nature biomedical engineering

Article

https://doi.org/10.1038/s41551-025-01608-0

Learning neuroimaging models from health system-scale data

Received: 27 August 2025

Accepted: 15 December 2025

Published online: 06 February 2026

Check for updates

Yiwei Lyu[1,7], Samir Harake[2,7], Asadur Chowdury[2,7], Soumyanil Banerjee[2], Rachel Gologorsky[2], Shixuan Liu[1], Anna-Katharina Meissner[3], Akshay Rao[2], Chenhui Zhao[1], Akhil Kondepudi[2,4], Cheng Jiang[2,4], Xinhai Hou[2,4], Rushikesh S. Joshi[2], Volker Neuschmelting[3], Ashok Srinivasan[5], Dawn Kleindorfer[6], Brian Athey[4], Vikas Gulani[5], Aditya Pandey[2], Honglak Lee[1] & Todd Hollon[1,2,4]

Neuroimaging is a ubiquitous tool for evaluating patients with neurological diseases. The global demand for magnetic resonance imaging (MRI) studies has risen steadily, placing substantial strain on health systems, prolonging turnaround times and intensifying physician burnout. These challenges disproportionately impact patients in low-resource and rural settings. Here we utilize data from a large academic health system to develop Prima, an AI foundation model for neuroimaging that supports real-world, clinical MRI studies as input. Trained on over 220,000 MRI studies, Prima uses a hierarchical vision architecture that provides general and transferable MRI features. Prima was tested in a 1-year health system-wide study that included 29,431 MRI studies. Across 52 radiologic diagnoses from major neurologic disorders, Prima achieved a mean diagnostic area under the curve (AUC) of 92.0%, outperforming other state-of-the-art general and medical AI models. Prima offers explainable differential diagnoses, worklist priority for radiologists and clinical referral recommendations. Prima demonstrates algorithmic fairness across sensitive groups. These findings highlight the transformative potential of health system-scale AI training and Prima's role in advancing AI-driven healthcare.

Health systems function as powerful data engines for developing medical foundation models[1,2]. Routine clinical operations generate vast volumes of electronic medical records, which can be used to train medical vision-language models (VLMs) in a manner analogous to the way that internet-scale data is used to train VLMs such as contrastive language-image pre-training (CLIP)[3], DALL-E[4] and Flamingo[5]. Globally, approximately 100 million magnetic resonance imaging (MRI) studies are performed annually, with 20–30% focused on neurological diseases. The demand for brain MRI studies surpasses the available neuroradiology services[6–8]. This imbalance has caused substantial healthcare challenges, including workforce shortages, increased workloads, burnout and more diagnostic errors[9–15]. In addition, health disparities in radiology have been exacerbated owing to limited resources and a contracting workforce[16]. Innovative technologies are needed to improve patient access to radiology services, especially in rural areas and low-/middle-income countries. A synergistic collaboration between AI and health systems is essential to address these challenges and improve healthcare delivery.

[1]University of Michigan Computer Science and Engineering, Ann Arbor, MI, USA. [2]University of Michigan Neurosurgery, Ann Arbor, MI, USA. [3]University of Cologne Neurosurgery, Cologne, Germany. [4]University of Michigan Computational Medicine and Bioinformatics, Ann Arbor, MI, USA. [5]University of Michigan Radiology, Ann Arbor, MI, USA. [6]University of Michigan Neurology, Ann Arbor, MI, USA. [7]These authors contributed equally: Yiwei Lyu, Samir Harake, Asadur Chowdury. e-mail: tocho@med.umich.edu

https://doi.org/10.1038/s41551-025-01608-0

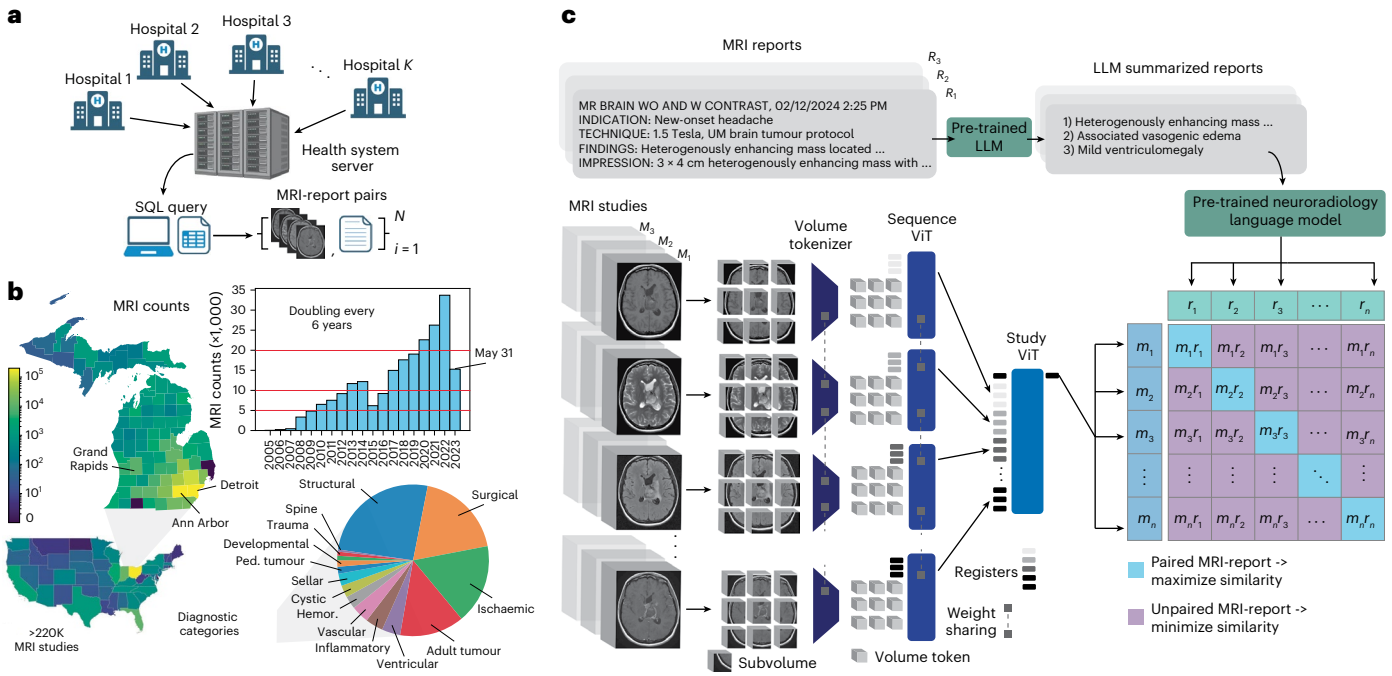


Fig. 1 | Overview of the UM-220K MRI dataset and Prima workflow. a, Over 220,000 brain MRIs were queried from our health system's PACS, forming the UM-220K dataset. This dataset includes MRI studies from multiple medical centres across the state and the United States. **b**, The distribution of MRI counts by county and state is presented. The number of MRIs archived in the PACS system has doubled approximately every 6 years over the past 2 decades, highlighting the growing demands on radiology and clinical services. The diagnostic categories reflect the standard operations of a large academic medical centre. Ped., pediatric; Hemor., hemorrhagic. **c**, Prima was trained using a CLIP framework and a hierarchical ViT architecture. Full MRI studies were divided into subvolumes, compressed into volume tokens using a tokenizer and processed by a sequence ViT to extract sequence-level features. Global sequence registers were passed to a study ViT to generate a study-level representation for alignment with radiology reports. Radiology reports were summarized using an LLM, and a pre-trained neuroradiology language model generated report representations. Finally, the MRI study embeddings and summarized report embeddings were aligned using a CLIP objective. Illustrations in **a** created with BioRender.com.

Prima is a general-purpose volumetric MRI VLM trained on health system-scale data, forming a foundation for addressing diverse radiologic and clinical prediction tasks. Traditional approaches to applying AI to MRI studies have relied on manually curated subsets of MRI sequences, such as the fluid-attenuated inversion recovery (FLAIR) sequence for lesion detection or T1-weighted images for dementia prediction[17,18]. These models are limited by partial radiologic information compared with a radiologist's interpretation of all MRI sequences. Like a radiologist, Prima integrates information from the clinical context, study indication and all MRI sequences to produce a comprehensive vector representation of the full study, enabling better performance across a broad range of prediction tasks. We demonstrate that Prima's learned representations perform strongly across multiple radiologic, clinical and biomedical research tasks. This versatility highlights Prima's potential in optimizing neuroimaging workflows, enhancing diagnostic accuracy and addressing systemic healthcare challenges.

Results

Health system-scale vision-language models

To create a large, diverse neuroimaging dataset for VLM development, we queried our health system's picture archiving and communication system (PACS) for all brain MRIs on 31 May 2023 (Fig. 1 and Supplementary Code 1). After data curation and quality assurance (Extended Data Fig. 1a), the UM-220K neuroimaging dataset contained 221,147 MRI studies with paired radiology reports (Extended Data Fig. 1a) from over 170,000 patients. UM-220K contains 5.6 million MRI sequences, 362 million MRI slices and 3.2 billion volume tokens (Fig. 1). UM-220K is the largest MRI dataset and includes all patients treated or referred to our health system and/or affiliated hospitals since the start of radiology digitization over two decades ago. We aimed to collect a neuroimaging dataset representative of the diverse patient populations and demographics encountered by tertiary health systems delivering primary and specialty care for the full spectrum of neurologic diseases (Extended Data Fig. 2). Manually annotating data at this scale is not feasible. With expert-engineered prompts and radiology reports (Supplementary Fig. 3), we leveraged HIPAA-compliant GPT-4 to label MRI studies for 52 radiologic diagnoses from the major neurologic disorders, including neoplastic, inflammatory, infectious and vascular lesions[19]. Our labelling strategy focused on selecting a diverse, clinically actionable subset of diagnoses to showcase Prima's ability to learn from health system data. The large language model (LLM) achieved an average annotation accuracy of 94.0 ± 1.1%, comparable to expert human annotators across diagnostic categories (Extended Data Fig. 3b).

We designed a hierarchical vision model to align with the MRI data structure, encompassing anatomic regions, MRI sequences and full studies. Prima's modular components were trained in three stages: volume tokenization, sequence/study feature learning and transfer learning for downstream tasks (Extended Data Fig. 1b). First, each MRI sequence was divided into subvolumes (Fig. 1). Inspired by the success of language tokenization and latent diffusion models[20], these subvolumes were transformed into latent volume tokens using a three-dimensional (3D) vector-quantized variational autoencoder (VQ-VAE). The VQ-VAE volume tokenizer is trained at a 16× compression rate using an $\mathcal{L}_1$ reconstruction objective. We achieved high-quality subvolume reconstructions across various MRI sequences, orientations and pathologies (Extended Data Fig. 4). Compressed volume tokens were pre-saved following model convergence for efficient downstream sequence and study-level training.

https://doi.org/10.1038/s41551-025-01608-0

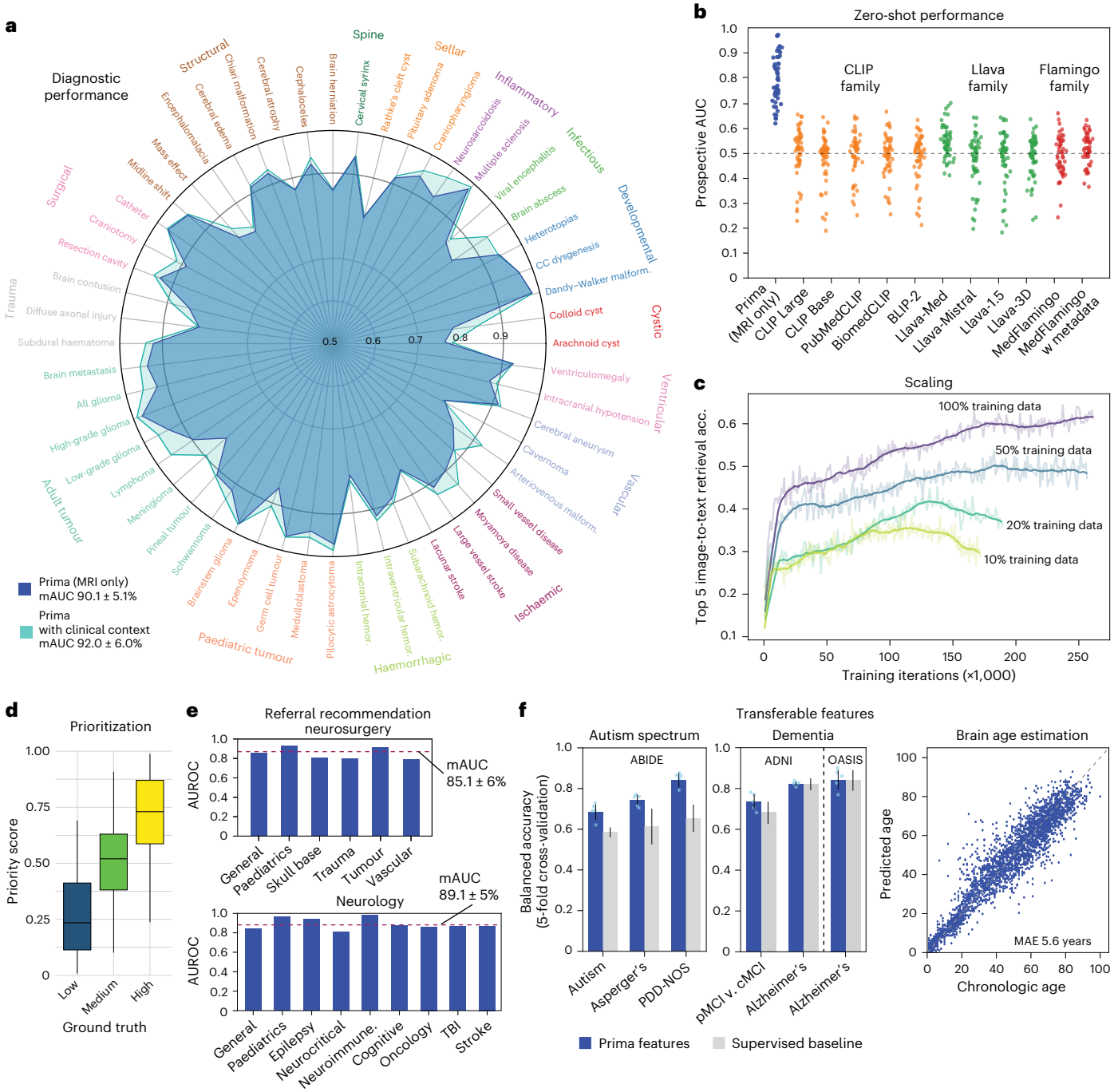


Fig. 2 | Overall performance. a, Radar plot of AUROC values for each diagnostic task in the testing cohort. Prima demonstrates strong performance across all major diagnostic categories. Similar to radiologists, Prima's diagnostic performance improves given the clinical context/indication. Malform., malformation; CC, corpus callosum. **b**, Vision-language alignment was evaluated via zero-shot classification performance. Each of the 52 diagnoses is shown as a dot for each model. MRI-only Prima outperforms both general and medical VLM models. **c**, Prima's generalization performance improves consistently with increased training data, consistent with neural network scaling laws[33]. Performance remains non-saturating with respect to training set size, suggesting that additional data will further enhance performance. Acc., accuracy. **d**, Box-and-whisker plots illustrating the prioritization of MRIs based on diagnostic findings (n = 29,431). Box plots show the median (centre line), interquartile range (box, 25th–75th percentile) and whiskers extending to the minimum and maximum values. Prima's priority scores correlate strongly with ground truth severity scores, yielding a Pearson's correlation coefficient of $\rho = 0.69$ (95% CI, 0.68–0.70, two-sided $P = 2.2 \times 10^{-16}$). TBI, traumatic brain injury. **e**, AUROC and mAUC values for referral recommendations in neurosurgery and neurology specialties, demonstrating high performance. **f**, Prima's ability to transfer features to out-of-domain tasks. Bar charts present 5-fold cross-validation results (mean ± standard deviation, n = 5) for autism spectrum disorder and dementia classification, where Prima outperforms supervised baselines from the literature[18,72]. Scatter plot shows brain age estimation results across the lifespan, based on clinical MRIs without exclusions.

The volume tokens from individual MRI sequences were then input into an MRI sequence vision transformer, ViT$_{seq}$ (ref. 21). Three-dimensional position embeddings were concatenated with volume token features to encourage position-aware feature learning. ViT$_{seq}$ is a multimodal ViT in two ways. First, the parameters are shared across all sequence types, orientations and protocols acquired during standard clinical operations. Second, ViT$_{seq}$ uses the sequence description metadata, such as 'AX_T2' or '_plus_COR_MPR', as a free text prompt for better

https://doi.org/10.1038/s41551-025-01608-0

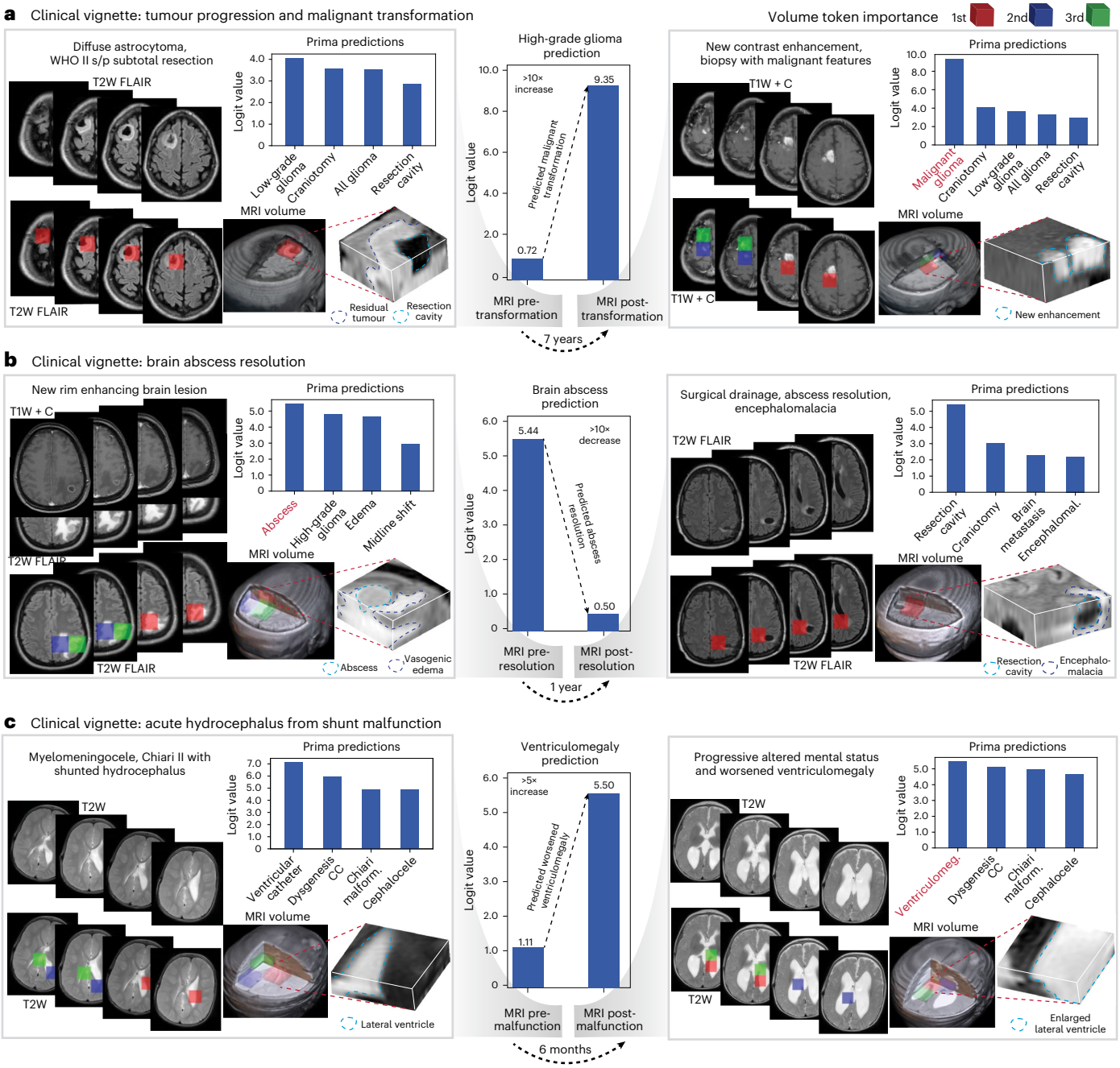


Fig. 3 | Explainable Prima predictions in clinical context. Three clinical vignettes demonstrate Prima's explainability using LIME. The left panels show patient MRIs at initial presentation (MRI Pre) with the top Prima logits and the Top-3 volume tokens identified by LIME. The centre bar charts depict changes in Prima logits between the initial presentation (MRI Pre) and after progression or intervention (MRI Post). The right panels show patient MRIs following their clinical courses (MRI Post). **a**, Clinical vignette of a diffuse low-grade glioma patient, status post (s/p) subtotal resection, who experienced tumour progression and malignant transformation 7 years after treatment. Prima accurately identified new regions of contrast enhancement, consistent with malignant glioma. **b**, Clinical vignette of a patient with a spontaneous brain abscess who underwent surgical drainage and antibiotic treatment, resulting in resolution. Encephalomal., encephalomalacia. **c**, Clinical vignette of a paediatric patient with a history of myelomeningocele and shunted hydrocephalus. At baseline, the patient had mild ventriculomegaly but presented with acute hydrocephalus following shunt malfunction. Prima accurately predicted the worsening of ventriculomegaly. Ventriculomeg., ventriculomegaly. Interactive demonstration can be found at prima.mlins.org.

feature extraction (Extended Data Fig. 3c). Finally, a set of registers were concatenated to the multimodal text-volume token sequence. ViTs are known to store global discriminative features in register tokens that can be used for downstream tasks[22]. The output registers from each MRI sequence are then input into a study ViT, ViT_{st}. The ViT_{st} aggregates MRI features using self-attention across the sequence registers. Study classification tokens were used to obtain the full MRI study representation.

Prima is trained using a CLIP framework[3]. The objective is to align a full MRI study representation with its corresponding radiology report[23–26]. However, raw radiology reports contain textual information that can minimize a CLIP objective but are not important for downstream diagnostic tasks, such as protocol information or radiologists' word choices. Moreover, radiology reports can be a source of bias and reduce algorithmic fairness[27]. An LLM, HIPAA-compliant GPT-3.5-turbo,

was prompted to summarize each report to distil and itemize the most important diagnostic findings, improving representation learning while minimizing bias and distribution shift (Supplementary Fig. 3). ViT_{seq} and ViT_{st} are trained jointly using summarized radiology report supervision with a pre-trained GPT-2 language model as the text CLIP text encoder[28]. Leveraging the inherent MRI data structure, we added a patient sequence discrimination objective that encourages the ViT_{seq} sequence representations to be similar within a patient's MRI study. MRI features such as brain morphology or pathologic lesions will be shared across MRI sequences and should have similar representations. The patient discrimination objective enforces that these shared features are consistently represented across sequences and improves model convergence (Extended Data Fig. 5e).

Clinical testing of Prima

We conducted a 1-year, clinical, offline, health system-wide diagnostic study to test Prima. All patients evaluated at our health system and who received a brain MRI between 1 June 2023 and 30 May 2024 were included as study subjects without exclusion. Our study was designed to reliably simulate the clinical setting in which Prima would be deployed for patient care. The testing cohort included 29,431 patients, exceeding the minimum calculated sample size of 22,338. The primary objective was to assess Prima's performance in a multi-label differential diagnosis task spanning 52 neuroradiologic diagnoses. Importantly, radiologists interpret MRI studies using the clinical context and study indications. We aimed to include this clinical context to allow Prima to make more informed diagnostic predictions. The clinical history and study indication from the patient's electronic medical record were embedded using an LLM (OpenAI's text-578 embedding-3-small). The clinical context embedding was concatenated with the MRI study features and used to train a multilayer perceptron (MLP) as a 52-way diagnostic classifier. Clinical context-aware Prima achieved a mean area under the receiver operating characteristic curve (AUROC) of 92.1 ± 5.5%, compared with 90.1 ± 5.0% with MRI-only Prima. AUROC scores ranged from 78.3% for arachnoid cysts to 99.7% for high-grade gliomas (Fig. 2a and Extended Data Fig. 5b).

We compared Prima with state-of-the-art general VLMs, including OpenAI's CLIP models[3] and Microsoft's LLaVa models[29]. We also compared Prima to state-of-the-art medical VLMs, including PubMedCLIP[30], BioMedCLIP[31] and Med-Flamingo[32]. Prima showed better vision-language alignment by outperforming all models on zero-shot diagnostic performance (Fig. 2b and Extended Data Fig. 5a). We compared Prima and open-source CLIP models using MLP probing, which includes training an MLP classifier on the UM-220K dataset and tested on the clinical cohort. Prima achieved top performance (Extended Data Fig. 5b), indicating the value of the UM-220K dataset and CLIP pre-training on 3D images. Foundation models show performance scaling laws with increased dataset size and compute budget[33]. Prima demonstrated consistent performance improvements with larger training datasets and compute budgets (Fig. 2c). Consistent improvement in MRI-report alignment was observed on Top-1 and Top-5 retrieval metrics. These results demonstrate that Prima has foundation model properties, and reported performance will continue to improve with additional health system training data and larger compute budgets. Additional subgroup, scaling, vision-language alignment and confidence calibration analysis[34] can be found in Extended Data Fig. 6.

Next, Prima was tested on two clinical tasks: radiologist's worklist prioritization and clinical referral recommendation. A classifier was trained on frozen Prima features to predict ground truth priority and clinical referrals, which were determined based on the radiologic diagnoses (Supplementary Table 3). For example, patients with evidence of subdural haematomas were assigned high priority, whereas patients with arachnoid cysts or unremarkable scans were assigned lower priority (Extended Data Fig. 7a). Prima's normalized priority scores were strongly correlated with three-tier ordinal priority scores (normal, medium and high), yielding a correlation coefficient of $\rho = 0.69$ (95% CI, 0.68–0.70, $P < 0.001$) (Fig. 2d and Extended Data Fig. 7a). Prima was then evaluated on referral recommendations to neurology and neurosurgery specialty care based on MRI features. For example, patients with newly diagnosed multiple sclerosis should be referred to a neuroimmunology specialist. Prima achieved an average neurosurgery referral AUROC of 85.1 ± 6.0% and neurology referral AUROC of 89.1 ± 5.0% (Fig. 2e). These results demonstrate how Prima can improve workflows and streamline clinical care.

Transferable Prima features

CLIP-learned visual representations can transfer effectively to various downstream tasks, including out-of-distribution scenarios[3]. We evaluated Prima's MRI representations using a linear evaluation protocol on three benchmarked neuroimaging tasks: autism spectrum prediction[35], dementia/Alzheimer's disease prediction[36,37] and brain age estimation[38]. These tasks are considered out-of-distribution because the summarized radiology reports lack information about the patient's clinical diagnoses or age. For autism spectrum and dementia predictions, Prima matched or exceeded performance of independent fully supervised and semi-supervised benchmarks on three publicly available datasets: ABIDE[35], ADNI[36] and OASIS[37] (ref. 18) (Fig. 2f). Using Prima features, brain age estimation yielded a mean absolute error (MAE) of 5.6 years on our testing dataset. These results are competitive with existing models trained end to end for brain age estimation on large, uncurated, clinical cohorts[39]. Our findings also demonstrate that Prima's performance transfers effectively to other public datasets and diagnostic tasks, including diffuse gliomas classification (BRATS[40]), brain metastasis prediction (UCSFMets[41] and NYUMets[42]) and acute strokes detection[43] (Extended Data Fig. 7b). Notably, previous models require extensive preprocessing, including skull stripping, sequence selection, resampling and segmentation. Prima's flexible architecture enables predictions from any sequence or sequence combination without preprocessing.

Explainable Prima predictions

Explainable AI is essential in healthcare to ensure safe, reliable and trustworthy predictions[44]. We assessed Prima's predictions using local interpretable model-agnostic explanations (LIME)[45]. LIME assigns importance scores to individual volume tokens in an MRI, with higher scores indicating greater contributions to Prima's predictions. If LIME highlights pathologic regions in an MRI with high importance scores, then Prima's diagnostic predictions are aligned with clinical reasoning. Figure 3 showcases three clinical vignettes illustrating Prima's value with LIME explanations: malignant brain tumour transformation, brain abscess resolution and acute hydrocephalus owing to shunt malfunction. Each vignette illustrates Prima's ability to generate accurate and trustworthy predictions across a patient's clinical course. For example, LIME revealed that Prima accurately identified regions of new contrast enhancement—a well-established radiologic marker of malignant transformation[46]—to predict the progression of a low-grade glioma into a malignant glioma. We quantitatively validated Prima by assessing its ability to assign high LIME scores to manually segmented brain tumour regions in the expert-annotated BraTS dataset. Prima achieved 98.0% Top-3 accuracy in selecting tokens within segmented brain tumour regions (Extended Data Fig. 7d). Extended Data Fig. 8 shows LIME visualizations for various pathologies, including paediatric, inflammatory, infectious and developmental lesions, underscoring Prima's versatility. As a multi-label classifier, we performed a multi-label analysis to demonstrate that Prima learns the co-occurrence between correlated diagnoses, such as brain contusion and midline shift (Extended Data Fig. 9a). Prima selects different volume tokens when making different diagnostic decisions for the same patient and MRI study. Prima correctly selects volume tokens in the posterior fossa when diagnosing a paediatric cerebellar brain tumour and selects tokens in the lateral ventricles to

https://doi.org/10.1038/s41551-025-01608-0

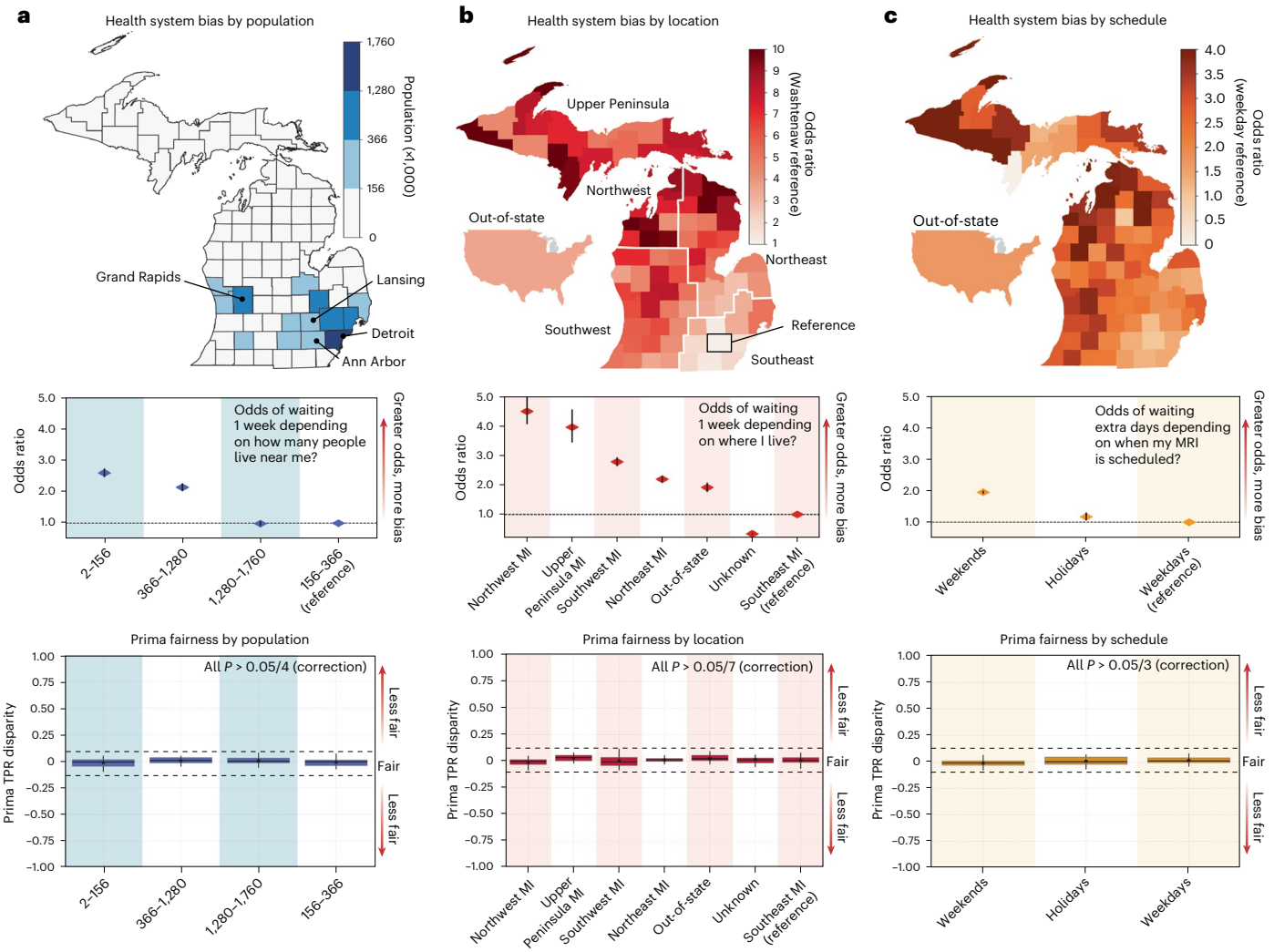


Fig. 4 | Health system bias and algorithmic fairness. a, State map showing county populations, grouped into quartiles with equal population sizes. Odds ratios for patients experiencing a 7-day turnaround time are shown for each quartile and computed over studies conducted in Michigan (n = 192,744 studies). Systemic biases were observed, particularly in sparsely populated regions (P < 0.001). Prima demonstrated algorithmic fairness, with minimal TPR disparity across these population groups. **b**, State map illustrating counties' odds ratios for a 7-day turnaround time based on location. The odds ratio plot is further divided by state regions (n = 241,372 studies). Systemic biases were prominent in rural areas, especially in Northwest Michigan and the Upper Peninsula (P < 0.001). Despite these biases, Prima maintained consistent fairness across the state and the United States. **c**, State map showing counties' odds ratios for MRI scheduling delays exceeding 2 days (n = 241,372 studies). Turnaround time biases were identified for weekend and holiday MRIs (P < 0.001). Prima exhibited minimal TPR disparity across these subgroups. *Note on statistical tests: in the forest plots, diamonds represent odds ratios, and error bars indicate 95% confidence intervals. Intervals not intersecting the dotted line are statistically significant, and P values are computed via a two-sided chi-square test of independence. Box-and-whisker plots show quartile values of TPR disparity, with the bold centre line indicating the median, black dots representing the mean and whiskers extending to data within 1.5 × the interquartile range. Each distribution is made up of 20 bootstraps of 200 samples each from the corresponding subgroup and a diagnostically stratified reference group that comes from our test set of 29,431 studies (see Methods for detailed description of the process). The P value for each TPR disparity distribution is a one-sided, non-parametric Mann–Whitney U statistical test where the null hypothesis is that the subgroup's TPR distribution is not less than the reference TPR distribution. Black dashed lines denote 10% TPR disparity.

diagnose associated ventriculomegaly from obstructive hydrocephalus (Extended Data Fig. 9b).

Health system bias and Prima fairness

Health system bias and health disparities are pervasive across all medical specialties[16,47]. Ensuring algorithmic fairness is critical for medical AI models to achieve equitable performance across sensitive attributes and mitigate existing disparities[27]. To assess Prima's algorithmic fairness, we examined a key modifiable source of health system bias in radiology—turnaround time. Turnaround time is the interval between when an imaging exam is performed and when the radiologist's report is accessible to the referring healthcare provider. Quick turnaround times are critical as timely diagnosis can impact patient care. Importantly, turnaround time is influenced by various health system factors, including imaging study complexity, radiologist workload, the need for specialized interpretation and overall health system efficiency. Final interpretation timestamps were used to calculate turnaround times in UM-220K. Owing to increasing radiology volumes, the average turnaround time at our health system has increased over time from a low of approximately 18 h in 2012 to a high of over 2.25 days in 2024 (Extended Data Fig. 2b). The turnaround time distribution showed a large right skew towards longer turnaround times, with the majority of turnaround time measured in patient-days found in the right tail of the distribution (Extended Data Fig. 2c). We identified three sensitive attributes that account for the tail distribution and lead to systemic biases affecting turnaround time: population density, geographic region and

scheduling. Patients in sparsely populated rural areas were 2–5 times more likely to experience a 7-day turnaround time compared with those in urban areas ($P < 0.001$; Fig. 4a). Patients who had an MRI scheduled during a weekend were twice as likely to wait 3 or more days ($P < 0.001$; Fig. 4a). Because the turnaround time is negligible for Prima, <3 s on a single graphical processing unit, we evaluated Prima's algorithmic fairness across sensitive groups to mitigate systemic biases. In Fig. 4, we compared the true positive rate (TPR) of sensitive groups with the study population, known as equalized opportunity[48]. Larger TPR disparity between the sensitive group and the study population indicates worse algorithmic fairness. Prima exhibited algorithmic fairness across the three primary sources of health system bias (Fig. 4). Further subgroup performance, equalized odds and intersectional analysis revealed equitable performance across patient risk factors (for example, race, sex and age), medical access factors (for example, insurance status and MRI manufacturer), reinforcing Prima's robustness and fairness (Extended Data Fig. 10 and Supplementary Fig. 4).

Discussion

Prima is a general-purpose neuroimaging VLM trained on health system-scale data, delivering general, scalable and equitable performance. Leveraging over 220,000 MRI studies (over 5.6 million 3D sequences) from diverse patient populations, Prima establishes a benchmark in radiologic diagnosis and clinical prediction. Our study underscores the potential of health system-scale models to improve clinical efficiencies and ease labour shortages. Unlike earlier neuroimaging models that rely on curated datasets and pre-selected sequences, Prima excels with large, uncurated imaging data, making it highly practical for real-world AI applications.

A phased certification pathway has been proposed for evaluated generalist medical AI systems that mirror the clinical training of physicians[49]. The first stage is evaluating AI systems for baseline competency through standardized testing and scenario analysis. Our study is limited to addressing only this initial stage of Prima's clinical certification. The reported diagnostic accuracy for expert neuroradiologists is over 94% (refs. 50–52). By evaluating via standardized testing benchmarks and clinical tasks, we aimed to show only that Prima has established a foundation of medical knowledge. Future work will explore integrating detailed clinical notes and electronic health record (EHR) data as input, advanced VLM tasks, including open-ended diagnosis, automated report generation and visual question answering. We aim for Prima's series and study representations to be seamlessly aligned with LLMs, enabling radiologist-level performance on complex interpretation tasks.

The broader impact of Prima extends beyond neuroimaging. Our proposed AI framework is broadly adaptable to other biomedical imaging modalities, such as computed tomography, radiography and ultrasound. We hope that our proposed framework can contribute to existing medical AI models for other organ systems and imaging modalities[53–56]. Prima's ability to provide comprehensive representations of clinical MRIs holds promise for advancing research in neuroimaging. Immediate research applications of Prima include brain phenotyping[57] and quantifying disease progression and treatment response[58]. Future model versions will incorporate genetic and clinicopathologic patient data to further improve predictions and explore pathophysiological insights.

Currently, Prima is limited to a single study for neuroimage interpretation. In future work, we will expand the input space to include multiple studies for pathology comparison through time. We will further improve the technology by designing model architectures and training objectives to improve computational efficiency and representation quality, allowing direct alignment between 3D tokens and specific descriptions within the report. We aim to design and integrate vision-only self-supervised training into Prima, thus ameliorating the need for language supervision.

In conclusion, Prima exemplifies the transformative potential of integrating health systems and medical foundation models to improve healthcare. As healthcare datasets grow and compute resources expand, model performance and utility are poised to scale, offering a pathway to AI-driven innovation in medicine.

Methods

Overall objectives and study design

The primary objective was to develop, optimize and evaluate a VLM trained on health system-scale data to achieve general and transferable representations of brain MRI studies. Our design emphasized (1) inclusive data criteria/'data in the wild', (2) minimize data preprocessing, (3) flexible vision-language modelling, (4) multimodal model input and (5) clinically informative and diverse prediction tasks. Owing to the heterogeneity of clinical MRI protocols, previous studies often had limited MRI sequence inclusion (for example, T1 only). We aimed to be data inclusive and develop a general vision-language modelling strategy to accommodate the full range of clinical MRI study protocols. Moreover, clinical MRI studies have an inherently hierarchical data structure: voxels > regions > sequences > studies. Utilizing this inherent structure, we developed hierarchical ViTs for brain MRI studies to achieve high-quality and transferable representations using radiology report supervision[3]. We performed a health system-scale clinical study of Prima performance to demonstrate that Prima can provide preliminary radiologic diagnoses, study triage/worklist prioritization and referral recommendations from MRI studies alone (full training and inference workflow in Extended Data Fig. 1). Finally, we aimed to show the algorithmic fairness of Prima and present those results in the context of known health system-level biases that results in healthcare disparities.

Data curation of UM-220K MRI dataset

Large-scale MRI data acquisition and curation was essential for study feasibility[59,60]. We queried Michigan Medicine's Sectra PACS systems via SQL queries to obtain all MRIs completed through 31 May 2023 that included head, brain, orbits, face or neck, resulting in 279,908 hits. Details of the SQL queries can be found in Supplementary Code 1. We then filtered the MRI dataset to ensure that all query images had (1) associated radiology reports, (2) a minimum of two MRI sequences and (3) non-corrupted data, resulting in 221,147 studies (detailed statistics in Extended Data Fig. 2). All studies were pushed to a HIPAA-compliant server. The MRI sequences were then converted to LPS (Left, Posterior, Superior) orientation, and images were rescaled to 256×256 pixels in the X and Y planes, and slice thickness was converted to 4 mm or greater in the Z plane.

MRI volume tokenization

A common processing step when using ViT architectures is splitting full images into smaller image patches, or vision tokens[21]. However, directly applying this patching strategy to 3D MRI volumes results in a prohibitively large number of tokens per sequence. Therefore, we designed a volumetric tokenization strategy that limits the number of tokens per MRI sequence while preserving diagnostic features by splitting each 3D MRI sequence into large $32 \times 32 \times 4$ patches (apply zero padding if needed) and compress each patch into a smaller token. Inspired by latent diffusion models[20], we train a VQ-VAE[61] to compress each large patch. The VQ-VAE consists of a 3D convolutional neural network (CNN) encoder (f), a quantization layer with a codebook of size 8,192 and a 3D CNN decoder. The encoder downsampled each patch to an $8 \times 8 \times 2$ volume with 2 feature dimensions, resulting in a compact embedding vector $\mathbf{z}_e \in \mathbb{R}^{256}$ (that is, $8 \times 8 \times 2 \times 2$) that serves as the input to the vision model. The codebook size was chosen to balance reconstruction quality and computational efficiency (Extended Data Fig. 4).

We favour using vector quantization via a discrete codebook because anatomic structures and pathologic features are

often shared across patients and pathologies. We demonstrate in Extended Data Fig. 4 that normal structures and radiographic diagnoses have similar embeddings both within and across patients. To ensure robustness across different imaging planes (for example, axial, coronal and sagittal), we apply a random permutation of the image axes during VQ-VAE training. This ensures that the 3D CNN encoder remains invariant to the imaging orientation, which is essential because clinical MRIs in the UM-220K dataset include multiple orientations. High-quality reconstructions across various imaging planes, spatial orientations and permutations are shown in Extended Data Fig. 4, and detailed training algorithm is shown in Supplementary Fig. 1. Volume tokenization effectively mitigates the challenge of handling large token counts in 3D MRI sequences while maintaining essential diagnostic information, enabling scalable and flexible vision-language model training.

MRI-report summarization

High-quality natural language annotations improve CLIP training[62]. Uncurated, clinical radiology reports can contain extraneous information, and non-diagnostic patterns in radiology reports (for example, personal writing styles) can lead to data leakage and bias. To minimize data bias and improve target quality during CLIP training, we used HIPAA-aligned GPT-3.5 to summarize reports of UM-220K in preparation for CLIP training[63]. GPT-3.5-turbo is known to provide high-quality, clinical-grade report summarizations for medical imaging[64,65]. Our report summarization aimed to (1) remove report text that may improve retrieval performance without improving classification performance, (2) homogenize reports and (3) minimize data leakage, bias or learning spurious correlations between the MRI-report pairs. The summarization criteria as well as the detailed GPT-3.5-turbo prompts are in Supplementary Fig. 3. MRI-report summarization leads to better text representations (Extended Data Fig. 3f) and substantially improved downstream classification performance with results shown in Extended Data Fig. 5d. In Extended Data Fig. 3d, we show that Prima, though trained on GPT-3.5 summarization, is aligned with summarization from neuroradiology experts.

MRI labelling with LLMs

We aimed to assign diagnostic labels to each MRI study in UM-220K using the clinical radiology reports. Language models have been used extensively to automate data annotation and radiology report labelling[64–67]. We selected 52 labels that spanned the full neurological disease spectrum to ensure diverse and clinically important predictive tasks. We detail the annotation process in Supplementary Text 1.8. To assess the quality of the automated, 'silver standard' annotations, we compared them to 'gold standard' annotations from an expert in neuroradiology (A.-K.M.) across a diverse subset of diagnostic classes, and results are shown in Extended Data Fig. 3b.

Hierarchical multimodal transformers

Following MRI volume tokenization, a vision model was trained to learn representations of MRI sequences and studies. We used a two-level hierarchical ViT (hViT). The sequence ViT (ViT_{seq}) was used to encode MRI sequence features, and the study ViT (ViT_{st}) was used to aggregate the sequence features and produce representations of full MRI studies.

We define each MRI study as $M = (stn, \{s_1, s_2, \ldots, s_m\})$, where stn is the study description (for example, 'MRI BRAIN WITH AND WITHOUT CONTRAST') and each s_i is a tokenized sequence in the study. Each tokenized sequence $s_i = (sn_i, Z_i)$, where sn_i is the sequence name (for example, 'AX_T1'), and $Z_i = \{z_i^1, z_i^2, \ldots, z_i^{n_i}\}$ is the set of VQ-VAE-encoded tokens in the sequence, each concatenated with a 30-dimensional sinusoidal positional embedding[68] based on its 3D coordinates within the sequence and a 3D one-hot vector indicating the sequence's original orientation (axial/sagittal/coronal). To improve efficiency and reduce memory requirement, we used pixel intensity filtering to remove the background tokens, and we define the filtering process as F, such that $F(Z_i) \subset Z_i$ is the post-filtering token subset. See Extended Data Fig. 1 for a schematic summary.

Similar to how radiologists interpret voxel intensities differently depending on the MRI sequence, we included the tokenized sequence name as part of the input to the sequence ViT. We use a sequence name encoder E_{sn}, a three-layer character-level transformer, to encode sequence names. We pre-trained E_{sn} with CLIP objective between $E_{sn}(SN_i)$ and $V(F(Z_i))$, where V is ViT model. Visualizations of the learned sequence name embeddings can be found in Extended Data Fig. 3c.

The input sequence to ViT_{seq} for each sequence s_i contains 3 parts: 20 register tokens (trainable parameters), $E_{sn}(SN_i)$ and $F(Z_i)$. The three parts are concatenated and input into ViT_{seq}. The encoded vector for each sequence ($r_i = \text{ViT}_{\text{seq}}(s_i)$) is a 1,024-dimensional vector obtained by the final layer output over the 20 registers concatenated together along feature dimensions then projected through a linear layer.

The input sequence to ViT_{st} for study M also contains 3 parts: 10 register tokens, $E_{stn}(STN)$ and $\{P(r_1), P(r_2), \ldots, P(r_m)\}$, where E_{stn} is the study name encoder (same architecture as E_{sn} but not pre-trained) and P is a linear projection layer. The three parts are concatenated together and fed into ViT_{st}. The encoded vector for the entire study M is the final layer output over the 10 register tokens concatenated along feature dimension to form a single 10,240-dimensional feature vector (1,024 dimensions × 10 registers).

We provide detailed specs of ViT_{seq} and ViT_{st} (layers/dimensions) in Supplementary Text 1.1.

Training objective on MRI-report pairs

We train hViT via a CLIP objective[3] between the hViT representation and the representations of the corresponding summarized text reports. The text encoding model, G, is a GPT-2 model pre-trained on radiology reports using an autoregressive next-word prediction objective. We found that pre-training the text encoder on the radiology report corpus improved VLM training efficiency (Extended Data Fig. 5f). For a batch of k MRI study-report pairs, $B = \{(M_1, R_1), (M_2, R_2), \ldots, (M_k, R_k)\}$, where $R_1, R_2, \ldots R_k$ are summarized reports, the CLIP objective is as follows:

$$v_i^M = P_M(\text{hViT}(M_i)), i \in [1, \ldots, k] \tag{1}$$

$$v_i^R = P_R(G(R_i)) \in [1, \ldots, k] \tag{2}$$

$$L_{\text{CLIP}_M} = \frac{1}{k}\sum_{i=1}^{k} -\log \frac{\exp(\text{sim}(v_i^M, v_i^R) \times \exp(\tau))}{\sum_{j=1}^{k} \exp(\text{sim}(v_i^M, v_j^R) \times \exp(\tau))} \tag{3}$$

$$L_{\text{CLIP}_R} = \frac{1}{k}\sum_{i=1}^{k} -\log \frac{\exp(\text{sim}(v_i^M, v_i^R) \times \exp(\tau))}{\sum_{j=1}^{k} \exp(\text{sim}(v_j^M, v_i^R) \times \exp(\tau))} \tag{4}$$

$$L_{\text{CLIP}} = L_{\text{CLIP}_M} + L_{\text{CLIP}_R} \tag{5}$$

Here sim is the cosine similarity, $\text{sim}(x, y) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$, and P_R and P_M are linear projection layers with an output dimension of 128. The temperature parameter, τ, was initialized to 0.07 and updated during training. We found that trainable temperature substantially improved optimization and overall performance. We also used various augmentations on both the reports and MRI images. See Supplementary Text 1.2 for augmentation details.

Self-supervised patient discrimination objective

In addition to the CLIP objective, we added a self-supervised patient discrimination objective that leverages the hierarchical structure of MRI studies. All sequences from an MRI study are of the same patient; therefore, neuroanatomic features are shared across MRI sequences from that patient. We developed a patient discrimination objective

that will enforce that sequences of the same patient will have similar representations from ViT_{seq}. Let s_i^j denote the ViT_{seq} representation for the jth sequence in M_i and let n_i denote the number of sequences in M_i. The patient sequence discrimination objective is as follows:

$$u_i^j = P_{\text{patdis}}(s_i^j) \tag{6}$$

$$L_{\text{patdis}} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i} -\log\frac{\sum_{j'=1}^{n_i}\exp(\text{sim}(u_i^j, u_i^{j'})/\tau_p)}{\sum_{i'=1}^{k}\sum_{j'=1}^{n_{i'}}\exp(\text{sim}(u_i^j, u_{i'}^{j'})/\tau_p)} \tag{7}$$

where P_{patdis} is a 2-layer MLP projection layer that maps ViT_{seq} outputs to the patient discrimination embedding space, and τ_p is a trainable temperature parameter initialized at 0.1. The numerator of equation (7) are the sequence similarities within the same study/patient. The denominator is the sequence similarities between all pairs of sequence representations. The final training objective for Prima is

$$L_{\text{train}} = L_{\text{CLIP}} + \lambda L_{\text{patdis}} \tag{8}$$

where λ is a hyperparameter. In our experiments, we set λ to 0.03. The full training algorithm is shown in Supplementary Fig. 2.

Evaluation metrics for vision-language alignment

During CLIP training, we used Top-1 and Top-5 retrieval accuracies to monitor vision-language alignment. These metrics evaluate the accuracy of Prima for matching MRI studies with their corresponding summarized radiology report. Top-1 and Top-5 retrieval accuracy metrics were defined as

$$\text{Top}-1\,\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left(\underset{j}{\text{argmax}}\,\text{sim}(i,j) = i\right) \tag{9}$$

$$\text{Top}-5\,\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(i \in \text{Top}-5(\text{sim}(i,j))) \tag{10}$$

where $N = 254$ and is a held-out validation set from UM-220K. To evaluate vision-language alignment on the clinical testing set, we randomly divide the testing set into 294 groups of 100, and then report Top-1 retrieval accuracy within each group, averaged across all groups.

Evaluation protocol for diagnostic tasks

Radiologic diagnoses are not mutually exclusive and cannot be treated as a multi-class classification task. Instead, we treat it as a multi-binary-label problem and trained a classification head C that takes as input the MRI study embeddings generated by hViT, v^M. C is a 3-layer MLP and outputs an L-dimensional vector, where L is the number of labels:

$$\hat{\boldsymbol{y}} = C(v^M),\ \ \hat{\boldsymbol{y}} = [y_1, y_2, \ldots, y_L],\ \ y_i \in [0,1] \tag{11}$$

Similar to other CLIP models[3], hViT is fixed and C is trained using a positive-weighted binary cross entropy loss:

$$\mathcal{L}_{\text{BCE}}^{\text{multi-label}} = -\frac{1}{L}\sum_{i=1}^{L}[p_i\, y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{12}$$

with p_i as the positive weight for the ith label obtained by $p_i = \frac{\text{num negative } i\text{th label in training set}}{\text{num positive } i\text{th label in training set}}$. The ground truth multi-hot vector, y, is provided by the LLM as described above:

$$\boldsymbol{y} = \text{LLM}(R),\ \ \boldsymbol{y} = [y_1, y_2, \ldots, y_C],\ \ y_i \in \{0,1\} \tag{13}$$

where R is the MRI study report. We selected the checkpoint C with the best performance for each task using a held-out validation set. We found that the above strategy successfully captures label co-occurrence and learns the semantic/differential diagnosis relationships between labels (Extended Data Fig. 9, where diagnoses were ordered using consensus clustering[69]).

Testing patient cohort and sample size calculation

We designed a 1-year diagnostic accuracy study and used the same inclusion criteria as shown in Extended Data Fig. 1. We performed a sample size calculation based on a parallel superiority trial with a binary outcome: differentiating normal versus abnormal MRIs[17], resulting in a minimal sample size of 22,338 MRIs (detailed process in Supplementary Text 1.3). The primary evaluation metric was mean AUROC across the radiologic diagnosis tasks. Testing patient cohort enrolment began on 1 June 2023 and completed on 30 May 2024. A total of 29,431 MRI studies were included, and our required minimal sample size was exceeded.

Model design ablations

We performed several ablation studies to optimize the design choices of the Prima architecture: inclusion of sequence name and study description, flat ViT encoder, 2D CNN encoder, 3D CNN encoder, unsummarized long reports and self-supervised patient discrimination loss. The results are shown in Extended Data Fig. 5d.

To ablate on the inclusion of sequence name and study descriptions, we evaluate the performance of the Prima model with no sequence name information (that is, replace all input sequence name with 'unk') and no study description (study name encoding not included in the input to ViT_{st}). We found a slight drop in performance after removing either sequence names or study descriptions, indicating that the inclusion of sequence name and study description does help with prediction accuracy, but Prima also does not rely heavily on these information.

We detail our implementation of flat ViT, 2D CNN and 3D CNN in Supplementary Text 1.4. Prima outperformed all three ablations, highlighting the importance of the hViT architecture.

We attempted to train Prima with unsummarized long radiology reports rather than summarized short reports. We found that the model trained with long reports suffers from overfitting owing to the model using non-diagnostic features in MRI reports to minimize the CLIP objective. We also trained a version of Prima without patient discrimination loss. As shown in Extended Data Fig. 5e, patient discrimination loss substantially accelerates model convergence. We performed an additional ablation called Simple ViT, where we train a flat ViT architecture with CLIP loss without any additional designs (no patient discrimination loss, sequence/study name or data augmentations). The model overfits during training owing to lack of inductive bias regarding the hierarchical structure of brain MRIs.

Explainable and trustworthy predictions

LIME[45] is a commonly used method to interpret the decision-making of black-box classification or regression models. It generates a score for each input feature indicating its 'contribution' towards each model prediction. A brief description of how the scores are obtained is in Supplementary Text 1.5. To determine the trustworthiness of Prima's predictions (for example, whether it predicts a tumour based on the MRI's tumour regions), we run LIME on Prima's prediction across diagnostic tasks. To isolate token contributions within each sequence in an MRI study, we only perturb tokens from one sequence of the study during LIME. Each input feature is a single 3D volume token, and the corruption process is simply token removal. For each LIME interpretation, 3,000 masked inputs were generated and the contribution score for each volume token was ranked and converted into colour-coded visualizations as shown in Fig. 3. Qualitative evaluation was performed across all diagnostic classes (Extended Data Fig. 8; also see the demo website (prima.mlins.org)). To quantitatively evaluate Prima's selection of the volume tokens within brain tumour regions, we used the BraTS

dataset[40] that includes dense semantic segmentation masks and measured the overlap rate between the top-K LIME volume tokens and the tumour segmentation masks (Extended Data Fig. 7d). Qualitative error analysis was completed for Prima to identify suboptimal performance in specific scenarios, such as uncommon MRI sequences or protocols, as shown in Supplementary Fig. 5.

Referral, acuity and age prediction training

For referral prediction tasks, we follow the same protocol as the diagnosis tasks: for R total referral tasks, we freeze the CLIP-trained sequence and study encoders, and we train a three-layer MLP that takes in the encoder output, and outputs an R-dimensional output where each dimension corresponds to the logit of a referral task (for example, referral to paediatric neurosurgeon). The MLP is trained with positive-weighted binary cross entropy loss. For each task, we take the checkpoint of the MLP with the best performance on the held-out validation set, and save for model testing. For acuity prediction, the CLIP-trained encodings are frozen and we train a three-layer MLP that takes in the encoder output and outputs a 3D vector that corresponds to three levels of acuity: normal, medium and high acuity. The MLP is trained with a categorical cross entropy loss on the training set. We ablated over alternative ordinal-based objectives, such as ordinal metric learning[70] and binary ordinal[71]. The checkpoint with the best validation performance is used for testing. Mappings between the radiologic diagnoses and the referral/acuity classes are in Supplementary Table 3. We follow the same protocol for age prediction as above, but the three-layer MLP outputs a scalar value for regression using an L2 objective.

Prima on public datasets

To further evaluate the generalizability and transferability, we evaluate Prima on several publicly available datasets. We divide the datasets into two groups. Group 1 aims to test generalizability and includes MRI datasets with one or more of the study diagnoses, namely in-domain. Group 2 aims to test transferability and includes MRIs with diagnoses outside of our study diagnoses, namely out-of-domain. These are out-of-domain because CLIP training with the radiology provides no supervision for these tasks. For each dataset in the first group, we directly use Prima and the diagnosis-specific MLP head to predict on MRI studies, and report TPR as the evaluation metric. The datasets that we include in group 1 are BRATS 2021 (adult glioma)[40], NYUMets (metastasis)[42], UCSF-BMSR (metastasis)[41] and Stroke (large vessel stroke)[43]. Group 2 includes datasets for autism spectrum disorder (ABIDE[35]) and dementia (ADNI[36] and OASIS-1[37]). For each task in each dataset, we obtain a single embedding vector for the MRI studies. ABIDE has one sequence per study, and we directly encode ABIDE sequences with ViT_{seq}. Otherwise, we use the full Prima encoder. We then trained a two-layer MLP that outputs a single logit for the task and trained with a binary cross entropy loss. We perform a five-way cross-validation on each task and report average validation performance and standard deviation in Fig. 2. Results were compared with recent baselines for ABIDE and ADNI tasks[18], and for OASIS-1 tasks[72].

VLM zero-shot benchmarking details

We evaluated the performance of several state-of-the-art, pre-trained VLMs as zero-shot baselines on the radiologic diagnosis tasks. We focus on three categories of publicly available pre-trained VLMs: CLIP family (CLIP-base[3], CLIP-large[3], PubMedCLIP[30], BioMedCLIP[31] and Blip-2[73,74]), Llava family (Llava-1.5-7B[75], Llava-Mistral, Llava-Med[76] and Llava-3D[77]) and MedFlamingo[54]. Other relevant neuroimaging models were excluded owing to lack of open-source implementation or publicly released model[59]. We randomly sample a subset of the clinical test set to evaluate the VLMs. For a diagnosis task with p positives in the testing set, we randomly sample $\min(p, 100)$ positive samples and $\min(p, 100)$ negative samples from the testing set to form a balanced subset.

Since the models only support 2D image inputs and each model family has different prediction mechanisms, we devised different methods to best extract zero-shot prediction scores to calculate area under the ROC curve (AUC) metric for each model family. We detail the methods in Supplementary Text 1.6.

We report the zero-shot performance of each model together with Prima's zero-shot performance on the same balanced subsets in Fig. 2b, with task-wise comparison in Extended Data Fig. 5a. We include best-performing prompts for each model for each task in Supplementary Table 6. We could not report performance of OpenAI's GPT VLMs because they refuse to give diagnostic predictions for MRI images.

MLP probing of 2D CLIP models

In addition to zero-shot evaluation, we compared the results of MLP probing of the 2D CLIP models, from 'CLIP family' above, over the UM-220K dataset to Prima's MLP-probing performance. For each CLIP model, the 2D vision encoder was used to encode each MRI slice. We then performed average pooling over the encodings to get the final study representation. Then, we fit the same 3-layer MLP classification head over UM-220K training data in the same manner as Prima. We report classification performance on testing data. The results are shown in Extended Data Fig. 5b. The 2D CLIP models show improved performance when trained on UM-220K dataset compared with zero-shot performance, indicating the value of health system-scale training. Baseline CLIP models underperform Prima owing to the hierarchical ViT design and 3D modelling.

Prima with clinical context

Radiologists use clinical context to better interpret MRI studies. Similarly, we hypothesized that Prima would benefit from having the clinical context when predicting radiologic diagnosis. To perform this evaluation, we first use GPT-4o-mini to parse the clinical history and indications from each patient's radiology report. We then use OpenAI's text-embedding-3-small model to obtain 1,536-dimensional embeddings for each patient's clinical context. Then, we concatenate each clinical embedding with the 10,240-dimensional MRI study embedding from Prima for each study, forming a 11,776-dimensional embedding for each study. We then performed the MLP probing over 52 diagnostic tasks and reported performance in Fig. 2a.

Bias and fairness criteria

We used odds ratios to measure relative disparity between a reference group and a sensitive group to evaluate potential systemic biases. A two-sided Fisher exact test was used to compute P values with multiple hypothesis correction. Null hypothesis was that the odds of longer turnaround time was not higher in the sensitive group compared with the reference group.

To evaluate the fairness of Prima, we used the equalized odds framework combined with intersectional and diagnostic subgroup analysis[48,78]. Algorithmic fairness under this framework uses a separation criterion defined over a sensitive attribute $A \in \{a, b\}$, a classifier $\hat{Y}$ and target variable Y, such that $\hat{Y} \perp A|Y$. The conditional independence states that the odds of predicting 'positive' or 'negative' are independent of the sensitive attribute conditioned on the target variable. The separation criteria for TPR, $\mathbb{P}\{\hat{Y} = 1|Y = 1\}$, and false positive rate (FPR), $\mathbb{P}\{\hat{Y} = 0|Y = 0\}$, are respectively

$$\mathbb{P}\{\hat{Y} = 1|Y = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1|Y = 1, A = b\} \tag{14}$$

$$\mathbb{P}\{\hat{Y} = 0|Y = 0, A = a\} = \mathbb{P}\{\hat{Y} = 0|Y = 0, A = b\} \tag{15}$$

Satisfying both criteria demonstrates equalized odds. Intersectional analysis includes combining sensitive attributes, such that equalized odds are computed over the intersection of patient sets defined by

individual attributes, for example, black females or Asians living in southwest Michigan. Finally, we evaluate these metrics for specific diagnoses, such as neoplastic lesions, ischaemic strokes and paediatric diagnoses. Complete algorithmic fairness experiments can be found in Extended Data Fig. 10 and Supplementary Fig. 4. As Prima is well suited for AI screening and decision support tool, TPR is the most important metric to evaluate and is called equalized opportunity as a fairness metric[79]; missing diagnoses are worse than predicting them when the patient is normal. Equalized opportunity evaluates if a classifier's TPR is equalized across sensitive attributes for model fairness and non-discrimination. TPR was calculated for a sensitive subgroup, G, as

$$\mathrm{TPR}_G = \frac{\sum_{i\in G}\mathbb{1}(y_i = 1 \wedge \hat{y}_i = 1)}{\sum_{i\in G}\mathbb{1}(y_i = 1)}, \tag{16}$$

and for the full study population as

$$\mathrm{TPR}_{\mathrm{Pop.}} = \frac{\sum_{i=1}^{N}\mathbb{1}(y_i = 1 \wedge \hat{y}_i = 1)}{\sum_{i=1}^{N}\mathbb{1}(y_i = 1)}. \tag{17}$$

Deviations from equalized opportunity were measured using TPR disparity:

$$\text{TPR disparity} = \mathrm{TPR}_G - \mathrm{TPR}_{\mathrm{Pop.}} \in [-1, 1] \tag{18}$$

with larger absolute values, |TPR disparity|, reflecting greater algorithmic bias. Examples of sensitive attributes that defined subgroups in our study were population density, geography, scheduling, race, ethnicity, sex, age and so on.

Equalized opportunity experiment design

For each fairness experiment, we used bootstrap sampling to estimate the sensitive subgroup TPR and the population TPR[79]. We randomly sample 200 patients with replacement from a subgroup and 200 diagnosis-matched patients from the study population with replacement. We then computed the TPR values for the bootstrap sample subgroup and the population. The process was repeated for 20 iterations. We performed P-value testing using a one-sided, non-parametric Mann–Whitney U statistical test. Our null hypothesis was that the subgroup TPR distribution was not less than the population TPR distribution. Lower TPR rates for the sensitive subgroups represent bias that causes harm through more false negatives.

Computational hardware and software

See Supplementary Text 1.7.

Ethics and inclusion statement

Our research was approved by the University of Michigan Institutional Review Board (HUM00229133). All MRI data were acquired under secondary data usage. The methods were carried out in accordance with the IRB's guidelines, regulations and policies. All human subjects that met inclusion criteria as stated above were included in the study.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Prima model parameters will be publicly available for investigational use only under an MIT licence. Institutional Review Board approval was obtained from the University of Michigan for MRI data collection. Restrictions apply to the availability of raw patient MRI imaging data, which were used with institutional permission through IRB approval for the current study, and are thus not publicly available. All data sharing between medical centres is regulated through data use agreements with the study authors. A similar data-sharing protocol may be established for interested investigators. Please contact the corresponding author (T.H.) for any requests for data sharing, and a response will be made within 2 weeks. All requests will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic and investigational purposes.

Code availability

All code was implemented in Python (version 3.9) using PyTorch (2.3.1) and Transformers (4.37.0) as the primary machine learning framework. All code and scripts to reproduce the training and inference of Prima are available via GitHub at MLNeurosurg/Prima under an MIT licence.

References

1. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
2. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
3. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.), Vol. 139 of *Proceedings of Machine Learning Research* 8748–8763 (PMLR, 2021).
4. Ramesh, A. et al. Zero-shot text-to-image generation. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.), Vol. 139 of *Proceedings of Machine Learning Research* 8821–8831 (PMLR, 2021).
5. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, Vol. 35 (eds Koyejo, S. et al.) 23716–23736 (Curran Associates, 2022).
6. Dreisbach, J. N. & Lukin, R. Where have all the neuroradiologists gone? *AJNR Am. J. Neuroradiol.* **22**, 1636–1638 (2001).
7. Rula, E. Y. Radiology workforce shortage and growing demand: something has to give. https://www.acr.org/Practice-Management-Quality-Informatics/ACR-Bulletin/Articles/July-2024/Radiology-Workforce-Shortage-and-Growing-Demand-Something-Has-to-Give (2024).
8. Christensen, E. W. et al. Association of state share of nonphysician practitioners with diagnostic imaging ordering among emergency department visits for medicare beneficiaries. *JAMA Netw. Open* **5**, e2241297 (2022).
9. Fawzy, N. A. et al. Incidence and factors associated with burnout in radiologists: a systematic review. *Eur. J. Radiol. Open* **11**, 100530 (2023).
10. Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M. & Kim, J. Long radiology workdays reduce detection and accommodation accuracy. *J. Am. Coll. Radiol.* **7**, 698–704 (2010).
11. Ivanovic, V. et al. Neuroradiology diagnostic errors at a tertiary academic centre: effect of participation in tumour boards and physician experience. *Clin. Radiol.* **77**, 607–612 (2022).
12. Ivanovic, V. et al. Factors associated with neuroradiology diagnostic errors at a large tertiary-care academic medical center: a case-control study. *Am. J. Roentgenol.* **221**, 355–362 (2023).
13. O'Neill, T. J. et al. Active reprioritization of the reading worklist using artificial intelligence has a beneficial effect on the turnaround time for interpretation of head CT with intracranial hemorrhage. *Radiol. Artif. Intell.* **3**, e200024 (2021).
14. Shin, H. J., Han, K., Ryu, L. & Kim, E.-K. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. *npj Digit. Med.* **6**, 82 (2023).

15. Alexander, R. et al. Mandating limits on workload, duty, and speed in radiology. *Radiology* **304**, 274–282 (2022).
16. DeBenedectis, C. M. et al. Health care disparities in radiology—a review of the current literature. *J. Am. Coll. Radiol.* **19**, 101–111 (2022).
17. Gauriau, R. et al. A deep learning-based model for detecting abnormalities on brain MR images for triaging: preliminary results from a multisite experience. *Radiol. Artif. Intell.* **3**, e200184 (2021).
18. Barbano, C. A., Brunello, M., Dufumier, B. & Grangetto, M. Anatomical foundation models for brain MRIs. *Pattern Recognition Letters* **199**, 178–184 (2026).
19. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/pdf/2303.08774 (2023).
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10684–10695 (2022).
21. Dosovitskiy, A. et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations* (OpenReview.net, 2021).
22. Darcet, T., Oquab, M., Mairal, J. & Bojanowski, P. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations* (eds Kim, B. et al.) 2632–2652 (2024).
23. Zhang, K. et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **181**, 1423–1433.e11 (2020).
24. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
25. Bannur, S. et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 15016–15027 (2023).
26. Wang, Y. et al. Enhancing vision-language models for medical imaging: bridging the 3D gap with innovative slice selection. *Neural Inf. Process. Syst.* **37**, 99947–99964 (2024).
27. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
28. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
29. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Proc. 37th International Conference on Neural Information Processing Systems* 34892–34916 (2023).
30. Eslami, S., Meinel, C. & De Melo, G. PubMedCLIP: how much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023* (eds Vlachos, A. & Augenstein, I.) 1151–1163 (ACL, 2023).
31. Zhang, S. et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. Preprint at https://arxiv.org/abs/2303.00915 (2023).
32. Moor, M. et al. Med-Flamingo: a multimodal medical few-shot learner. In *Proc. 3rd Machine Learning for Health Symposium* (eds Hegselmann, S. et al.) 353–367 (PMLR, 2023).
33. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
34. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proc. 34th International Conference on Machine Learning* 1321– 1330 (PMLR, 2017).
35. Di Martino, A. et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
36. Petersen, R. C. et al. Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. *Neurology* **74**, 201–209 (2010).
37. Marcus, D. S. et al. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**, 1498–1507 (2007).
38. Lee, J. et al. Deep learning-based brain age prediction in normal aging and dementia. *Nat. Aging* **2**, 412–424 (2022).
39. Bashyam, V. M. et al. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).
40. Baid, U. et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. Preprint at https://arxiv.org/abs/2107.02314 (2021).
41. Rudie, J. D. et al. The University of California San Francisco Brain Metastases Stereotactic Radiosurgery (UCSF-BMSR) MRI dataset. *Radiol. Artif. Intell.* **6**, e230126 (2024).
42. Oermann, E. et al. Longitudinal deep neural networks for assessing metastatic brain cancer on a massive open benchmark. *Nat. Commun.* **15**, 8170 (2024).
43. Liu, C.-F. et al. A large public dataset of annotated clinical MRIs and metadata of patients with acute stroke. *Sci. Data* **10**, 548 (2023).
44. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
45. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?' Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Krishnapuram, B. et al.) 1135–1144 (2016).
46. Smith, J. S. et al. Role of extent of resection in the long-term outcome of low-grade hemispheric gliomas. *J. Clin. Oncol.* **26**, 1338–1345 (2008).
47. Waite, S., Scott, J. & Colombo, D. Narrowing the gap: imaging disparities in radiology. *Radiology* **299**, 27–35 (2021).
48. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, 2023).
49. Rajpurkar, P. & Topol, E. J. A clinical certification pathway for generalist medical AI systems. *Lancet* **405**, 20 (2025).
50. Ivanovic, V. et al. Impact of shift volume on neuroradiology diagnostic errors at a large tertiary academic center. *Acad. Radiol.* **30**, 1584–1588 (2023).
51. Babiarz, L. S. & Yousem, D. M. Quality control in neuroradiology: discrepancies in image interpretation among academic neuroradiologists. *AJNR Am. J. Neuroradiol.* **33**, 37–42 (2012).
52. Wu, M. Z., McInnes, M. D. F., Macdonald, D. B., Kielar, A. Z. & Duigenan, S. CT in adults: systematic review and meta-analysis of interpretation discrepancy rates. *Radiology* **270**, 717–735 (2014).
53. Azizi, S. et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779 (2023).
54. Moor, M. et al. Med-Flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)* 353–367 (PMLR, 2023).
55. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
56. Blankemeier, L. et al. Merlin: a vision language foundation model for 3D computed tomography. Preprint at https://www.researchsquare.com/article/rs-4546309/v1 (2024).
57. Elliott, L. T. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).
58. Kickingereder, P. et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* **20**, 728–740 (2019).
59. Wood, D. A. et al. A self-supervised text-vision framework for automated brain abnormality detection. Preprint at https://arxiv.org/abs/2405.02782 (2024).

60. Ghosh, S., Poynton, C. B., Visweswaran, S. & Batmanghelich, K. Mammo-CLIP: a vision language foundation model to enhance data efficiency and robustness in mammography. In *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention* 632–642 (Springer, 2024).
61. van den Oord, A., Vinyals, O. & Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, Vol. 30 (eds Guyon, I. et al.) (2017).
62. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at https://arxiv.org/pdf/2204.06125 (2022).
63. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
64. Chien, A. et al. AI-assisted summarization of radiologic reports: evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. *AJNR Am. J. Neuroradiol.* **45**, 244–248 (2024).
65. Ranjit, M., Ganapathy, G., Manuel, R. & Ganu, T. Retrieval augmented chest X-ray report generation using OpenAI GPT models. In *Proc. Machine Learning for Healthcare Conference* (eds Deshpande, K. et al.) 650–666 (PMLR, 2023).
66. Adams, L. C. et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* **307**, e230725 (2023).
67. Titano, J. J. et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* **24**, 1337–1341 (2018).
68. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) Vol. 30 (Curran Associates, Inc., 2017).
69. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
70. Kondepudi, A. et al. Foundation models for fast, label-free detection of glioma infiltration. *Nature* **637**, 439–445 (2025).
71. Cheng, J., Wang, Z. & Pollastri, G. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* 1279–1284 (2008).
72. Saratxaga, C. L. et al. MRI deep learning-based solution for Alzheimer's disease prediction. *J. Pers. Med.* **11**, 902 (2021).
73. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. International Conference on Machine Learning* 19730–19742 (PMLR, 2023).
74. Chen, Q. & Hong, Y. MedBLIP: bootstrapping language-image pre-training from 3D medical images and texts. In *Proc. Asian Conference on Computer Vision* (eds Cho, M. et al.) 2404–2420 (2024).
75. Liu, H., Li, C., Li, Y. & Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition* 26296–26306 (2024).
76. Li, C. et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, Vol. 36 (eds Oh, A. et al.) 28541–28564 (Curran Associates, Inc., 2023).
77. Zhu, C., Wang, T., Zhang, W., Pang, J. & Liu, X. LLaVA-3D: a simple yet effective pathway to empowering LMMs with 3D-awareness. In *Proc. IEEE/CVF International Conference on Computer Vision* 4295–4305 (2025).
78. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, Vol. 29 (eds Lee, D. et al.) (2016).
79. Vaidya, A. et al. Demographic bias in misdiagnosis by computational pathology models. *Nat. Med.* **30**, 1174–1190 (2024).

Acknowledgements

We would like to thank K. Eddy (University of Michigan), G. Laderach (University of Michigan), B. Palen (University of Michigan) and M. Bhalli (University of Michigan) for providing technical support, D. Hanauer (University of Michigan) for support with the University of Michigan Electronic Medical Record Search Engine (EMERSE) and A. Rosenzweig (University of Michigan) for scientific guidance. This work was supported by the following National Institutes of Health (NIH) funding sources: K12NS080223 (T.H.). This work was supported by the Chan Zuckerberg Initiative (CZI), Frankel Institute for Heart and Brain Health (T.H.), the Mark Trauner Brain Research Fund, the Zenkel Family Foundation (T.H.), Ian's Friends Foundation (T.H.) and the UM Precision Health Investigators Awards grant programme (T.H.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This research was also supported, in part, through computational resources and services provided by Advanced Research Computing, a division of Information and Technology Services at the University of Michigan.

Author contributions

Y.L., S.H., A.C. and T.H. conceived and designed the experiments, performed the experiments, analysed the data, contributed materials and analysis tools, and wrote the paper. S.B., R.G., S.L., A.-K.M., A.R., C.Z., A.K., C.J., X.H. and R.S.J. performed the experiments, analysed the data, and contributed materials and analysis tools. V.N., A.S., D.K., B.A., V.G., A.P. and H.L. conceived and designed the experiments.

Competing interests

The authors declare no competing interests.

Additional information

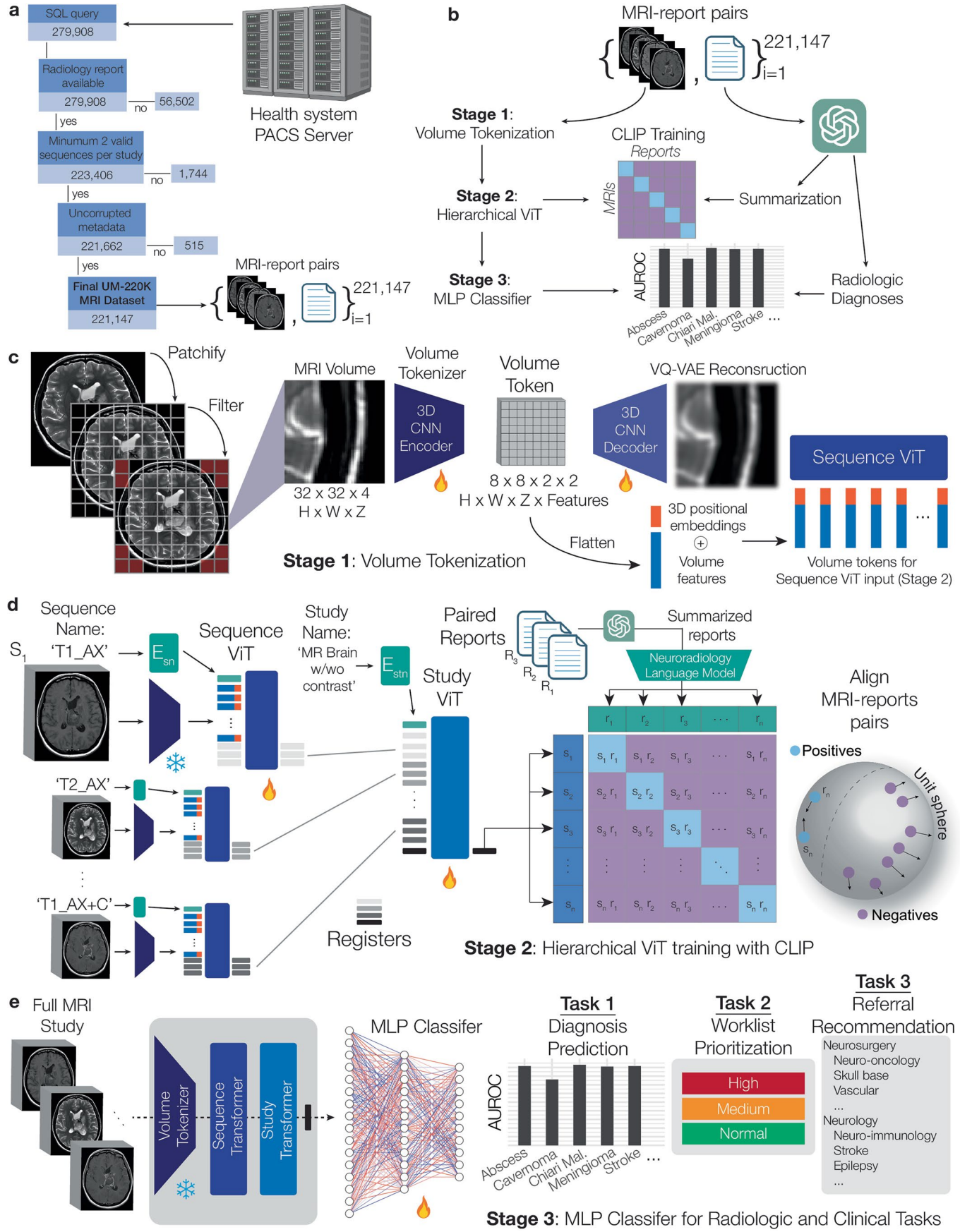
Extended data is available for this paper at https://doi.org/10.1038/s41551-025-01608-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-025-01608-0.

Correspondence and requests for materials should be addressed to Todd Hollon.

Peer review information *Nature Biomedical Engineering* thanks Eric Oermann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41551-025-01608-0

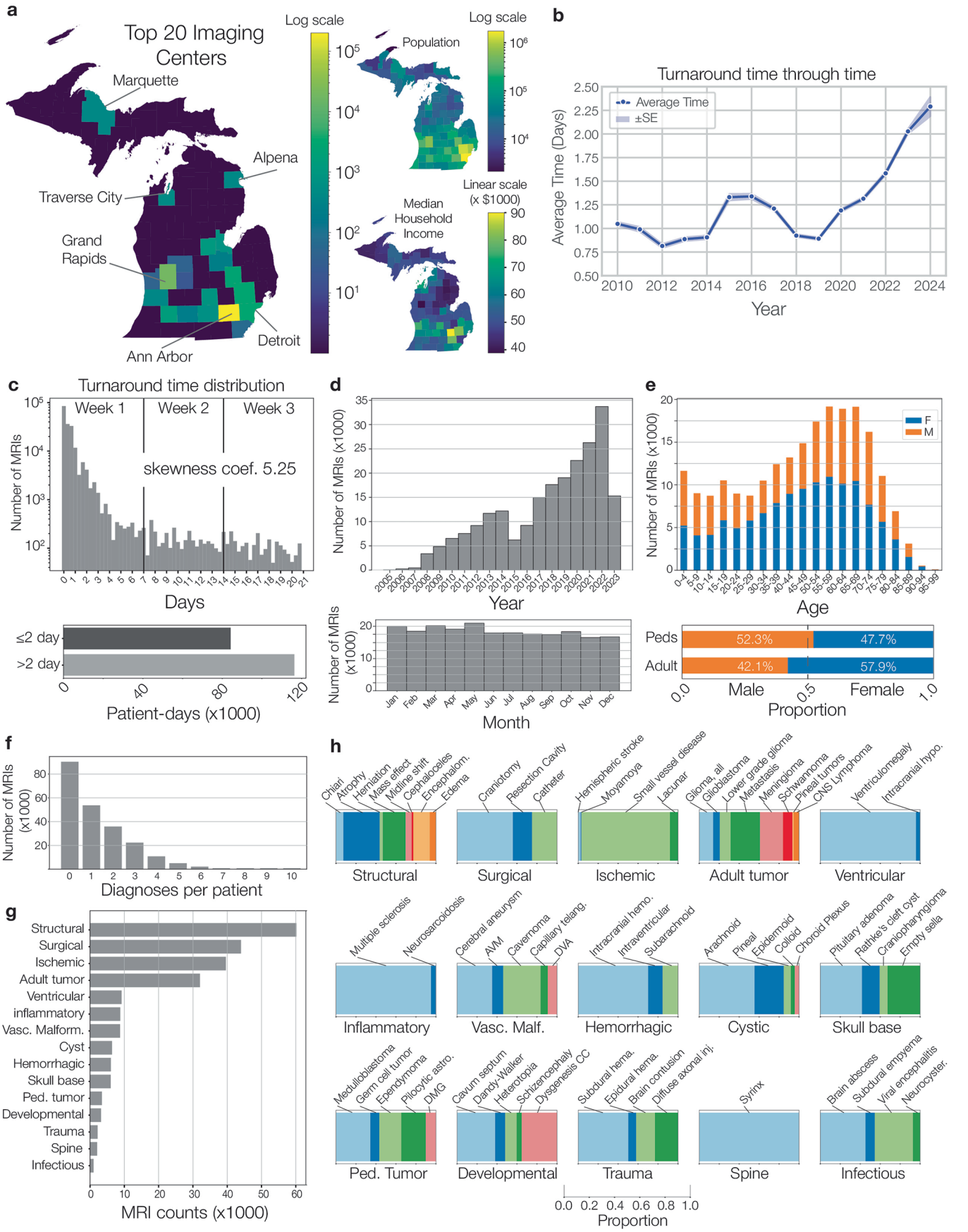


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 1 | Expanded Workflow and Prima Architecture. a, Our health system Sectra PACS server was queried for all cranial MRIs. We then filtered these MRIs based on the availability of an associated radiology report and having a minimum of 2 series per study. We then ensured that all metadata was present, resulting in a total of 221,147 UM-220K dataset. **b**, Overview of the stages of training Prima on UM-220K, which includes volume tokenization, hierarchical ViT training with CLIP objective function, and transfer learning to predict radiologic diagnoses. An LLM provides radiology report summarization and diagnostic labels for reliable, accurate, and scalable vision-language modeling. **c**, Volume tokenization stage involves dividing each MRI volume into smaller subvolume patches of shape 32x32x4, removing background tokens, and encoding each subvolume using a VQ-VAE encoder. The latent VQ-VAE tokens are then passed forward to the sequence ViT with the concatenated positional encodings. **d**, The hierarchical ViT is trained using a CLIP objective on frozen volume token features. The sequence ViT is a multimodal transformer that takes as input both the volume tokens and the embedded free-text sequence description. The series registers are passed forward to the study ViT that outputs a single representation for the full MRI study. The paired reports are summarized and passed through a pre-trained neuroradiology model to align the MRI study and the paired report. **e**, A transfer learning strategy is used such that the volume tokenizer, sequence and study transformers are frozen, and an MLP is trained on the learned study features for radiologic and clinical task prediction. Illustrations in **a**–**e** created with BioRender.com.

https://doi.org/10.1038/s41551-025-01608-0

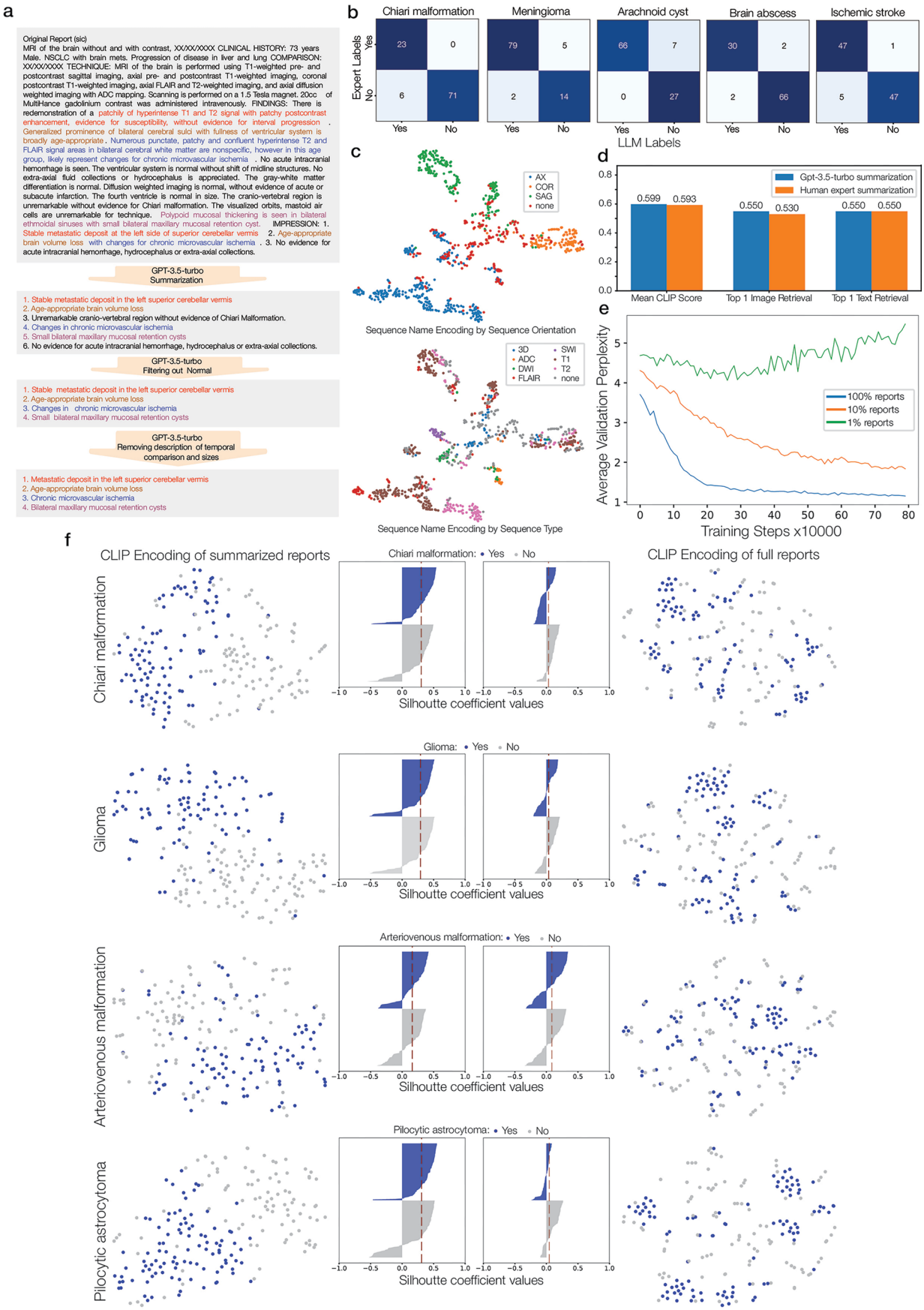


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 2 | Overview of UM-220K Dataset. Descriptive details of the UM-220K dataset are shown here. **a**, The majority of MRI data is collected from patient's with permanent residence in the state of Michigan. A geospatial map shows the counties with the top 20 imaging centers across the state. The top centers are located in populous regions and regions with higher median income based on census data (www.census.gov). The Upper Peninsula and Northern Michigan are lower resource settings, rural areas, and most susceptible to experiencing longer turnaround times (Fig. 4). **b**, Average turnaround time through time. We observed a steady year-to-year increase in turnaround time since 2019. This increase correlates with increasing MRI demand and imaging volumes at our health system. **c**, The distribution of turnaround times. The distribution shows a severe right skew, with a Fisher-Pearson coefficient of skewness of 5.25. The majority of turnaround time measured in patient-days is in the right tail distribution greater than 2 days. These results prompted us to target turnaround time as a metric of algorithmic fairness. **d**, Distribution of MRI counts through time and divided by month of year. We observed a consistent increase in the number of MRIs/year. **e**, Age and sex distribution of the UM-220K. **f**, Distribution of patients by the number of diagnoses per patient, including patients with no diagnoses. **g**, Distribution of diagnostic categories and (**h**) the granular radiology diagnoses for each category. The aims was to have a broad set of diagnostic categories that spanned the full diagnostic spectra and to include clinically meaningful and actionable diagnostic classes.

https://doi.org/10.1038/s41551-025-01608-0

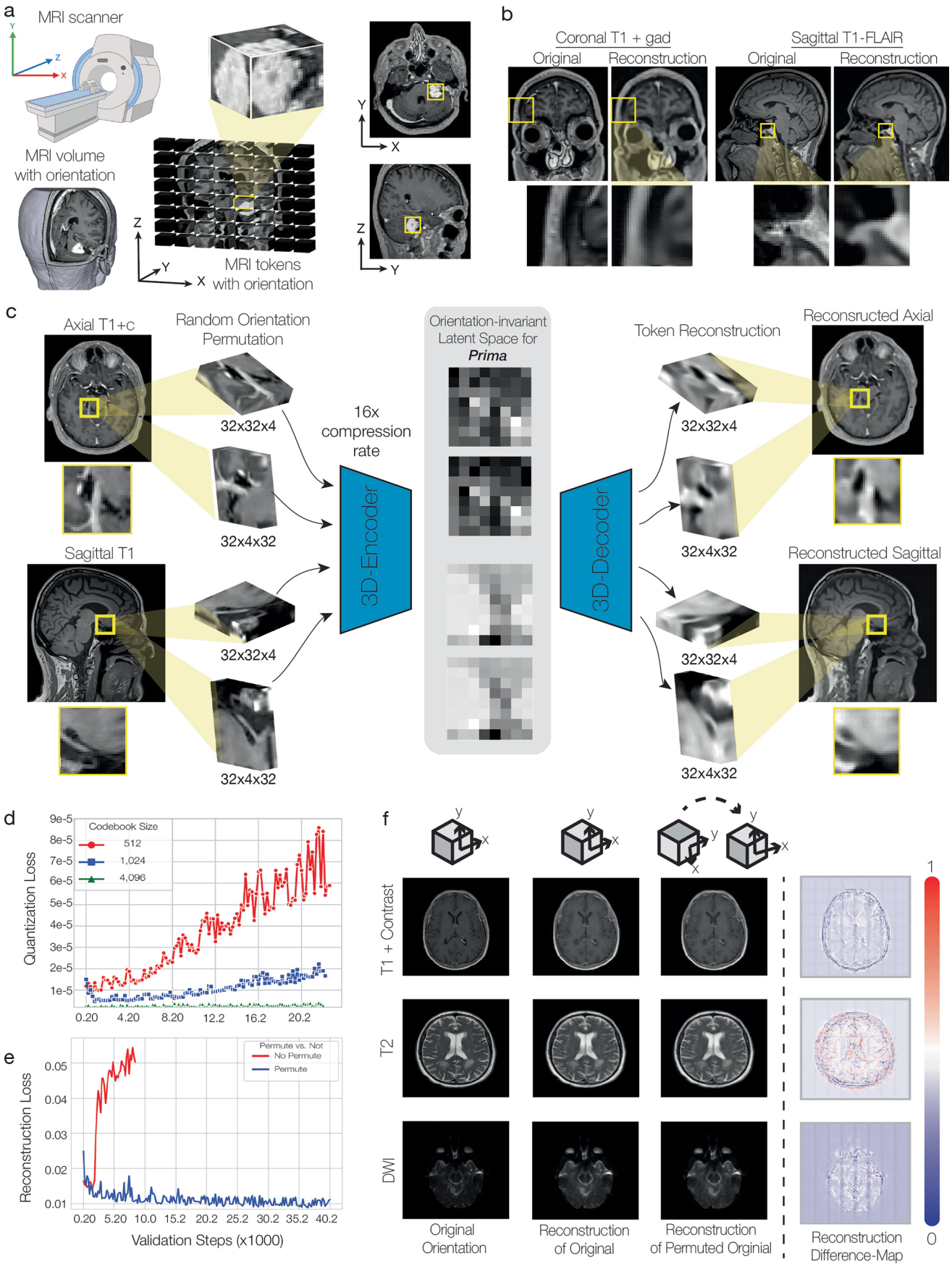


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 3 | LLM annotations and neuroradiology language models. **a**, Example of an original MRI report with highlighting of the major findings in the MRI study. GPT-3.5-turbo is then prompted in stages to summarized the report to achieve an itemized report summary of positive findings. Additional details regarding prompting are in Supplementary-Data-Fig. 2. **b**, The report classification performance of GPT-4 for 5 diagnoses from the different diagnostic categories: structural, adult tumor, cystic lesions, infectious, and vascular ischemic. Prompting details of GPT-4 classification are in Supplementary-Data-Fig 3. **c**, tSNE visualizations of the sequence name encoding after training the sequence name encoder, E_{sn}, using a CLIP objective. The model correctly encodes sequence names based on both imaging planes and sequence type. This effectively prompts ViT_{seq} for better MRI sequence feature extraction. **d**, Comparison on Prima CLIP-score and retrieval performance between GPT-3.5-turbo summarized reports and human expert summarized reports on 100 prospective studies. We observe no statistically significant difference between the two summarizations on CLIP scores (P=0.62, paired two-sided t-test) or retrieval metrics (P > 0.80, McNemar's test), indicating that Prima, although trained on GPT-3.5-turbo summarized reports, is aligned with radiologist summarization. **e**, Line plot shows the average validation perplexity during training of the neuroradiology language model given different percentages of the report data. Neuroradiology language model was trained using next-word prediction and benefits from increased training data, approaching the lower perplexity bound of 1 with 100% of the MRI reports. **f**, tSNE plots of the CLIP encoded summarized reports versus full reports. The summarized reports show better label-conditional clustering compared to the full reports. These results demonstrate the challenge of using full reports for CLIP objective, which contain extraneous or non-informative textual details that can degrade visual representation learning. The center panels are Silhouette plots to quantify cluster quality. The red dotted line is the average Silhouette coefficient. Larger values represent better clustering results.

https://doi.org/10.1038/s41551-025-01608-0

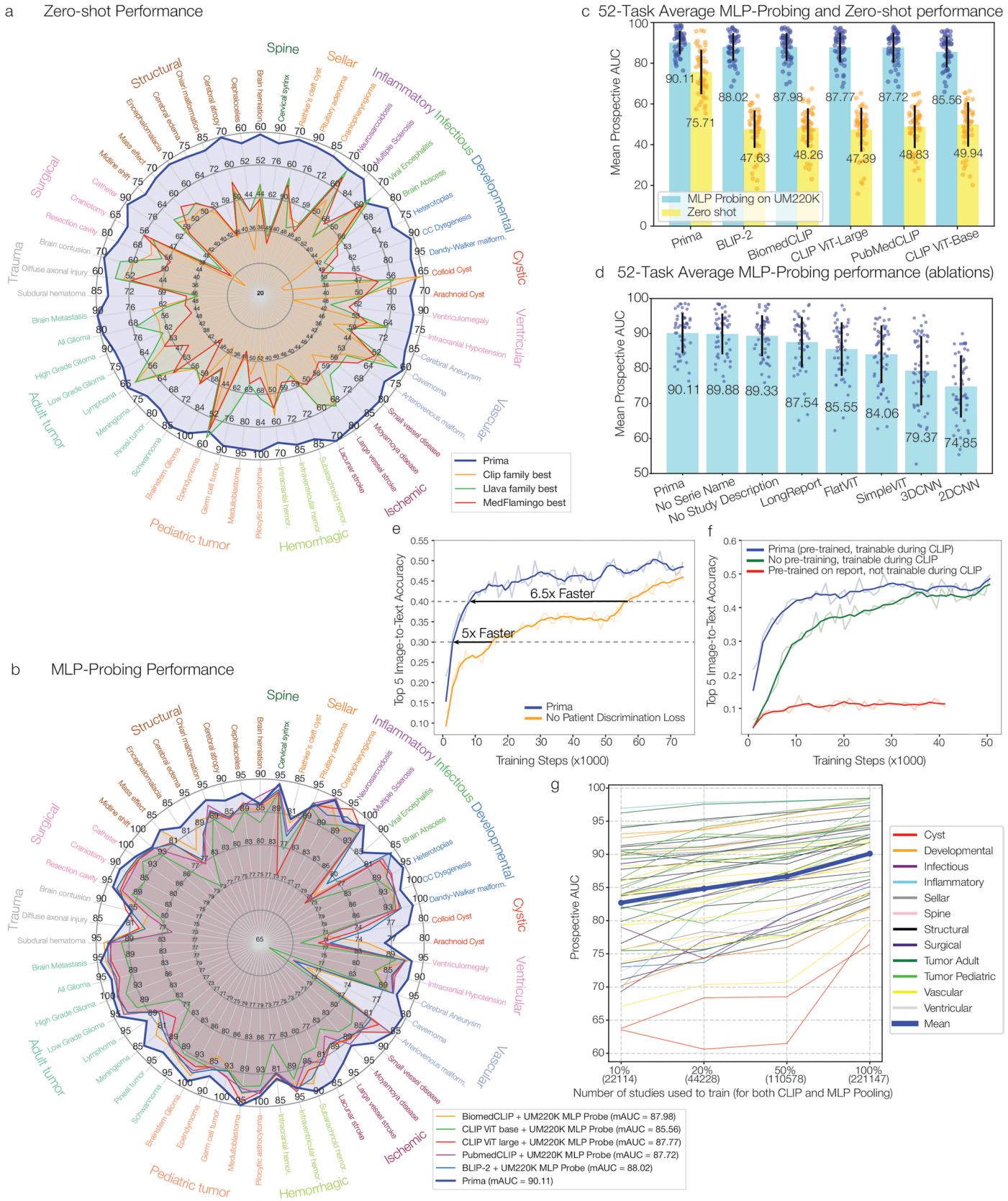


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 4 | MRI volume tokenization. a, MRI scanners acquire images with specified orientations (for example LAS, RAS, etc) and planes (for example axial, coronal, sagittal). The MRI tokens will have the same orientation and plane as the source MRI sequence after patching. **b**, Examples of VQ-VAE reconstructions in different MRI sequences and orientations. **c**, Because Prima takes as input multiple different orientations and imaging planes, the volume tokenizer should be orientation invariant, meaning the representation of the same anatomic region should not change if imaged in axial versus coronal plane or LAS versus RAS orientation, for example. We used two strategies: random orientation permutations and 3D-CNN encoders. Our VQ-VAE volume tokenizer is encouraged to encode each volume token equivalently across all orientations under a reconstruction loss. Examples of MRI subvolumes are shown in different orientations after permutation. The latent volume tokens with near-equivalent latent encodings are shown in the center panel. with the reconstructions after the decoder on the right. **d**, Ablation study over the codebook sizes shows the quantization loss validation curves. Larger codebook sizes led to less overfitting and better reconstructions. **e**, Reconstruction losses with and without random orientation permutations. Random permutations regularized the VQ-VAE and resulted in higher-quality reconstructions and lower reconstruction losses. **f**, Examples of reconstructions before and after orientation permutations for different MRI sequences. Reconstructions are perceptually equivalent after forward pass through the VQ-VAE model regardless of orientation or imaging plane. Subtle reconstruction differences can be seen on difference maps. Illustrations in **a** created with BioRender.com.

https://doi.org/10.1038/s41551-025-01608-0

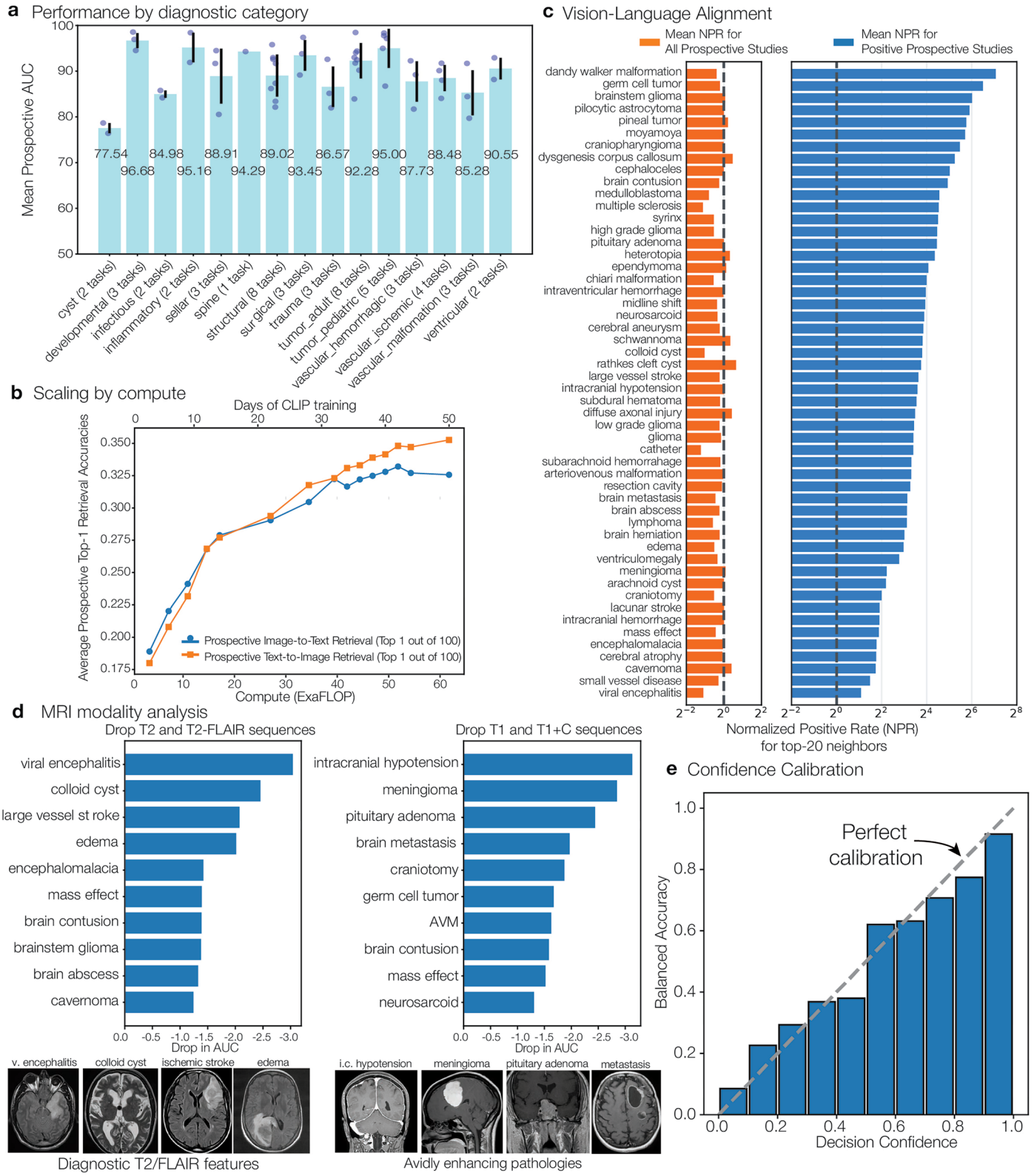


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 5 | MRI-report contrastive pretraining. a, Radar plot of prospective zero-shot performance (in AUC) comparison between Prima and open-source pre-trained VLM families across 52 tasks. The baseline performances reported for each VLM family is from the best-performing model in the family for each task. Zero-shot Prima substantially outperforms all baselines across the majority of tasks. **b**, Radar plot of prospective MLP-probing performance on UM-220K across 52 tasks, between Prima MRI study representations and average-pooled study representations from open-source pre-trained 2D CLIP-like models, all probed over UM-220K training data. Prima outperforms the 2D CLIP models, highlighting the importance of CLIP pre-training on 3D volumes and whole studies. **c**. Bar plot of Prima performance against baselines zero-shot performance and MLP-probing performance after training a diagnostic classifier on UM-220K data, over n=52 tasks. Error bars indicate standard deviation across AUC of each task. **d**, Mean prospective AUC across model design choices over n=52 tasks. Error bars indicate standard deviation across AUC of each task. Prima architecture outperformed other model designs. **e**, Top 5 validation set image-to-text retrieval accuracy of Prima with and without the patient discrimination loss. We see over a 5x speed-up in training time when using the patient discrimination loss. **f**, Ablation experiments over the neuroradiology language model. We found that pre-training on radiology reports resulted in more efficient training and that updating the language model during CLIP training was essential for MRI representation learning. **g**, AUC performance of Prima on each of the 52 tasks with various amounts of training data for both CLIP training and MLP probing. We have not observed an upper bound on performance and the data provides evidence that additional MRI data will continue to improve performance. These results emphasize the importance of health system-scale training and the health system-as-data engine framework.

https://doi.org/10.1038/s41551-025-01608-0

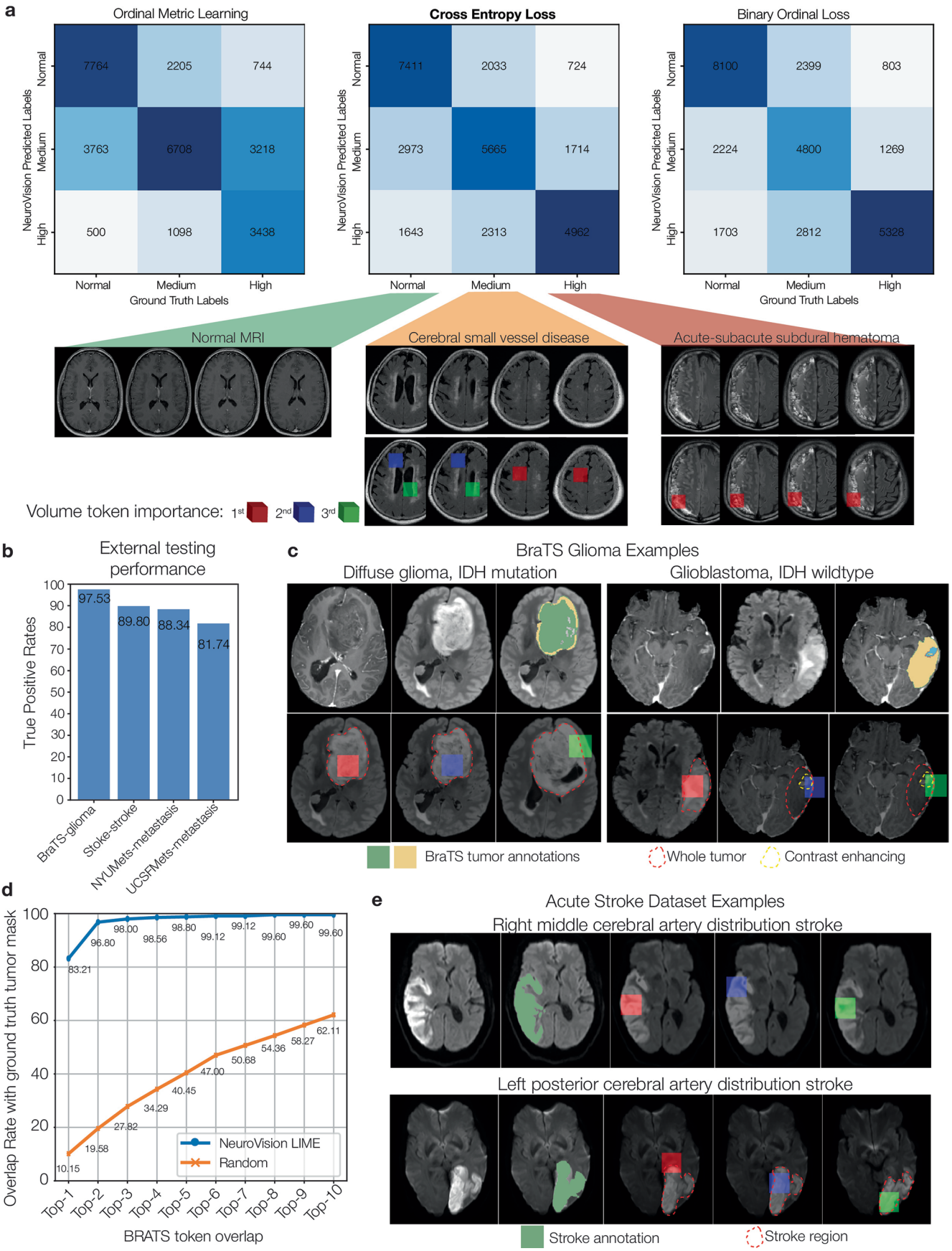


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 6 | Extended Prima performance results. a, Performance in each diagnostic category averaged over the diagnostic tasks. Error bars indicate standard deviation across AUC of each task within the category. **b**, Average Top-1 retrieval accuracy on the prospective testing cohort by days of CLIP training and amount of computation in ExaFLOPs. **c**, We show the normalized positive rate (NPR) of the top 20 nearest retrospective neighbors for prospective instances on each prediction task. NPR is calculated by $\frac{\text{average positive rate in top 20 nearest retrospective neighbors}}{\text{positive rate in retrospective set}}$. which indicates how many times as likely the top 20 nearest neighbors of a study are to have same positive label, compared to the positive rate of the full dataset. We compute NPR averaged across all prospective studies (orange) and across positive prospective studies only (blue). The expected NPR value for randomly distributed examples is 1 (dotted line). This analysis demonstrates that Prima embeddings tend to group studies with the same diagnoses closer together. **d**, We investigate Prima's use of MRI sequence modalities for radiologic diagnosis by dropping T1-weighted and T2-weighted modalities during inference, and showed the top 10 diagnoses with the highest drop in performance. For diagnoses that produce T2 hyperintensity such as encephalitis, acute ischemia, and cerebral edema, Prima shows an expected drop in diagnostic performance, indicating Prima is using known radiologic features to diagnose specific pathologies. We make similar observations when dropping T1-weighted modalities. This shows that Prima is utilizing MRI sequences for radiologic diagnosis, concordant with human interpretation. **e**, A reliability diagram shows that Prima's diagnostic outputs are well calibration [34]. The sigmoid prediction logit is the confidence score. We then calculate the balanced accuracy of the predictions with a confidence within each 0.1-binned interval. A calibrated classifier should have a confidence score that matches the accuracy within each interval (gray dashed line).

https://doi.org/10.1038/s41551-025-01608-0

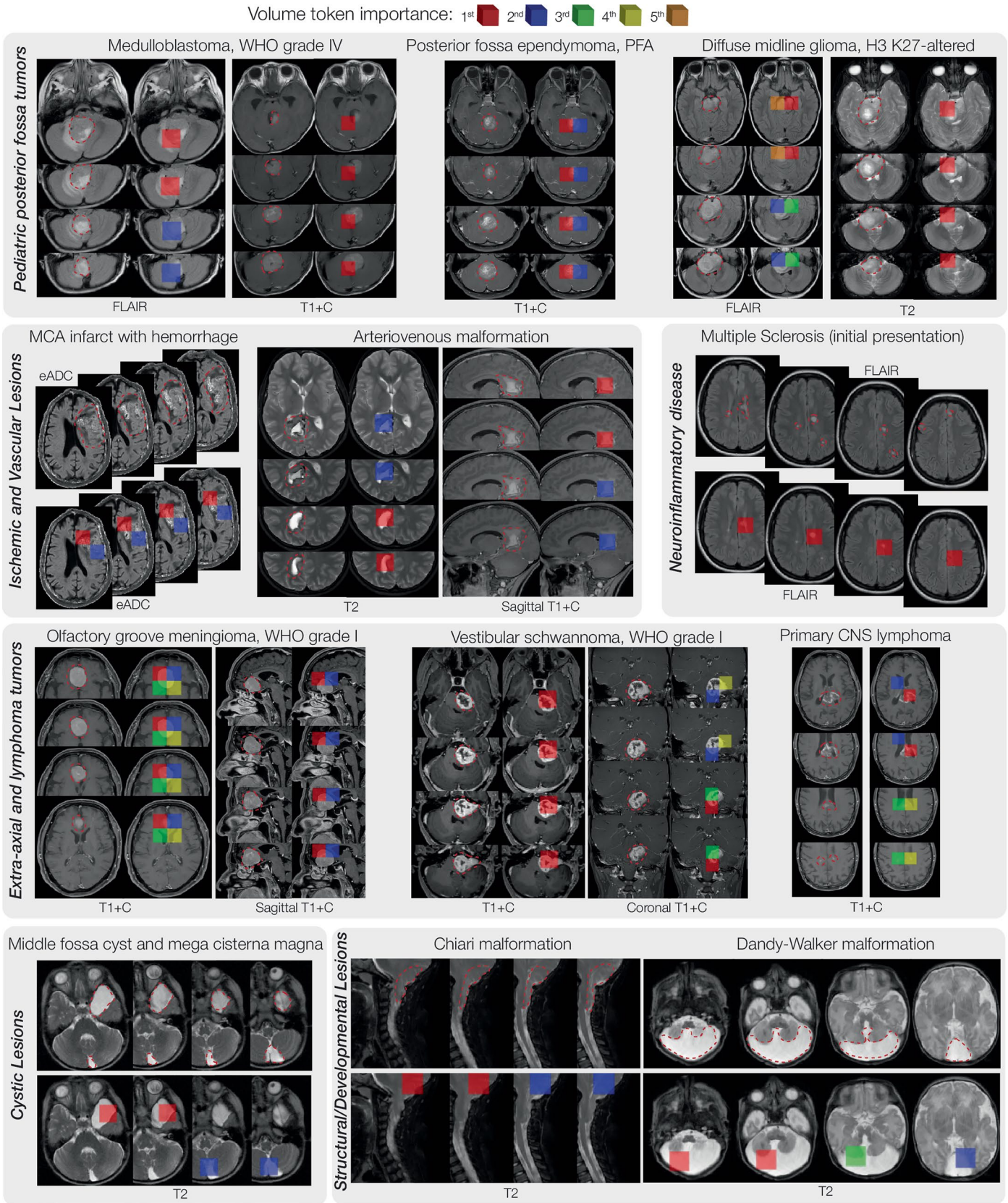


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

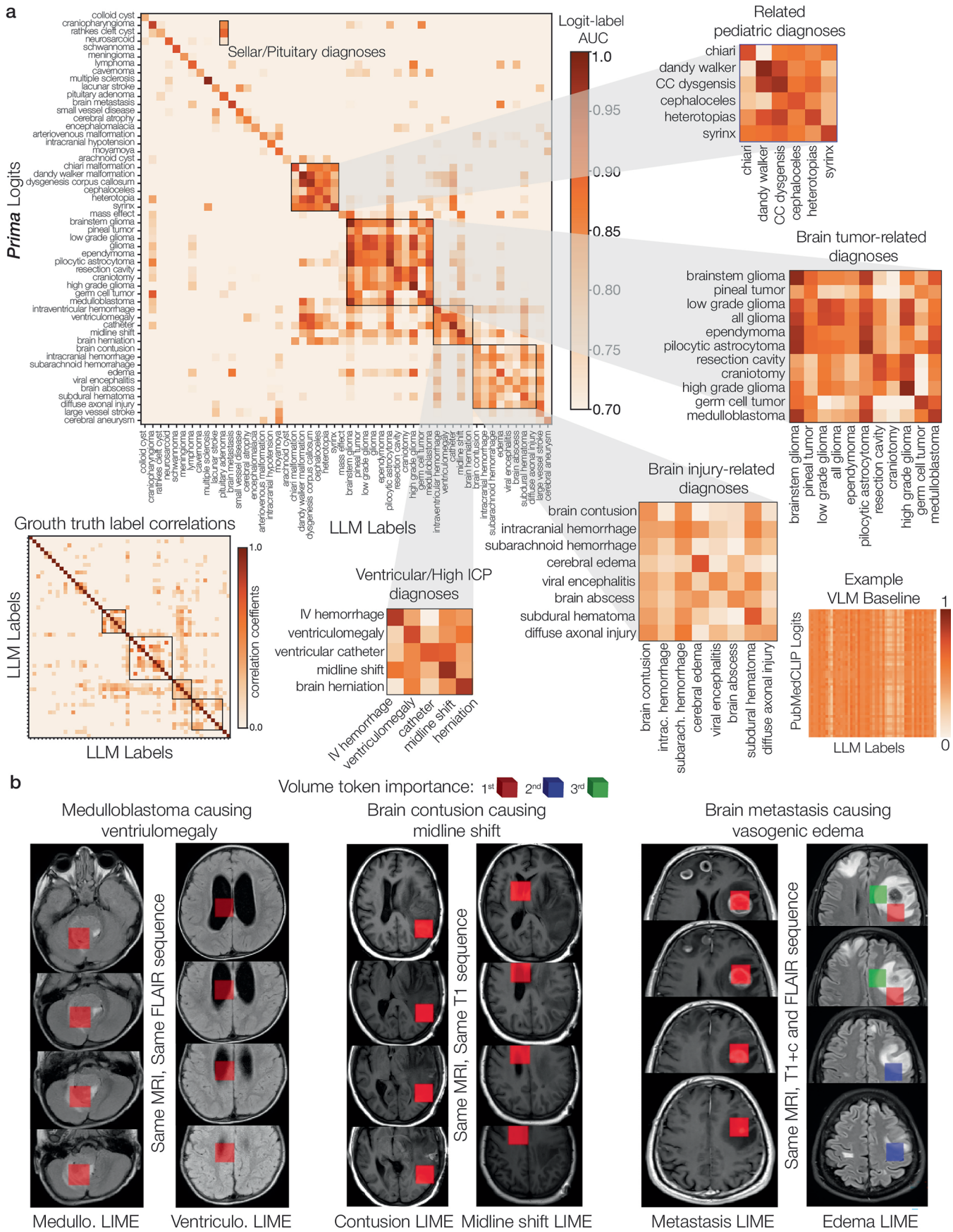
Extended Data Fig. 7 | Extended clinical task and transfer learning results. **a**, Confusion matrices for ordinal metric learning, cross entropy, and binary ordinal losses for MRI prioritization. Cross entropy provided the overall best results. Examples of normal, medium, and high acuity are shown. Most importantly for triage and acuity assessment, misclassification rates were lowest for normal-high discrimination. **c**, External testing performance on BraTS, Stroke, NYUMets, and UCSFMets datasets. We see true positive rates on par with our prospective testing cohort. **c**, LIME importance scores extended to the BraTS dataset, identifying volume tokens within the externally annotated tumor regions. An example of both a lower grade diffuse glioma, IDH mutation, and a higher grade glioblastoma, IDH wildtype, MRI are shown. Prima correctly identified the FLAIR hyperintense regions as evidence of tumor infiltration. Contrast enhancing regions of the glioblastoma were most important for high grade glioma classification. **d**, Quantitative evaluation of LIME volume token selection versus the segmented tumor regions. We show that Top-3 accuracy for Prima selecting volume tokens within the ground truth segmented tumor region was 98.0%. These results provide external, quantitative evaluation of trustworthy predictions from Prima. **b**, LIME visualizations for our external acute stroke dataset. An example of an acute middle cerebral artery distribution stroke and a posterior cerebral artery distribution stroke are shown. Prima assigned high LIME score to regions of diffuse restriction on DWI images, indicative of acute ischemia.

https://doi.org/10.1038/s41551-025-01608-0



Extended Data Fig. 8 | Diverse Clinical Examples of Prima Explainibility. We show a diverse selection of patients from our prospective testing cohort with associated LIME importance scores on the volume tokens. The top 5 tokens are color-coded according to the legend above. Prima correctly identifies the pathologic regions in all the clinical scenarios presented above, including pediatric posterior fossa tumors, vascular malformations, ischemic lesions, adult brain tumors, brain cysts, and developmental abnormalities. High LIME score tokens that localize to the pathologic regions demonstrate trustworthy Prima predictions. Red dashed line represented expert annotated pathologic regions. Full interactive demonstration with LIME visualizations and predictions can be found at prima.mlins.org.

https://doi.org/10.1038/s41551-025-01608-0

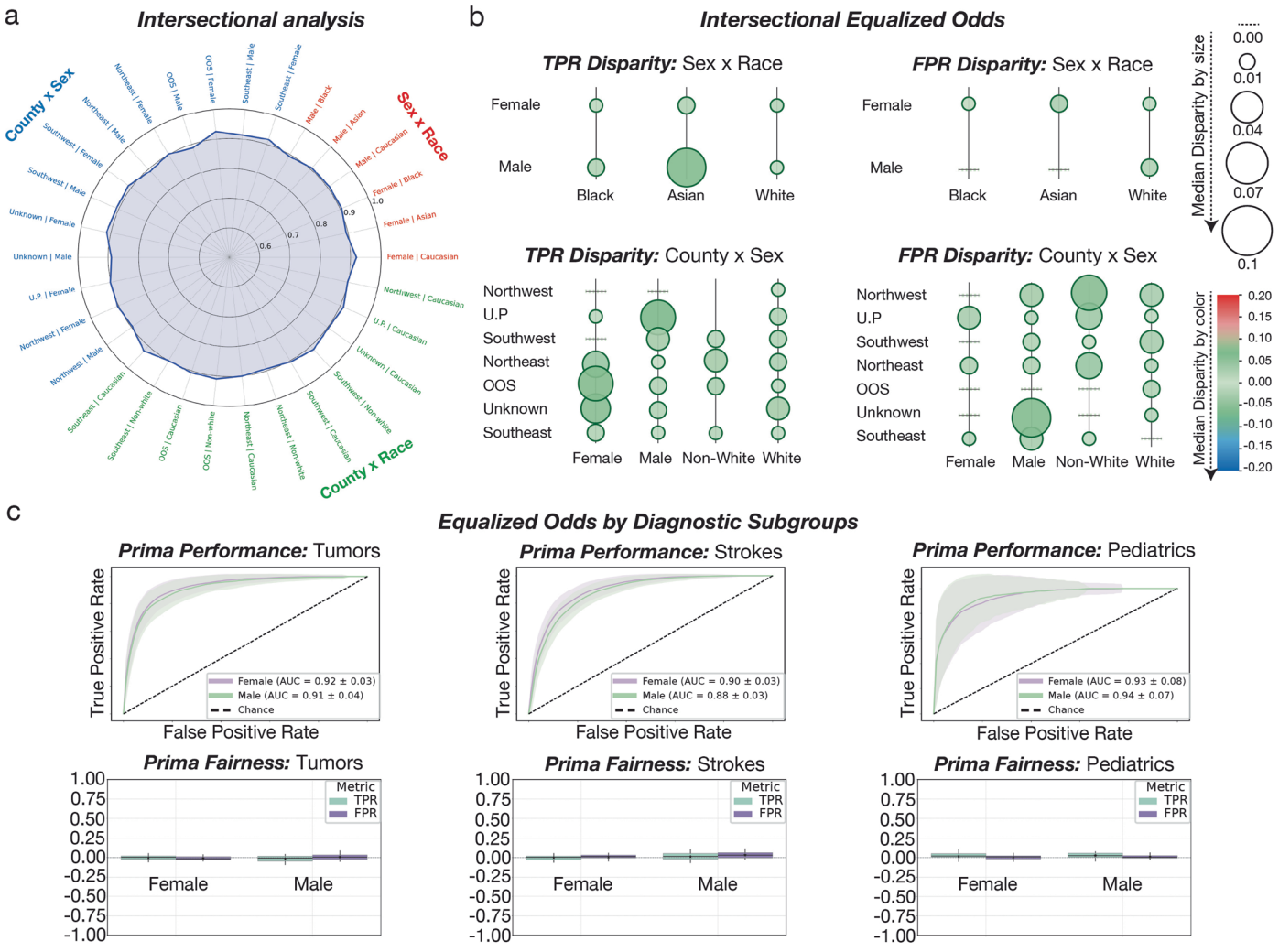


Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41551-025-01608-0

Extended Data Fig. 9 | Multi-label analysis of Prima. a, We performed a multi-label classification analysis across all 52 diagnoses. The Prima logit-label matrix above shows the AUC value for each logit-label binary comparison. The diagnoses were ordered using consensus clustering for easier visualization [69]. The lower left-hand matrix is the correlation matrix of the ground truth labels with the same ordering. Prima learned the label co-occurrence relationships, such as ventriculomegaly often co-occurs with ventricular catheters or brain tumors co-occur with mass effect. It also correctly captures the semantic similarity of diagnoses within differential diagnoses. We observe higher AUC values for related structural pediatric diagnoses, brain tumor diagnoses, and brain injury/trauma. These findings demonstrate that Prima has correctly modeled the multi-label classification problem while learning the semantic relationships between related diagnoses. **b**, Multi-label LIME analysis shows that Prima attends to different pathologic regions of the same MRI sequence depending on the diagnostic prediction. High LIME scores are assigned to tokens within the posterior fossa when the LIME scores are computed for 'medulloblastoma' prediction. Conversely, high LIME scores are assigned to the enlarged ventricles when the LIME scores are computed for 'ventriculomegaly' prediction. We see similar patterns investigating the relationship between 'brain contusion' and 'midline shift' labels, and 'brain metastasis' and 'vasogenic edema' labels. Prima demonstrates trusthworthy multi-label classification.

https://doi.org/10.1038/s41551-025-01608-0



Extended Data Fig. 10 | Intersectional and equalized odds fairness analysis. **a**, Intersectional fairness analysis of sex, race, and geographic region on our prospective testing cohort that includes a diverse patient population. Radar plot shows each intersectional group and mean AUC values for each group. Prima shows minimal variance in diagnostic performance across intersectional groups. **b**, Intersectional equalized odds analysis demonstrate minimal disparity in TPR or FPR across intersectional groups. TPR and FPR disparity was less than 0.1 threshold we used for clinical significance. **c**, Equalized odds by diagnostic subgroup are shown, including brain tumors, strokes, and pediatric diagnoses. Similar to the demographic features shown in Fig. 4, there is minimal disparity between diagnostic subgroups. Prima displays algorithmic fairness on intersectional and equalized odds analysis of our diverse prospective testing cohort. The error bands in the AUC plots are standard deviations across individual tasks within each diagnostic subgroup (n=13 tasks for tumors, n=6 tasks for strokes, and n = 3 tasks for pediatrics). Box plots show the median (center line), interquartile range (box: 25th-75th percentile), and whiskers extending to the minimum and maximum values, plotted over 20 bootstrapped iterations of 200 samples from each subgroup.

nature portfolio

Corresponding author(s): Todd C. Hollon

Last updated by author(s): Oct 6, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected from entire University of Michigan health system PACS service. We have queried for all brain MRI from SECTRA data warehouse to narrow our search and only studies with report were included in final training cohort. We provide our SQL query as part of our supplemental materials.
Data analysis	Sample size calculations were completed in R (4.3.0) using the epiR package (2.0.63) epi.sssupb function. All other code was implemented in Python (version 3.9) using PyTorch (2.3.1) and Transformers (4.37.0) as the primary machine learning framework. Additional data analysis tools include numpy (2.0.2), scikit-learn (1.6.0), matplotlib (3.9.4), and pandas (2.2.3). All code and scripts to reproduce the training and inference of Prima are available on GitHub at https://github.com/MLNeurosurg/Prima under an MIT license.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The Prima model parameters will be publicly available for investigational use only under an MIT license. Institutional Review Board approval was obtained from University of Michigan for MRI data collection. Restrictions apply to the availability of raw patient MRI imaging data, which were used with institutional permission through IRB approval for the current study, and are thus not publicly available. All data sharing between medical centers is regulated through data use agreements with the study authors. A similar data-sharing protocol may be established for interested investigators. Please contact the corresponding author (T.C.H.) for any requests for data sharing, and a response will be made within 2 weeks. All requests will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic and investigational purposes.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	In accordance with the Declaration of Helsinki, we evaluated the performance and fairness of our model (PRIMA) across several sensitive attributes and subgroups including patient sex. Patient sex was defined as the biological classification of male or female and was obtained from our institutional PACS-interfacing Sectra Datawarehouse system where it is reported in a patient's medical record along with the corresponding MRI study. Given that this study exclusively consisted of secondary analysis of existing data, the need for individual informed consent was waived during our IRB approval process. Protection of patient identity was paramount and not released publicly. In this study, we evaluated model fairness using an "equality of opportunity" criterion, comparing the true positive rate (TPR) of male and female patients compared to the total population of patients. TPR was reported in a disaggregated fashion for males and females to clearly demonstrate the similarity of model performance in each group.
Reporting on race, ethnicity, or other socially relevant groupings	According to the Declaration of Helsinki, it is imperative to provide appropriate representation in medical research to underrepresented groups, particularly those with distinct healthcare needs, healthcare access, and potential benefits from the research findings. As a result, we further evaluated our model fairness in socially relevant groupings including patient race and area-of-residence. This information was obtained from our institutional data warehouse which contains patients' self-reported race and county-based residence. With respect to race, included group classifications were African American, Asian, and Caucasian. As for counties, all Michigan counties were included for the determination of odds of a 1-week MRI turnaround read relative to the most represented county. Counties were subsequently grouped into geographically informed regions. To account for possible confounding effects, our assessment of subgroup-stratified TPR was conducted separately for race and county groups. Additionally, comparisons of each subgroup TPR to the population TPR were further stratified by diagnosis-distribution to account for the possible impact of diagnosis on model performance.
Population characteristics	Several sensitive attributes were reported for our study population including age, sex, race, insurance type, and area-of-residence. Diagnostic attributes were also collected across 52 diagnostic classes and reported for each patient. The demographic and diagnostic distributions of our data are included as supplemental data tables.
Recruitment	All MRIs and associated reports were obtained via a SQL-query of our institutional Sectra Data Warehouse which interfaces with our clinical PACS system. This study was a secondary analysis of existing data. We address the risk of patient selection bias by querying all MRIs included in the PACS system without exclusion of any geographic regions or patient groups.
Ethics oversight	This study protocol was approved by our Institutional Review Board at University of Michigan.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We acquired 279,908 clinical studies with over 170000 unique patients within our study cohort. These studies includes about 5.6 million sequences and over 362 million images. In addition, we have 29,435 MRI studies included in the prospective testing set. The prospective sample size was justified through statistical calculations detailed in the methods section.

Data exclusions	We describe our training data filtering process in details in Extended data figure 1. We excluded all studies with no available associated radiology reports, less than 2 valid sequences, or incomplete metadata, resulting in a final training set of size 221,147.
Replication	The model trained on the training set was used to make predictions on the prospective test set, which is completely mutually exclusive from the training set temporally. The model achieves high AUC on classification tasks on the prospective test set, indicating that the model's prediction capabilities can be reproduced on MRI studies outside the training set.
Randomization	The training set and the prospective test set was determined by MRI acquisition time. The prospective test set includes all studies acquired between June 1, 2023 and May 30, 2024, while the training set consists of studies acquired prior to June 1, 2023. During CLIP training, a randomly selected held-out of size 254 was selected from the training set; during classification head training, a balanced set of 100 positive/100 negative studies was randomly selected from the training set for each class as validation.
Blinding	Blinding was not relevant to our study. We performed no clinical trials and the study is non-interventional.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Our prospective study was a non-interventional, diagnostic study so it does not meet criteria for clinical trial registration.
Study protocol	Our prospective study was a non-interventional, diagnostic study, not a clinical trial. The detailed study evaluation protocol is included in the Methods sections of our submission.
Data collection	Our dataset included all previous studies completed or recorded (Outside MRIs) at University of Michigan Health System.
Outcomes	We evaluated our trained model on a prospective test set of MRIs, and reported prediction performance on many downstream tasks, including 2 diagnostic predictions (arachnoid_cyst, colloid_cyst, dandy_walker_malformation, dysgenesis_corpus_callosum, heterotopia, brain_abscess, viral_encephalitis, multiple_sclerosis, neurosarcoid, craniopharyngioma, pituitary_adenoma, rathkes_cleft_cyst, syrinx, brain_herniation, cephaloceles, cerebral_atrophy, chiari_malformation, edema, encephalomalacia, mass_effect, midline_shift, catheter, craniotomy, resection_cavity, brain_contusion, diffuse_axonal_injury, subdural_hematoma, brain_metastasis, glioma, high_grade_glioma, low_grade_glioma, lymphoma, meningioma, pineal_tumor, schwannoma, brainstem_glioma, ependymoma, germ_cell_tumor, medulloblastoma, pilocytic_astrocytoma, intracranial_hemorrhage, intraventricular_hemorrhage, subarachnoid_hemorrahage, lacunar_stroke, large_vessel_stroke, moyamoya, small_vessel_disease, arteriovenous_malformation, cavernoma, cerebral_aneurysm, intracranial_hypotension, ventriculomegaly), 15 referral predictions (neurosurgery pediatric, neurosurgery skull base, neurosurgery general, neurosurgery trauma, neurosurgery tumor, neurosurgery vascular, neurology pediatric, neurology epilepsy, neurology neurocritical, neurology neuroimmunology, neurology neurocognitive, neurology oncology, neurology trauma, neurology stroke, neurology general), and patient acuity prioritization. Additional tasks evaluated on external public datasets include brain age prediction, Autism/ADHD, and Alzhimer's/dimentia. Our model achieves strong performance across all tasks evaluated.

Plants

Seed stocks — *Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes — *Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication — *Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

Magnetic resonance imaging

Experimental design

Design type: Clinical MRI

Design specifications: Our dataset include various sets of clinical protocol such as brain, stroke, pituitary gland protocol etc.

Behavioral performance measures: Our study did not involve behavioral performance measures.

Acquisition

Imaging type(s): T1, T2, DWI, FLAIR, ADC, SWI, MPRAGE

Field strength: 1.5T-3T

Sequence & imaging parameters: The sequence and imaging parameters varied a lot within our study cohort due to health-system wide clinical datasets.

Area of acquisition: Our major focus for this study were head, brain, orbits, neck areas of acquisition.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software: The MRI data were processed using following software and frameworks; Python (v3.9), SimpleITK (v2.2.1), OpenCV (v4.1.2.30), numpy (v1.24.3), pydicom (v2.3.0), nibabel (v4.0.1), Pytorch (2.3.1+cu121)

Normalization: The data was not normalized. We did not perform normalization because we really want to create a model that can generalize to all types of MRI sequences from all types of data sources and machines.

Normalization template: No normalization was performed.

Noise and artifact removal: No noise or artifact were removed as a preprocessing step. Our VQVAE model has some smoothing artifact after reconstruction. We had additionally reoriented all the data to LPS orientation.

Volume censoring: The volume weren't censored (neither defaced or skull stripped), although the metadata were removed from raw dicom files and stored separately. The original volume accession numbers were replaced with UID.

Statistical modeling & inference

Model type and settings: Our model consists of three parts: a VQ-VAE model that converts each volume token into token representations, hierarchical multimodal transformers that generates a single representation vector for an entire MRI study, and classification heads that makes predictions based on encoded study vectors. The hierarchical multimodal transformers includes several transformer-based parts, including a ViT that encodes each sequence into a representation vector, a ViT that takes in all sequence vectors of a study and generates a study representation vector, and a causal transformer that encodes sequence names. The classification head is a 3-layer MLP that is trained separately for different downstream tasks.

Effect(s) tested: Main downstream prediction tasks include 52 diagnostic predictions (arachnoid_cyst, colloid_cyst, dandy_walker_malformation, dysgenesis_corpus_callosum, heterotopia, brain_abscess, viral_encephalitis, multiple_sclerosis, neurosarcoid, craniopharyngioma, pituitary_adenoma, rathkes_cleft_cyst, syrinx, brain_herniation, cephaloceles, cerebral_atrophy, chiari_malformation, edema, encephalomalacia, mass_effect, midline_shift, catheter, craniotomy, resection_cavity, brain_contusion, diffuse_axonal_injury, subdural_hematoma, brain_metastasis, glioma, high_grade_glioma, low_grade_glioma, lymphoma, meningioma, pineal_tumor, schwannoma, brainstem_glioma, ependymoma, germ_cell_tumor, medulloblastoma, pilocytic_astrocytoma, intracranial_hemorrhage, intraventricular_hemorrhage, subarachnoid_hemorrahage, lacunar_stroke, large_vessel_stroke, moyamoya,

small_vessel_disease, arteriovenous_malformation, cavernoma, cerebral_aneurysm, intracranial_hypotension, ventriculomegaly), 15 referral predictions (neurosurgery pediatric, neurosurgery skull base, neurosurgery general, neurosurgery trauma, neurosurgery tumor, neurosurgery vascular, neurology pediatric, neurology epilepsy, neurology neurocritical, neurology neuroimmunology, neurology neurocognitive, neurology oncology, neurology trauma, neurology stroke, neurology general), and patient acuity prioritization. Additional tasks evaluated include brain age prediction, Autism/ADHD, and Alzhimer's/dimentia.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See Eklund et al. 2016)

Cluster wise (voxels are clustered into volumn tokens before being fed into our model). Each volume token is clustered by location of voxels, i.e. the entire 3D sequence (256x256xH voxels) is cut up into volume tokens, where each token is a cube of 32x32x4. The z dimension is further scaled up to final shape of 32x32x8.

Correction

When training classification heads, we hold-out a balanced set of training instances as validation set, which is not used to train the classification heads. We select the checkpoints of the classification heads with the highest performance on the held-out validation set to be included in the final model.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

Our model consists of three parts: a VQ-VAE model that converts each volume token into token representations (for dimension reduction), hierarchical multimodal transformers that generates a single representation vector for an entire MRI study, and classification heads that makes predictions based on encoded study vectors. The hierarchical multimodal transformers includes several transformer-based parts, including a ViT that encodes each sequence into a representation vector, a ViT that takes in all sequence vectors of a study and generates a study representation vector, and a causal transformer that encodes sequence names. The classification head is a 3-layer MLP that is trained separately for different downstream tasks. The model is trained together with a language encoder (a pre-trained causal transformer based on gpt-2 model) with CLIP objective on 221,147 pairs of MRI-study and corresponding summarized radiology reports, with additional patient discrimination loss and augmentations. The classification heads are trained separately for each downstream prediction task with different objectives (binary cross entropy loss with positive weighting for all diagnostic and referral tasks, as well as Autism and dimentia; MSE loss for age prediction; and cross entropy loss / binary ordinal metric loss for acuity prioritization). Metrics reported for predictive analysis includes Area Under ROC Curve (AUC), Balanced Accuracy, Retrieval Accuracy, and MAE.