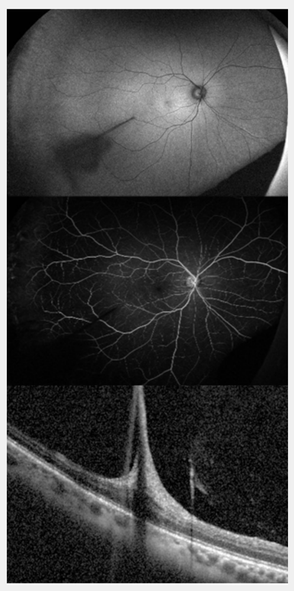
Towards Evaluating Capabilities of Vision Language Models in Ophthalmology

Rishi Ramessur [1,2] Peter Thomas [2] Luciana D'Adderio [1] Sotirios A Tsaftaris [1]
Steven McDonagh [1]

[1]The University of Edinburgh, United Kingdom [2]Moorfields Eye Hospital, London, United Kingdom

Motivation

- **Promise**: Multi-modal neural models can interpret images + text and support interactive queries (e.g., community Optometrists).
- **Challenge**: Safe deployment requires clinical validation and reliable uncertainty estimation - both underdeveloped.
- **Solution**: A generalisable validation framework combining uncertainty estimation with selective prediction.
- **Impact**: Targeting low-risk, high-value tasks that reduce administrative burden and streamline patient pathways in Ophthalmology, the UK's largest outpatient specialty.



Uncertainty estimation

Defining uncertainty

- **Aleatoric uncertainty** refers to irreducible inherent noise in the data-generating process.
- **Epistemic uncertainty** is reducible and arises from a lack of knowledge about the model or the data, typically from insufficient training data, poor model fit, or limited prior information.

Measuring uncertainty

Calibration assesses whether an uncertainty metric's predicted probability correlates with observed probability. If a VLM assigns 80% confidence that a referral is "urgent," then $\sim$ 8 in 10 cases should truly be urgent; if only 5 in 10 are, it is poorly calibrated, posing risks for triage.

Uncertainty estimation methods (priority-ranked)

- **Ensemble methods** train multiple model instances and estimate predictive uncertainty from variation in their outputs due to differences in learned parameters.
- **Bayesian methods** infer uncertainty indirectly, modelling distributions over weights/ measuring variance across predictions from multiple forward passes with dropout enabled at inference.
- **Sample consistency methods** estimate uncertainty by measuring variability in model responses across multiple stochastic forward passes to a perturbed, but semantically similar, prompt.
- Deterministic uncertainty estimation
  - **Confidence elicitation** directly prompts the model to estimate certainty for the response to a given prompt.
  - **Token level probabilities** aggregate probabilities assigned to each token in a model's generated response.
- **Mechanistic interpretability** probes causal relationships in internal model architecture.
- Emerging directions:
  - Prompt tuning can influence model calibration without modifying the underlying network.
  - LoRA-ensembles enhance calibration without incurring the computational costs of traditional ensembles.
  - Conformal prediction ensure the true label y lies in the prediction set with $\geq (1-\alpha)$ probability, without altering model weights - ideal for API-restricted VLMs.
  - Temperature scaling offers a post-hoc solution to calibration by modifying the softmax output probabilities.
  - Training a density estimator in the latent space of encoder embeddings and using it during inference to adjust the softmax outputs [1].

Use cases

Referral classification: Regional variation in hospital coding affects reimbursement. Extracting diagnosis, complexity, and expected activity from referrals enables auditing and standardisation of coding across providers, and strengthens financial oversight.

Leveraging imaging modalities for referral precision: Classify retinal colour fundus photographs (CFP) and Optical Coherence Tomography (OCT) scans as *emergency*, *urgent*, *routine*, or *no referral*. Combining with coding data may further strengthen referral classification.

Text-based information extraction for audit: Test VLMs' ability to handle longitudinal clinical text (notes, letters) with uncertainty estimation, to automate case identification for clinical audit.

Figure 1. Use cases for a selective prediction pipeline in Ophthalmology

Tasks

Extraction	Reasoning
Parse referral fields (e.g., demographics, symptoms, duration)	Standardise units (convert Snellen visual acuity to LogMAR; standardise refraction data)
Extract visual acuity, eye pressure, laterality	Infer urgency, target subspecialty, diagnosis
Identify imaging modality (CFP, OCT)	Cross-reference referral with clinical notes
Identify and label missing information	Standardise abbreviations and full form

Table 1. Separating extraction from reasoning ensures reasoning is applied only to reliable structured outputs.

Connect with me on LinkedIn!

Datasets

Datasets for model fine-tuning

Fine-tuning VLMs on limited ophthalmic datasets (referrals, notes, CFP/OCT) helps adapt them to domain-specific tasks, but small sample sizes increase risks of overfitting and bias. Uncertainty estimation becomes critical (at training time to highlight underrepresented or noisy data; at inference time to flag unreliable predictions), supporting safer model adaptation.

Datasets for clinical validation

Referral classification: A de-identified corpus of referral data from the Moorfields Single Point of Access pathway, covering North Central and North East London since June 2023.

Medical image classification: A corpus of de-identified and labelled Moorfields retinal CFP and OCT images from different imaging machines from 2012-2024.

Clinical information extraction:

1. Open access datasets:
   480 glaucoma appointment notes [2].
   US EHR entries: AAO IRIS registry [3] (access request permitting).
   All of Us – genomic and EHR data [4].
2. A corpus of de-identified patient letters from Moorfields' EHR across all sub-specialties from 2012-2024.

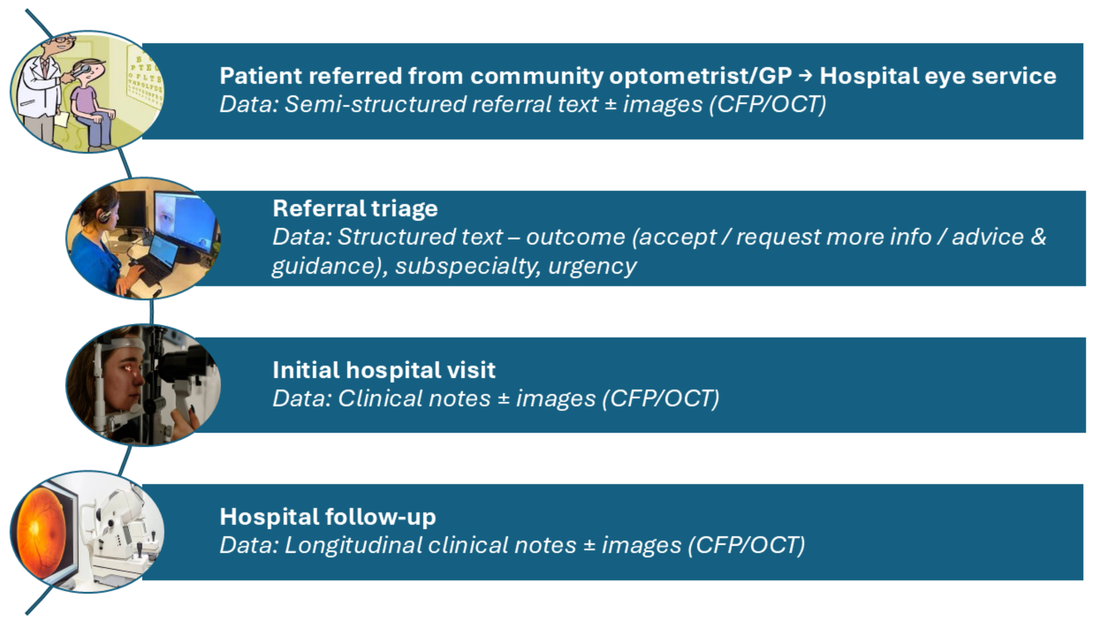


Figure 2. Our work uses real-world referrals, imaging, and text notes - capturing complexities of clinical practice often absent from synthetic benchmarks. These conditions test VLM robustness, generalisability, and safety, while approved datasets may also serve as benchmarking resources for the wider research community.

Technology transfer

Regulating VLMs in healthcare must balance transformative potential with safety, equity, and privacy. Unlike deterministic processes, VLMs show stochastic behaviour - identical inputs can yield variable outputs - creating unique risks which require continuous monitoring, adaptive governance, and new validation standards beyond one-size-fits-all frameworks. We plan to:

1. **Regulatory Landscape Review**: Analyse existing frameworks (MHRA, EU MDR, FDA), with focus on model adaptation, safety, and explainability.
2. **Case Studies**: Evaluate successes and failures in AI-based vision models' regulatory approval, highlighting roles of uncertainty estimation.
3. **Stakeholder Engagement**: Interview developers, regulators, clinicians, administrators to identify deployment barriers: validation, interpretability, liability, workflow integration.
4. **Uncertainty in Decision-Making**: Human-in-the-loop studies to test how uncertainty estimates affect clinical trust, adoption, and triage decisions.
5. **Commercial and Ethical Pathways**: Assess IP, licensing, and commercialisation strategies, alongside ethical issues of bias, privacy, and responsible use.
6. **Technology Transfer Framework**: Propose a structured approach to deploy multimodal models, with uncertainty estimation central to regulatory approval and clinical trust.

Significance of work

- Benchmark VLMs on real-world ophthalmic referral/CFP/OCT/EHR data.
- Ophthalmology-specific tasks defined to evaluate model performance.
- Selective prediction framework to allow models to abstain when uncertain.
- Translational framework: human-AI interaction, stakeholder input, regulatory analysis.
- Integration pathways into referral and EHR workflows for safe deployment.
- Early use in low-risk settings to build evidence for wider, scalable adoption.

Acknowledgements and References

Funding from University of Edinburgh UKRI AI Centre for Doctoral Training in Biomedical Innovation and Causality in Healthcare AI (CHAI) Hub. The authors declare no conflicts of interest.

[1] Ha Manh Bui and Anqi Liu. Density-softmax: Efficient test-time model for uncertainty estimation and robustness under distribution shifts, 2024.

[2] Jimmy S Chen et al. Development of an open-source annotated glaucoma medication dataset from clinical notes in the ehr. *Translational Vision Science & Technology*, 11(11):20–20, 2022.

[3] DW Parke Ii, F Lum, and WL Rich. The iris® registry: purpose and perspectives. german version. *Der Ophthalmologe*, 113:463–468, 2016.

[4] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.