Investigating Demographic Bias in Medical AI on Drusen Detection

Andreas Pattichis[1], Pedro Moreira[1], Ahmet Çalış[1], and Dariana Dorin[1]

Universitat Pompeu Fabra, Barcelona, Spain
{andreas.pattichis01, pedro.moreira01, ahmet.calis01, dariana.dorin01}@estudiant.upf.edu

Abstract. Artificial intelligence holds significant promise for healthcare diagnostics, yet demographic biases related to age and sex pose challenges for equitable implementation. This meta-research study explored the impact of these biases on the diagnostic accuracy of OpenAI's GPT-4 model in detecting Drusen using fundus images from the Brazilian Multilabel Ophthalmological Dataset. The study evaluated model performance across different age groups (young, middle-aged, older) and sexes (male, female) in two experiments. Despite small sample sizes, findings suggested a performance bias favoring older adults and female patients. These trends underscore the need for further investigation to ensure fair, effective AI healthcare outcomes across demographic groups.

Keywords: AI in healthcare · AI bias · Medical imaging · Drusen detection · Diagnostic accuracy · GPT-4

1 Introduction

Artificial intelligence (AI) is rapidly transforming healthcare by enhancing diagnostics and patient care. Advanced machine learning algorithms and deep neural networks enable AI systems to analyze complex medical data with precision, identifying subtle patterns that may go unnoticed by even the most experienced clinicians [4, 16]. Designed to support rather than replace human expertise, these AI tools assist clinicians in making more accurate and timely diagnoses. Research has demonstrated AI's effectiveness across various medical fields, including dermatology, radiology, and cardiology. In these areas, AI helps identify diseases such as skin cancer, diabetic retinopathy, and cardiovascular abnormalities by analyzing extensive imaging datasets [4, 1]. As AI technologies continue to advance, they promise to improve diagnostic accuracy, increase healthcare efficiency, and expand access to quality care [16].

This technological advancement is particularly impactful in ophthalmology, where AI-driven fundus imaging has revolutionized the diagnosis of retinal diseases. Fundus imaging captures detailed views of the retina and optic disc, essential for identifying and monitoring eye conditions such as diabetic retinopathy, glaucoma, and age-related macular degeneration (AMD) [15, 1]. By analyzing

these images, AI algorithms can detect disease markers with sensitivity comparable to or even surpassing that of human clinicians, enabling earlier diagnoses and better treatment outcomes [1]. This early detection capability allows AI to reduce diagnostic burdens on healthcare providers and facilitate regular, automated screenings for at-risk populations.

While specialized AI applications, such as those used in fundus imaging, have significantly advanced diagnostic processes, the broader landscape of AI in healthcare continues to evolve with the development of more versatile models. In this context, Large Language Models (LLMs) like OpenAI's GPT-4 have emerged as powerful tools that complement specialized systems. These models, known for their multimodal capabilities, can process and generate both text and image data, enabling them to support a wide range of tasks, including medical tasks and even beyond image analysis alone [3]. For instance, GPT-4 has shown promising results in Medical Visual Question Answering (VQA) by interpreting and responding to queries based on medical images [7]. Additionally, benchmarking studies indicate that GPT-4V(ision) can assist in ophthalmic multimodal image analysis, despite certain clinical limitations [18]. Integrating LLMs like GPT-4 with specialized imaging AI holds the potential to enhance diagnostic workflows, providing more comprehensive support to clinicians. While these models are not yet ready for fully automated classifications in medical settings, their potential suggests that LLMs will increasingly become essential components of medical imaging and diagnostic processes.

However, the deployment of AI in healthcare comes with significant challenges. One major concern is demographic bias in AI models, which can result in unequal diagnostic outcomes across different population groups. The reliability of AI systems depends heavily on the quality and diversity of their training data. Biases in medical imaging datasets—whether related to demographics (e.g., age, gender, ethnicity) or technical factors (e.g., imaging equipment variability)—can lead to inaccurate diagnoses [11].

Studies have highlighted the serious implications of biased AI models in healthcare. Esteva et al. warned that although AI can achieve high accuracy in tasks like skin cancer detection, biases in datasets may cause misdiagnoses, especially among underrepresented groups [4]. Similarly, Oakden-Rayner noted that limited demographic diversity in radiology datasets can reduce the generalizability of AI models, posing risks to patient safety in diverse clinical settings [10]. These issues are further underscored by Buolamwini and Gebru's "Gender Shades" project, which revealed significant accuracy disparities in commercial AI models for darker-skinned individuals and women, highlighting the ethical concerns of deploying biased AI in sensitive applications [2].

In medical imaging, various types of biases have been identified that could worsen existing health disparities. Rengan et al. pointed out that insufficient demographic representation in radiology datasets could increase inequities in healthcare outcomes, as AI models may perform unevenly across different patient groups [12]. Gichoya et al. advocated for implementing fairness audits and continuous assessments to ensure that AI models provide equitable healthcare benefits

[5]. Additionally, Kheiri et al. found that site-specific biases in histopathology images led to overly optimistic performance estimates, showing how data acquisition variability can affect model accuracy [6]. Stanley et al. suggested that data imbalances can distort AI predictions, resulting in disparities across various subgroups [14]. Saw and Ng emphasized the importance of robust data governance in medical AI to reduce these biases, promoting transparency, accountability, and ethical oversight in AI development and deployment [13].

AI applications in ophthalmology, particularly fundus imaging, are especially vulnerable to demographic biases due to the critical role of early disease detection in conditions like diabetic retinopathy and AMD [9]. Drusen, which are deposits associated with aging and retinal diseases, can vary significantly in appearance based on demographic factors such as age and sex [17]. As a result, biases in AI models for fundus imaging could lead to differences in diagnostic accuracy, potentially impacting patient care. Ensuring that AI systems are trained on diverse and representative datasets is crucial to minimize these biases and promote equitable healthcare delivery for all patient populations.

Despite promising advancements, a significant research gap remains in understanding demographic biases within LLMs applied to medical imaging. As the integration of LLMs like GPT-4 into healthcare continues to rise, it is crucial to identify and address these biases to prevent disparities in healthcare outcomes across different demographic groups. Currently, limited studies have explored how biases related to age and sex in LLMs affect their performance in specialized tasks such as medical imaging. This study aims to bridge this gap by focusing on Drusen detection in fundus imaging. Specifically, we investigate: *"How does demographic bias—specifically related to age and sex—impact the diagnostic accuracy of AI models in detecting ophthalmic conditions using fundus imaging, and what are the implications of this bias for equitable healthcare outcomes?"*. By exploring these questions, our research seeks to contribute to the development of more fair and effective AI-driven diagnostic tools in ophthalmology.

2 Methodology

2.1 Study Design

This meta-research study investigates demographic biases in AI diagnostics, specifically focusing on the GPT-4 model's performance in detecting Drusen in fundus images. GPT-4 was selected for this study due to its accessibility and its multimodal capabilities, which make it an emerging tool for medical imaging analysis. While more specialized models, such as Med-Gemini and Med-PaLM 2, have demonstrated superior performance in certain clinical applications, they are not publicly accessible [3]. As such, GPT-4 serves as a representative general-purpose LLM for exploring demographic biases in medical imaging, with the goal of informing the design of future equitable AI systems.

The primary objective is to identify and quantify performance disparities across different age and sex groups. A secondary objective is to understand the implications of these biases for equitable healthcare outcomes.

We hypothesize that due to GPT-4's lack of specific training for fundus imaging, the model may exhibit suboptimal performance overall. However, our primary focus is on identifying biases rather than evaluating performance, so this should not be a significant issue. Specifically, we expect the model to struggle more with detecting Drusen in younger individuals, where manifestations are less prevalent and pronounced compared to older adults. There should not be any additional bias towards a certain age group or gender group.

2.2 Dataset Description

The dataset used in this study is the Brazilian Multilabel Ophthalmological Dataset (BRSET), introduced by Nakayama et al. [8]. BRSET consists of 16,266 color fundus photographs from 8,524 Brazilian patients, annotated with comprehensive demographic and clinical information. Key features of the dataset include patient age, sex, comorbidities, and multiple ophthalmological conditions such as diabetic retinopathy, age-related macular degeneration (AMD), and Drusen. The dataset is multi-labeled, allowing for the classification of multiple conditions per image.

Drusen was selected as the condition of interest for this study because it is the second most common condition in the dataset, following increased cup-to-disc ratio. While increased cup-to-disc ratio has more counts, it is also more challenging to identify due to its subtle features, particularly for a general-purpose model like GPT-4 that is not specialized in medical imaging. By focusing on Drusen, which is more easily identifiable, we ensure that the study can effectively examine potential demographic biases without the additional complication of evaluating the model's ability to handle inherently difficult conditions. This approach aligns with our goal of analyzing biases rather than overall diagnostic performance. Figure 1 illustrates the distribution of disease conditions within the BRSET dataset.

2.3 Data Preprocessing and Sampling

Data preprocessing involved several key steps to ensure the quality and representativeness of the samples used in the experiments. Initially, all identifiable patient information was removed to maintain confidentiality. The dataset was then filtered to include only images labeled as "Adequate". This filtering resulted in a subset of images deemed suitable for reliable diagnostic analysis. Prior to stratification, a thorough examination was conducted to identify any missing data within the dataset. It was confirmed that there were no null values in the key demographic and diagnostic columns, ensuring the completeness and integrity of the data used in subsequent analyses.

To facilitate unbiased evaluation, the dataset was stratified into demographic groups based on age and sex. Age groups were defined as follows: *Young* (under 30 years), *Middle-aged* (30 to 60 years), and *Older* (over 60 years). were chosen based on established clinical findings that the prevalence and presentation of Drusen vary significantly with age, particularly increasing in older populations

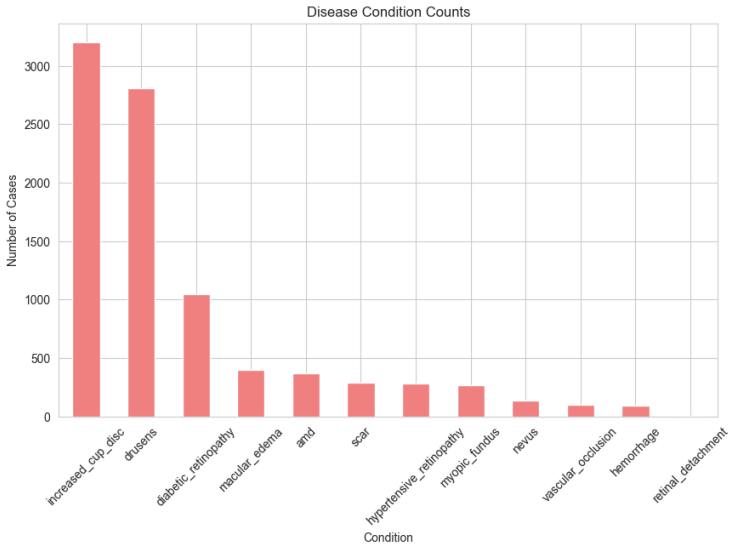

Fig. 1. Distribution of Disease Conditions in the BRSET Dataset

[17]. This stratification ensures a meaningful analysis of bias by reflecting real-world demographic patterns observed in ophthalmic conditions. For sex-based analysis, groups were categorized as *Male* and *Female*. Balanced sampling was employed to select 50 images per demographic subgroup for each experiment, ensuring comparable distributions of the Drusen condition across groups. Specifically, the sampling strategy aimed to include approximately half of the samples with Drusen present and half without, within each demographic category. This approach minimizes interfering factors and allows for a clearer assessment of potential biases related to age and sex.

Table 1 summarizes the distribution of Drusen condition across demographic groups. Overall, the data reflects a balanced sampling strategy, with slight variations in the presence of Drusen observed across age and gender groups.

Table 1. Drusen Condition Data Distribution by Demographic Groups

Demographic Group	**Presence of Drusen (Count)**
Young Adults (Age $<$ 30 years)	20
Middle-Aged Adults (Age 30–60 years)	25
Older Adults (Age $>$ 60 years)	27
Male Participants	30
Female Participants	31

2.4 Experimental Design

The experimental framework includes two primary experiments designed to assess demographic biases in GPT-4's diagnostic performance: one focusing on age groups and the other on sex groups. Each experiment involved the classification of 50 fundus images per demographic subgroup using GPT-4, with careful sampling to ensure comparable distributions of the Drusen condition.

For both age and sex experiments, 50 images were randomly selected per demographic subgroup. The selection ensured a balanced representation of Drusen presence, with a predetermined number of images containing Drusen within each group as specified in Table 2. This was achieved by first sampling the required number of images with Drusen and then filling the remaining slots with images without Drusen.

GPT-4's multimodal capabilities were accessed via OpenAI's API using a generated API key. Each fundus image was encoded in base64 format and included in the API request payload. A standardized prompt was utilized to ensure consistency across all classifications. The prompt instructed GPT-4 to respond exclusively with a binary indicator (1 for Drusen present, 0 for absent), without any additional explanations. Importantly, the model was not provided with any demographic information about the images. This allowed us to check for biases in the results based on age and sex without influencing the model's responses.The prompt that was used was the following:

> "You are an AI ophthalmology assistant analyzing a fundus image to check exclusively for the presence of Drusen. There may be other conditions present, but they are not relevant to this task. If Drusen is present, respond with the digit 1. If Drusen is not present, respond with the digit 0. Only respond with the single digit 1 or 0; provide no explanations or additional text. This is a strict format requirement."

It is important to acknowledge the practical limitations of this study, particularly the financial constraints associated with accessing GPT-4's API. As university students not funded for this research, these constraints necessitated the selection of a small sample size of 50 images per subgroup and limited each image to a single analysis to control costs. While these limitations restricted the scope of the analysis, the experimental design was sufficient to identify potential trends and assess demographic biases, laying a foundation for future, more expansive studies.

The model's responses were recorded alongside the actual labels from the dataset, forming the basis for computing evaluation metrics and bias assessment.

The two experiments were structured as follows:

- **Experiment 1: Age-Related Performance Evaluation** – Assess GPT-4's diagnostic accuracy across three age groups: Young, Middle-aged, and Older.
- **Experiment 2: Sex-Related Performance Evaluation** – Evaluate GPT-4's diagnostic accuracy between Male and Female groups.

Table 2 summarizes the demographic groups and the distribution of Drusen presence within each group.

Table 2. Summary of Experiments and Demographic Groups

Experiment	Demographic Group	Total Images	Drusen Present/Absent
Age-Related	Young (<30 years)	50	20 Present / 30 Absent
	Middle-aged (30–60 years)	50	25 Present / 25 Absent
	Older (>60 years)	50	27 Present / 23 Absent
Sex-Related	Male	50	30 Present / 20 Absent
	Female	50	31 Present / 19 Absent

2.5 Evaluation Metrics and Bias Assessment

For evaluation, we used a set of metrics to evaluate GPT-4's diagnostic performance and demographic biases. These metrics were calculated separately for each demographic subgroup to identify disparities.

Performance metrics included accuracy, precision, recall (sensitivity), F1 score, false positive rate (FPR), and false negative rate (FNR). Accuracy measured the proportion of correctly identified cases. Precision was the ratio of true positive predictions to total positive predictions. Recall represented the proportion of actual Drusen cases correctly identified. The F1 score balanced precision and recall. FPR indicated the proportion of non-Drusen cases incorrectly identified as Drusen, while FNR represented the proportion of Drusen cases not detected.

Bias metrics included demographic parity (selection rate) and calibration. Demographic parity assessed the proportion of individuals predicted to have Drusen within each demographic group. Calibration measured the agreement between predicted probabilities and actual outcomes.

3 Results

The performance of GPT-4 in detecting Drusen across different demographic groups is summarized in Tables 3 and 4 for the age-related and sex-related experiments, respectively.

3.1 Experiment 1: Age-Related Performance Evaluation

The results indicate that GPT-4's performance varies across different age groups. The model shows higher recall and selection rates in Older adults, suggesting a bias towards detecting Drusen in this group. The FPR also increases with age, indicating a higher likelihood of incorrectly identifying non-Drusen cases as Drusen in Older adults. This trend highlights the need for further refinement to ensure consistent performance across all age groups. The selection rate increases

with age, from 34% in Young to 60% in Older groups. Both TPR (Recall) and FPR increase with age, indicating higher detection and false positive rates in Older adults. This implies that the model is more likely to predict Drusen in older individuals, potentially leading to underdiagnosis in younger patients.

Table 3. GPT-4 Diagnostic Performance Metrics by Age Group

Metric	**Young**	**Middle-aged**	**Older**
Accuracy	62%	64%	62%
Precision (Presence)	52.94%	64.00%	63.33%
Recall (Presence)	45.00%	64.00%	70.37%
F1 Score (Presence)	48.65%	64.00%	66.67%
False Positive Rate (FPR)	26.67%	36.00%	47.83%
False Negative Rate (FNR)	55.00%	36.00%	29.63%
Demographic Parity	34%	50%	60%
Calibration	48.65%	64.00%	66.67%

3.2 Experiment 2: Sex-Related Performance Evaluation

In the sex-related experiment, GPT-4 demonstrates higher recall and selection rates in Female patients, indicating a bias towards detecting Drusen in this group. The FPR is also higher in females, suggesting a greater likelihood of incorrectly identifying non-Drusen cases as Drusen. These findings underscore the importance of addressing sex-related biases to ensure equitable diagnostic performance. Females have a higher selection rate (66%) compared to males (44%). Females exhibit higher TPR and FPR, indicating better detection but more false positives. This implies that the model may over-predict Drusen in females and under-predict in males, potentially leading to unequal healthcare outcomes based on sex.

Table 4. GPT-4 Diagnostic Performance Metrics by Sex Group

Metric	**Male**	**Female**
Accuracy	60%	68%
Precision (Presence)	72.73%	72.73%
Recall (Presence)	53.33%	77.42%
F1 Score (Presence)	61.54%	75.00%
False Positive Rate (FPR)	30.00%	47.37%
False Negative Rate (FNR)	46.67%	22.58%
Demographic Parity	44%	66%
Calibration	61.54%	75.00%

3.3 Overall Analysis

Overall, GPT-4 demonstrated reasonable diagnostic performance across all demographic groups, with accuracy ranging from 60% to 68%. However, the model exhibited biases in both age and sex-related experiments. In the age-related experiment, the model showed higher recall and selection rates in Older adults, accompanied by increased FPR, indicating a bias towards detecting Drusen in this group. In the sex-related experiment, the model showed higher recall and selection rates in Female patients, along with a higher FPR, suggesting a preferential detection in females. These findings highlight the need for further refinement of the model to ensure equitable diagnostic performance across all demographic groups.

4 Discussion

This study assessed demographic biases in GPT-4's Drusen detection using the BRSET dataset. Two experiments focusing on age and sex revealed biases favoring older adults and female patients.

In the age-related experiment, GPT-4 exhibited higher diagnostic accuracy and recall in Older adults compared to Young and Middle-aged groups (Table 3). This trend aligns with the clinical understanding that Drusen manifestations are more pronounced in older populations, potentially making them easier for AI models to detect. However, the accompanying increase in FPR among Older adults raises concerns about over-prediction, which could lead to unnecessary follow-up procedures and patient anxiety. On the other hand, the lower recall in Young individuals suggests a risk of underdiagnosis, potentially delaying critical interventions for this subgroup.

In the sex-related experiment, the model performed better in Female patients, as shown by higher Recall and Demographic Parity (Table 4). This suggests a bias that could be due to dataset imbalances or inherent model tendencies. While higher Recall lowers the risk of missing Drusen cases in females, the increased FPR presents similar issues as in the age-related analysis. The lower Recall in Male patients highlights the need for model improvements to ensure fair diagnostic capabilities across sexes.

Given that our study is based on a small subset, the findings are not statistically significant but do raise important questions. These issues should be addressed promptly before considering the broader use of AI models in medical tasks to prevent biases and incorrect classifications. Identifying and addressing these biases early on is crucial to ensure that AI models like GPT-4 do not perpetuate existing disparities in medical diagnostics. This is why it is important to conduct such studies from the beginning and identify potential biases early. By doing so, we can work towards refining these models to promote equitable healthcare outcomes and prevent any unintended consequences that could arise from biased AI systems. The identified potential biases have significant implications for equitable healthcare delivery. AI models like GPT-4, if deployed

without addressing these biases, could unintentionally extend existing disparities in medical diagnostics. The preferential detection in certain demographic groups may lead to unequal healthcare outcomes, undermining the foundational goal of AI in promoting universal health equity.

Our findings resonate with prior studies highlighting demographic biases in AI-driven healthcare tools. For instance, Buolamwini and Gebru's "Gender Shades" project identified performance disparities in AI models based on demographic factors [2]. Additionally, Rengan et al. emphasized the exacerbation of health inequities due to insufficient demographic representation in training datasets [12]. This study extends these concerns to the application of LLMs in medical imaging, underscoring the pervasive nature of demographic biases across different AI architectures.

While the study provides valuable insights into demographic biases in GPT-4's diagnostic capabilities, several limitations must be acknowledged. The experiments utilized only 50 images per demographic subgroup, resulting in a relatively small sample size and limiting each image to a single analysis. This approach was chosen to stay within budget constraints, as the API calls required payment. Consequently, the small sample size may affect the statistical robustness and generalizability of the findings. Future studies should employ larger and more diverse datasets, as well as multiple tests per sample, to validate these results. Additionally, GPT-4 was not specifically trained for medical image analysis, which may have influenced its performance metrics. Specialized models trained explicitly on ophthalmic imaging tasks might exhibit different bias profiles. The BRSET dataset itself may have inherent biases related to the Brazilian population's specific demographics. These biases could have influenced the model's performance across different groups. Furthermore, the study focused only on Drusen detection. Expanding the analysis to include multiple ophthalmic conditions could provide a more comprehensive understanding of the model's bias landscape.

5 Conclusions

This study investigated demographic biases in OpenAI's GPT-4 model for Drusen detection using fundus images from the BRSET dataset. Through two focused experiments examining age and sex demographics, we found potential biases favoring older adults and female patients. Specifically, GPT-4 demonstrated higher diagnostic accuracy and recall in older adults and female patients, accompanied by increased FPR in these groups. In contrast, the model showed reduced recall in younger and male patients, indicating a risk of underdiagnosis and potential unequal healthcare outcomes.

These biases underscore the critical need for equitable AI healthcare solutions. If unaddressed, such biases can reinforce existing disparities in medical diagnostics, undermining the foundational goal of AI to enhance universal health equity. The study highlights the importance of early bias detection and mitigation strategies to ensure that AI models provide fair and reliable diagnostic support across all demographic groups.

Building upon these findings, future research should focus on expanding and diversifying datasets to enhance the model's generalizability and statistical robustness. Incorporating a broader range of demographic variables will provide a more comprehensive understanding of AI biases. Additionally, developing and training models specifically for ophthalmic imaging tasks can help reduce inherent biases. Implementing advanced bias mitigation strategies, such as adversarial training and conducting regular fairness audits, is essential for ensuring equitable performance across all demographic groups.

Addressing these biases is crucial for promoting equitable healthcare outcomes. Ensuring that AI models like GPT-4 perform fairly across all patient groups helps prevent existing disparities in medical diagnostics, thereby supporting the overarching goal of achieving universal health equity through technology.

References

1. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., Folk, J.C.: Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ digital medicine **1**(1), 39 (2018)
2. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
3. Corrado, G., Barral, J.: Advancing medical ai with med-gemini (May 2024), https://research.google/blog/advancing-medical-ai-with-med-gemini/, google Research Blog
4. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
5. Gichoya, J.W., McCoy, L.G., Celi, L.A., Ghassemi, M.: Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ health & care informatics **28**(1) (2021)
6. Kheiri, F., Rahnamayan, S., Makrehchi, M., Asilian Bidgoli, A.: Bias in histopathology datasets: A comprehensive investigation on possible factors. Research Square (June 2024), https://doi.org/10.21203/rs.3.rs-4559295/v1, pREPRINT (Version 1)
7. Liu, Y., Li, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Cui, L., Tu, Z., Wang, L., Zhou, L.: A systematic evaluation of gpt-4v's multimodal capability for chest x-ray image analysis. Meta-Radiology p. 100099 (2024)
8. Nakayama, L.F., Goncalves, M., Zago Ribeiro, L., Santos, H., Ferraz, D., Malerbi, F., Celi, L.A., Regatieri, C.: A brazilian multilabel ophthalmological dataset (brset). PhysioNet **version 1.0.0** (2023), https://doi.org/10.13026/xcxw-8198
9. Nakayama, L.F., Matos, J., Quion, J., Novaes, F., Mitchell, W.G., Mwavu, R., Hung, C.J.Y.J., Santiago, A.P.D., Phanphruk, W., Cardoso, J.S., et al.: Unmasking biases and navigating pitfalls in the ophthalmic artificial intelligence lifecycle: A narrative review. PLOS Digital Health **3**(10), e0000618 (2024)
10. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM conference on health, inference, and learning. pp. 151–159 (2020)

11. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342
12. Rengan, V., Lalwani, D., Bhat, S., Sundaram, P.M.: Enhancing dataset quality for ai in radiology: Challenges and solutions. Journal of Gastrointestinal and Abdominal Radiology (2024)
13. Saw, S.N., Ng, K.H.: Current challenges of implementing artificial intelligence in medical imaging. Physica Medica **100**, 12–17 (2022)
14. Stanley, E.A., Souza, R., Winder, A.J., Gulve, V., Amador, K., Wilms, M., Forkert, N.D.: Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. Journal of the American Medical Informatics Association **31**(11), 2613–2621 (2024)
15. Ting, D.S.W., Pasquale, L.R., Peng, L., Campbell, J.P., Lee, A.Y., Raman, R., Tan, G.S.W., Schmetterer, L., Keane, P.A., Wong, T.Y.: Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology **103**(2), 167–175 (2019)
16. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine **25**(1), 44–56 (2019)
17. Williams, M.A., Craig, D., Passmore, P., Silvestri, G.: Retinal drusen: harbingers of age, safe havens for trouble. Age and ageing **38**(6), 648–654 (2009)
18. Xu, P., Chen, X., Zhao, Z., Shi, D.: Unveiling the clinical incapabilities: a benchmarking study of gpt-4v (ision) for ophthalmic multimodal image analysis. British Journal of Ophthalmology (2024)