.013_{\pm 0.076}$ (SVM), $-0.004_{\pm 0.010}$ (LDA), and $-0.014_{\pm 0.036}$ (AdaBoost). For recall metrics—MESHAgent achieves $\overline{\text{Recall}}$ of $-0.005_{\pm 0.042}$ (SVM), $+0.020_{\pm 0.033}$ (LDA, outperforming human-selected features), and $-0.009_{\pm 0.040}$ (AdaBoost). The superior recall stems from MESHAgents' ability to identify complementary phenotypes that better capture disease variance in linear discriminant space in Fig. 3. MESHAgents-discovered phenotypes achieved superior results in specific conditions like Stroke (AUC +0.221 in SVM), PVD (AUC +0.042 in LDA), and Diabetes. This suggests that MESHAgents' automatic phenotype selection captures clinically relevant patterns without requiring extensive domain knowledge. The classification results across Fig. 3 further validate that MESHAgents-identified parameters maintain robust diagnostic value across different conditions, demonstrating the generalizability and scalability of our approach. These findings suggest that automatic phenotype discovery through multi-agent reasoning can effectively complement expert feature selection, potentially reducing subjective bias while maintaining clinical relevance in cardiovascular studies.

4 Conclusion

The framework could be further enhanced through the integration of vector embedding bases. For knowledge integration, an embedding library of medical imaging literature could be incorporated into our existing agent communication protocols, allowing phenotype-confounder evaluations to reference specific

studies. In general, `MESHAgents` framework demonstrates advantages of multi-agent reasoning for medical imaging phenotype analysis compared to traditional approaches that rely on predefined criteria.

Acknowledgements W.Z. is supported by the JADS programme and the UKRI Centre for Doctoral Training in AI for Healthcare (EP/S023283/1). M.Q. is supported by the EPSRC DeepGeM Grant (EP/W01842X/1) and the Dame Julia Higgins Postdoc Collaborative Research Fund. W.B. is supported by the EPSRC DeepGeM Grant (EP/W01842X/1), CVD-Net Programme Grant (EP/Z531297/1) and the BHF New Horizons Grant (NH/F/23/70013). S.N. is supported by the National Institutes of Health (R01-HL152256), the European Research Council (PREDICT-HF 864055), the British Heart Foundation (RG/20/4/34803), the EPSRC (EP/X012603/1 and EP/P01268X/1), the Technology Missions Fund under the EPSRC (EP/X03870X/1), and the Alan Turing Institute. P.M.M. acknowledges generous personal support from the Edmond J. Safra Foundation and Lily Safra, an NIHR Senior Investigator Award, Rosalind Franklin Institute, and the UK Dementia Research Institute, which is funded predominantly by the UKRI Medical Research Council. B.K. acknowledges the HPC resources provided by NHR@FAU under the NHR projects b143dc and b180dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the DFG - 440719683. Additional support was received by the ERC - project MIA-NORMAL 101083647, DFG 513220538, 512819079, and by the state of Bavaria (HTA). This research was conducted using the UK Biobank Resource under Application Number 18545. We thank all UK Biobank participants and staff.

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

References

1. Sing CF, Stengård JH, Kardia SL. Genes, environment, and cardiovascular disease. Arteriosclerosis, Thrombosis, and Vascular Biology. 2003;23(7).
2. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nature Reviews Cardiology. 2015;12(11).
3. Rozanski A, Others. Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. Circulation. 1999.
4. Skelly AC, Dettori JR, Brodt ED. Assessing bias: the importance of considering confounding. Evidence-based spine-care journal. 2012;3(01):9-12.
5. Mukherjee P, Shen TC, Liu J, Mathai T, Shafaat O, Summers RM. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. Nature Medicine. 2022;28(6).
6. Qiao M, Wang S, Qiu H, De Marvao A, O'Regan DP, et al. CHeart: A conditional spatio-temporal generative model for cardiac anatomy. IEEE Transactions on Medical Imaging. 2023.
7. Qiao M, McGurk KA, Wang S, Matthews PM, O'Regan DP, Bai W. A personalized time-resolved 3D mesh generative model for unveiling normal heart dynamics. Nature Machine Intelligence. 2025:1-12.
8. Touvron H, Lavril T, et al. Llama: Open and efficient foundation language models. ArXiv preprint. 2023;abs/2302.13971.
9. OpenAI. GPT-4 Technical Report. ArXiv preprint. 2023;abs/2303.08774.
10. Lu P, Peng B, Others. Chameleon: Plug-and-play compositional reasoning with large language models. arXiv preprint arXiv:230409842. 2023.
11. Park JS, O'Brien JC, Others. Generative agents: Interactive simulacra of human behavior. In: UIST. Association for Computing Machinery; 2023. .
12. Zhang W, Zang C, Schmidt M, Blythman R. BiD: Behavioral agents in dynamic auctions; 2024.
13. Zhang W, Zang C, Kainz B. Truth or deceit? A Bayesian decoding game enhances consistency and reliability; 2024.
14. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956).
15. Singhal K, Azizi S, Tu T, Mahdavi S, Wei J, et al. Large language models encode clinical knowledge. Nature. 2023 07;620.
16. Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. AlpaCare: Instruction-tuned Large Language Models for Medical Application; 2023.
17. Bao Z, Chen W, Xiao S, Ren K, Wu J, et al.. DISC-MedLLM: Bridging general large language models and real-world medical consultation; 2023.
18. Bai W, Suzuki H, Huang J, Francis C, Wang S, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Medicine. 2020;26(10).
19. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? arXiv preprint arXiv:220708143. 2022.
20. Schmidt HG, Rikers RM. How expertise develops in medicine: knowledge encapsulation and illness script formation. Medical Education. 2007;41(12).
21. Tarale P, Rietman E, Siegelmann HT. Distributed multi-agent lifelong learning. Transactions on Machine Learning Research. 2025.
22. Zheng G, Others. Towards a collective medical imaging AI: Enabling continual learning from peers. In: MIDL; 2024. .

23. Harris E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. JAMA. 2023.
24. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Computing Surveys. 2023;55(12):1-38.
25. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972).
26. Tang X, Zou A, Zhang Z, Li Z, Zhao Y, Zhang X, et al. Medagents: Large language models as collaborators for zero-shot medical reasoning. arXiv preprint arXiv:231110537. 2023.
27. Chen X, Jin Y, Mao X, Wang L, Zhang S, Chen T. RareAgents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. arXiv:241212475. 2024.
28. Lu M, Ho B, Ren D, Wang X. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In: Findings of the Association for Computational Linguistics: EMNLP; 2024. p. 5747-64.
29. Li B, Yan T, Pan Y, Luo J, Ji R, et al. MMedAgent: Learning to use medical tools with multi-modal agent. arXiv:240702483. 2024.
30. Li Y, Zhang Y, Sun L. MetaAgents: Simulating interactions of human behaviors for LLM-based task-oriented coordination via collaborative generative agents. arXiv:231006500. 2023.
31. Wang Z, Mao S, Wu W, Ge T, Wei F, Ji H. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. NAACL. 2024.
32. Wu Q, Bansal G, Zhang J, Wu Y, Zhang S, Zhu E, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. arXiv:230808155. 2023.
33. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving factuality and reasoning in language models through multiagent debate. In: ICML; 2023. .
34. Fu Y, Peng H, Khot T, Lapata M. Improving language model negotiation with self-play and in-context learning from ai feedback. arXiv:230510142. 2023.
35. Yang Z, Zhang Z, Zheng Z, Jiang Y, et al. Oasis: Open agents social interaction simulations on one million agents. arXiv preprint arXiv:241111581. 2024.
36. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. arXiv:230308774. 2023.
37. Anthropic. Claude 3.5; 2023. https://www.anthropic.com/claude.
38. Wei J, Wang X, et al. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS. 2022;35:24824-37.
39. Strocchi M, Augustin CM, Gsell MA, Karabelas E, Neic A, Gillette K, et al. A publicly available virtual cohort of four-chamber heart meshes for cardiac electromechanics simulations. PloS one. 2020;15(6):e0235145.
40. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. 1997;55(1):119-39.
41. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics. 1936;7(2):179-88.
42. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273-97.