This is an Accepted Manuscript of an article published in *NEJM AI* on 13th August 2024, available at: https://doi.org/10.1056/AIoa2400353.

Methodology for independent evaluation of algorithms for automated analysis of medical images for trustworthy and equitable deployment of clinical AI in diverse population screening programmes

Jiri Fajtl[1], Roshan A Welikala[1], Sarah Barman[1], Ryan Chambers[2], Louis Bolter[2], John Anderson[2], Abraham Olvera-Barrios[3], Royce Shakespeare[4], Catherine Egan[3], Christopher G. Owen[4], Adnan Tufail[3], Alicja R. Rudnicka[4]

On behalf of the ARIAS Research Group

1. School of Computer Science and Mathematics, Kingston University, London, UK
2. Diabetes and Endocrinology, Homerton Healthcare NHS Foundation Trust, London, UK
3. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK
4. Population Health Research Institute, St George's University of London, London, UK

Author for correspondence: Professor Alicja R Rudnicka, email: arudnick@sgul.ac.uk

Abstract

Background: Deployment of algorithms in healthcare screening programmes has been hindered by a lack of agreed methodology to evaluate trustworthiness and equity. We outline transferable methodology for independent evaluation of algorithms, using a routine, high volume, multi-ethnic national diabetic eye screening programme as an exemplar. Automated retinal image analysis systems (ARIAS), including artificial intelligence (AI), for detection of diabetic retinopathy (DR), could substantially increase image-grading capacity.

Methods: Twenty-five vendors with or pending CE Class IIa ARIAS for DR detection from retinal images were invited. Sample data (6,268 images) were provided to confirm ARIAS outputs could be replicated in a trusted research environment. Consecutive routine screening encounters between 1st January 2021 to 31st December 2022 at the North East London Diabetic Eye Screening Programme were curated for evaluation. Sample size calculations focussed on precision for detection of severe DR by population subgroups, particularly ethnicity. Vendor algorithms did not have access to human grading data or other metadata during image processing. We report technical and operational considerations relevant to implementation and evaluation in large scale population screening.

Results: 8/25 eligible vendors participated. In total, 202,886 encounters were evaluated (1.2 million images) from 32% white, 17% black and 39% South Asian ethnic groups, including ~25,000 cases requiring referral to ophthalmology for review/treatment. Image resolutions varied from 150x300 to 6000x4000 pixels. Time from study invitation to ARIAS installation and algorithm verification ranged from 96 to 460 days; image processing required between 13.5 hours to 105 days.

Conclusions: This paper provides the framework for transparent, equitable, robust and trustworthy evaluation of clinical AI in screening, to inform standards in healthcare prior to deployment. This multiple-vendor study compared ARIAS, at scale, on a range of images with different characteristics from a population of different ethnicities, wide age range, levels of deprivation and spectrum of DR.

(Abstract - 299 words)

(Manuscript text 2989 words excluding abstract, tables & references)

Introduction

Diabetes mellitus (DM) is a rising global burden, affecting 1 in 10 adults globally. Worldwide rates have quadrupled over the past two decades, with approximately 537 million currently affected and is projected to rise to 784 million in 2045.[1] Diabetes related health costs are substantial[2 3] with the majority being spent on complications. Diabetic retinopathy (DR) remains a leading cause of blindness among the working age population.[4 5] Early detection through annual screening for referrable DR and treatment can prevent or delay sight loss.[4] The English National Health Service (NHS) Diabetic Eye Screening Programme (DESP) faces challenges due to rising diabetes prevalence,[6 7] requiring manual grading of ~13 million images each year. Automated retinal image analysis systems(ARIAS), that usually utilise artificial intelligence (AI) approaches, provide alternative technologies to detect those with medium-high risk of developing sight threatening DR, [8-13] resulting in substantially expanding grading capacity in screening programmes.[8-10]

Vendor-led studies typically provide over-optimistic estimates when compared with independent evaluations of the same ARIAS in different clinical settings/computational environments.[11] To address this issue, comparisons between multiple ARIAS need to be undertaken on the same dataset, using a consistent computational platform that as closely as possible reflects the conditions of the real world where they will be deployed. Furthermore, it is necessary to assess algorithmic fairness across diverse population subgroups before deploying algorithms in the intended healthcare settings.[14-16] Direct ARIAS performance comparisons are hampered by single ARIAS evaluations on different populations (often lacking diversity), limited sample size, unspecified pre-selection or pre-processing of retinal images, image capture systems, grading protocols and reference standards.[13 17]

There is a need for vendor independent head-to-head comparisons of AI systems on large, real-life, population data to provide impartial, direct, precise comparisons of different ARIAS across diverse population subgroups. Using the largest, most ethnically diverse North-East London NHS-DESP we describe a sustainable platform for independent evaluation of state-of-the art ARIAS (including AI systems) licenced as a medical device, i.e., with FDA or CE Class IIa certification. We describe transferable principles and methodology that could be adopted for other disease or healthcare settings to ensure algorithms continue to meet pre-defined standards across population groups prior to impact on patient care pathways.

Methods

To ensure this evaluation was transparent, objective, and free from data distortion and corruption at each stage, we adhered to the following generalisable principles:

- Engage with vendors equitably to maintain study result integrity
- Evaluate under a consistent computational environment
- Evaluate on realistic data corresponding to real-life deployment
- Safeguard patient privacy and protect the curated data

Our use case was based on the evaluation of CE-marked ARIAS for DR detection prior to commissioning for deployment in NHS-DESP. Although UKCA registered ARIAS would have been eligible for inclusion in our study, UKCA registration was not fully operational during the study period. Our research team is independent of any conflict of interest in relation to any ARIAS provider.

Identification of CE marked ARIAS

There is no open database to search for potential CE marked ARIAS vendors. We identified ARIAS vendors by combining information from our previous studies,[8-10 18] a recent review commissioned by the UK National Screening Committee,[13]presentation of our project at conferences and communication with stakeholders/colleagues in the field. We also searched various medical device databases with limited access functionality, including EUDAMED[19] and FDA.[20] Out of 54 ARIAS vendors, 22 were identified as potentially eligible CE Class IIa ARIAS for DR detection from retinal images and were invited to participate in our study. Supplemental Appendix A details enrolment processes.

Population setting - data sources and curation

The North East London DESP located at the Homerton Healthcare NHS Foundation Trust was the location for this large-scale evaluation. This screening centre offers annual screening to over 100,000 people living with diabetes aged 12 years or older and uniquely contains a high proportion of patients from different ethnic groups (with high representation of white, black and South Asians) in one of the most deprived regions in the country, with a spectrum of diabetic eye disease and a wide age range.[10 21]

Retinal image capture for screening encounters followed National Screening Committee protocols.[22 23] A range of fundus cameras were used including, for example, Canon 2CR (Canon, Tokyo, Japan),

Canon 2CR-Dgi (Canon), Canon EOS (Canon), TOPCONnect (Topcon, Tokyo, Japan) cameras. Additional retinal images were captured to ensure images of sufficient quality for grading or to document other/peripheral pathology. Non-retinal images (e.g., crystalline lens, eyelids, hands) to document anterior segment pathology or confirm camera functioning when the patient could not be photographed were also taken as needed. Retinal images and metadata were recorded in an OptoMize (NEC Software Solutions UK Limited) database at the North East London DESP in accordance with NHS screening and grading protocols.[22 24-26]

Test dataset of retinal images

A dataset of 1000 expired encounters (6268 images) with pseudonymised encounter IDs, encompassing the range of images captured within the North East London DESP (including non-retinal images, high and low quality retinal images) was shared with the ARIAS vendors to test the required software Application Programme Interface (API) for compatibility and confirm their ARIAS could process the range of images generated in a typical screening programme. It was also used to confirm outputs obtained by vendors aligned with that obtained by researchers within their respective computational environments. No associated meta-data or human grading data were provided to vendors. This test dataset was based on encounters collected from an earlier time period, before the evaluation dataset, from individuals that are not present in the evaluation dataset.

Evaluation dataset of retinal images

The evaluation dataset of images was curated from circa 200,000 consecutive screening episodes (~1.2 million images) between 1st January 2021 and 31st December 2022, with patient IDs pseudonymised. Images were exported in jpeg format (as stored in the OptoMize database) for the evaluation from all screening pathways that contained images. All personal data were stripped from this evaluation dataset in accordance with data protection. These data (or any prior data from this screening centre) have not been used to develop any ARIAS.

Curated dataset of individual characteristics and DR grading data

Data on human grading outcomes and associated sociodemographic data were exported by the Homerton clinical care team from the OptoMize database using the same pseudonymised ID used in the evaluation dataset, for screening encounters between 1st January 2021 and 31st December 2022. These data were exported to a separate location and not shared with vendors. Data extracted additionally included routinely recorded information, such as age, sex, type and duration of diabetes,

self-reported ethnicity and index of multiple deprivation which is a small-area measure of relative deprivation across the UK.[27] These data were subsequently linked using pseudonymized IDs to the ARIAS post-processing outputs on the image evaluation dataset.

Trusted Research Environment (TRE)

A TRE was set up at the Homerton Healthcare NHS Foundation Trust to host the necessary hardware and datasets. This included obtaining approval for the TRE, procuring the server, setting up remote access to the TRE, ensuring a high-security level by contracting independent security assessment of the server and network audit, and putting in place operational procedures for server monitoring, TRE remote access, user access management, data transfer between the research team and the TRE, and server decommissioning at the end of study. To obtain the necessary approvals for the TRE, as per the Data Sharing Framework in North East London, a Data Protection Impact Assessment (DPIA) was submitted and approved by the NHS North East London Information Governance Steering Group (IGSG) and Data Access Group. Supplemental Appendix B provides additional technical specifications.

Evaluation protocol

Vendors enrolled in the study were provided with instructions on ARIAS software packaging and data formats (Supplemental Appendix C). The evaluation proceeded as follows:

1. Vendors signed a Non-Disclosure Agreement (NDA) restricting the utilization of the test dataset of images for ARIAS testing only, with a mandatory deletion upon study completion.
2. Encrypted test dataset was shared with the vendors over a per-vendor password protected cloud space. The test dataset contained images and pseudonymised encounter IDs only.
3. Vendors were tasked with packaging their software into a Linux Docker container and implementing a CSV file-based input and output API.
4. ARIAS software received from vendors. Software NDA signed if requested by vendor.
5. ARIAS functionality and data formats were tested on TRE server on the test dataset. Issues were reported to vendors requesting rectification, returning to step 3 above.
6. ARIAS passing the runtime test were tested for inference consistency between the server and the vendor's runtime by comparing the test dataset outcomes. Our project team and the vendor exchanged output results to verify inference consistency.

7. Vendors whose ARIAS passed the consistency check proceeded to the main evaluation. In case of test failure, vendors addressed the issues and provided an updated ARIAS, returning to step 3.
8. ARIAS whose correct installation was confirmed by vendors was placed into a queue to process the evaluation dataset.

We processed the evaluation dataset through each ARIAS in turn to avoid potential performance issues from sharing data/TRE resources. Figure 1 outlines the ARIAS evaluation steps.

Sample size

To assess equity in ARIAS performance by population subgroups, there is a need to provide similar certainty for estimates of performance across sub-groups of the population. Sample size calculations considered precision, as defined by the 95% confidence interval (CI), of detection rates for the rarest and most serious DR, proliferative DR (R3 as per National Screening Committee grading criteria[24]) within population subgroups of age, sex, ethnicity and quintiles of index of multiple deprivation. Since one of the smallest subgroups were ethnicity, we placed emphasis on powering the study to give similar level of ***precision*** (i.e. width of 95% CI) on estimates of detection for the rarest and most serious DR (R3) by subgroups of ethnicity. Table 1 provides 95% confidence intervals for a range of hypothesized detection rates (sensitivities) from 90% to 100% by the number of R3 cases set to 600, 300 or 100 cases. Given the known prevalence of the three main ethnic groups at the screening centre [10 21] and prevalence of proliferative DR in each group (0.5-1%),[9 10] curating 100,000 consecutive encounters would have resulted in approximately 200-300 cases of proliferative DR in each of white and South Asian subgroups but only 100-150 in Black-African-Caribbean, resulting in up to 5 percentage points variation in the lower bound of the 95%CI. However, with 200,000 encounters the number of proliferative DR increases to approximately 600 in whites and South Asians and 300 in Black-African-Caribbean and the lower bound of the 95%CIs for detection would differ by approximately one percentage point across ethnic group (Table 1). A priori we agreed that a 1.0 percentage point difference or less in the lower bound of the 95% confidence interval was ***equitable*** for precision across subgroups of ethnicity for the rarest and most serious DR. Hence target sample size was set to 200,000 encounters. It follows that precision for more common DR outcomes will also be equitable across groups. Algorithmic fairness will be examined by comparing systematic differences between the point estimates of test performance across subgroups of the population.

Reference standards

NHS-DESP retinal images were assessed by up to three trained human graders (primary, secondary and tertiary) following national standards.[24 28] Grading classifications hierarchy are, no observable retinopathy (R0), mild non-proliferative retinopathy (R1), no observable maculopathy or non-referable maculopathy (M0), ungradable images (U), moderate-severe non-proliferative retinopathy (R2), referable maculopathy (M1) and proliferative retinopathy (R3). In the English NHS-DESP retinopathy grades R0M0, R1M0 are non-referable retinopathy and grades R2, M1, R3 are referable retinopathy. The commensurate Early Treatment Diabetic Retinopathy Study (ETDRS) retinopathy grade scores are; R0 equivalent to "no apparent retinopathy"; R1 ETDRS scores 20-35 inclusive; R2 ETDRS scores 43-53 inclusive; R3 ETDRS scores 61+.[24 28 29] The final human grade in the worst eye served as the reference standard (representing current clinical practice) to confirm required sample size by ethnic subgroup (Supplemental Appendix D outlines ARIAS outputs sought).

Planned statistical analysis

A future publication will examine ARIAS test performance by population subgroups as compared with human graders. Here we outline our a-priori statistical analyses plan. ARIAS outputs from evaluation data will be merged with the curated human grading outcomes and sociodemographic data using the pseudonymised ID for linkage. ARIAS outcome will be defined as test positive (classified as DR present and ungradable/technical failure) or test negative (classified by ARIAS as DR absent). ARIAS sensitivity (detection rate), false positive rates (i.e. 100% - specificity as %), positive predictive values, negative predictive values and likelihood ratios with corresponding 95% confidence intervals (logit-transformed 95% confidence intervals or binomial exact confidence intervals in the presence of values of 100%) will be estimated overall and by self-described ethnicity (white, black, South Asian, Other/Unknown), age groups (<30, 30 to <45, 45 to <60,60 to <75, and ≥75 years), sex, and quintiles of index of multiple deprivation for each DR grade and for combined grades of referable DR (R2, M1 and R3) and non-referable DR (R0 and R1). Multiple variable logistic regression models will quantify the strength of evidence for statistical heterogeneity in ARIAS test positive vs test negative outcomes by subgroups of the population for each DR grade. In view of the sample size and number of statistical significance tests, the p-value will be <0.0001 for tests of heterogeneity. ARIAS ungradable/non-assessable rates will be examined by the same population subgroups.

.

Results

Invites were sent to 22 ARIAS vendors on May 26th 2022, of which five responded to full participation and three additional eligible vendors subsequently joined the study before the enrolment cutoff (Figure 2). Eight vendors participated and their enrolment times ranged from one to 69 days, the software preparation phase averaged 121 days, with time to software delivery ranging from 48 to 269 days (Figure 3a). Verification of the test dataset, including rectification of technical or operational issues, took on average 101 days, with the shortest being 8 and longest 218 days (Figure 3a). The ARIAS in Linux Docker images ranged in size from 6GB to 90GB. As per the vendors' description, all ARIAS algorithms utilised neural networks (AI), were capable of offline (local with no internet connectivity) and cloud execution and can leverage GPU accelerators.

The evaluation dataset contained 202,886 encounters. The number of images per encounter ranged from 1 to 64, with the most common being 6, followed by 4 images per encounter (Figure 3b). Up to three and above ten images per encounter were typically outliers, either being non-retinal images, test images or repeated retinal images to check camera set-up. The evaluation dataset comprised 1,175,423 images, taking up 3TB of storage. Image resolutions varied from 150x300 to 6000x4000 pixels, with 2736x1824 being the most common. File sizes ranged from 142KB to 3.8MB (Figure 3c). The test dataset had similar characteristics with 1,000 encounters and a total of 6,268 images, occupying 9GB.

Time for ARIAS to process the evaluation dataset in the TRE ranged from 13.5 hours to 105 days, equivalent to 240 milliseconds to 45 seconds per encounter. Leveraging the processing capabilities of the server, the evaluation was expedited by parallelizing the execution of certain ARIAS algorithms that were not utilizing the server's resources. Parallelization involved breaking down the evaluation dataset into smaller batches and running multiple instances of the same algorithm concurrently, reducing the longest runtime from 105 days to 15 days. With parallelization, the total processing time for all ARIAS algorithms was on average 47 days, 18 hours. However, multiple re-runs were required during testing to resolve ARIAS execution issues.

The DESP population characteristics overall and by ethnic group for the 202,886 encounters are provided in Table 2. The target sample size for the number of proliferative DR cases within each ethnic group was achieved, with near 25,000 cases of referable DR.

Discussion

We have outlined our platform and methodology for an independent evaluation of AI systems for screening medical images with a specific use case for DR screening where most CE marked algorithms exist. Our use case is ARIAS for use in the NHS-DESP. We provide details of the sample size calculations needed for equitable precision in test performance metrics across population subgroups, with a statistical analysis plan that is transparent and robust. As far as the authors are aware, this is the largest evaluation to date, including over 200,000 encounters, ~1.2 million images with near 25,000 cases of referable DR. Importantly, there is good representation across population subgroups, including ethnicity, age, levels of deprivation, with a spectrum of diabetic eye disease. Our platform can provide updated information on ARIAS performance at scale within a short time frame, as well as providing valuable feedback to vendors where an algorithm might benefit from improvement. We placed importance on building good working relationships with vendors, by providing clear and transparent information about the aims, purpose, and process of the evaluation. In contrast to previous work,[11] our goal is for open label publishing of ARIAS performance, in accordance with our earlier work[8-10]. To ensure the same communications with all vendors, we replicated email invitations to all vendors, and shared an anonymous Q&A document addressing vendors' queries prior to formal enrolment. To maximise participation, we aided with software installation and execution as needed.

However, the study faced several challenges, primarily stemming from the unpredictable timelines associated with ARIAS software delivery and bug fixes. The absence of a standardized API for ARIAS, coupled with the requirement for offline ARIAS execution to comply with NHS data governance standards, necessitated additional effort to devise and implement proprietary solutions. The introduction of a standard API and the adoption of an established computation platform, such as a GDPR-compliant cloud-based solution would have simplified testing and deployment, reducing costs and researcher time associated with the evaluation. The set-up of the TRE environment locally with intricate setup and mandated security procedures, including remote access via a proxy server and restrictions on file transfers on the server, introduced substantial time overheads to the evaluation process. Several ARIAS vendors encountered difficulties when adapting their cloud-based ARIAS to run offline, leading to multiple cycles of bug tracking and fixing. Moreover, some vendors indicated that their ARIAS could only process images with a resolution above 1024x1024 pixels, and the evaluation dataset contained images with lower resolutions but that did not seem to affect ARIAS processing encounters. However, this raises the need to reopen discussion around standardization of

image capture formats within the DESP.[30 31] Moreover, this study did not include comparative cost effectiveness of ARIAS approaches, which should be included in future evaluations.

We believe conducting ARIAS evaluations in a cloud-hosted TRE would have avoided many of the functionality issues encountered during this study. Cloud-based TRE approaches could allow vendors to develop and test on a readily accessible platform, while allowing fast and flexible remote access by research team. This would be a future proof platform solution for on-going evaluations, which are needed given rapid developments of ARIAS, as well as providing a readily accessible 'real life' test bed for potential deployment.

We believe the approach outlined here not only provides a model for evaluation of ARIAS for diabetic eye screening, but for AI use in other healthcare imaging domains, providing governmental, NHS, lay and healthcare provider stakeholders an exemplar of equitable methodology of clinical AI prior to moving to the next stage of implementation and commissioning. Our approach aligns with a recent independent governmental review on equity in medical devices that was triggered by the poorer performance of pulse oximeters in darker skinned individuals.[16] The report concluded that AI medical devices should be tested in the eventual deployment setting and in the real-world context in which such devices are to operate. Although this study was based at one screening centre, it is one of the largest NHS screening centres, serving one of the most diverse and deprived populations in the country. Key features of our study included multiple vendor participation, large appropriately powered, diverse, clinically relevant dataset using the same computational platform and executed by a vendor-neutral research team. These study design features support the generalizability of our findings to other screening sites. We believe having a research team that is independent of any commercial interests working with commercial stakeholders, can encourage investment in health service provision and together with our public engagement activities[32] lead to trust in innovation and technological advance.

Ethical approval

All data were managed in accordance with UK NHS information governance requirements and adhered to the principles of the Data Protection Act 2018. The methodologies adhered to the Homerton Healthcare NHS Foundation Trust Security Policy and Information Governance Policy, and GDPR principles relating to processing of personal data. The study has been approved by the Health Research Authority (IRAS project ID 265637). The NHS Health Research Authority toolkit (http://www.hra- decisiontools.org.uk/ethics/) identified that Research Ethics Approval was not required for this project as all data were pseudonymized and with findings presented in aggregate form. A Data Protection Impact Assessment was submitted to Homerton Healthcare NHS Foundation Trust Information Governance Lead and approved by Information Governance Team in December 2022.

Collaborators:

The Artificial Intelligence / Automated Retinal Image Analysis Systems (ARIAS) Research Group

John Anderson, Sarah Barman, Louis Bolter, Ryan Chambers, Lakshmi Chandrasekaran, Umar Chaudhry, Cathy Egan, Jiri Fajtl, Aaron Lee, Abraham Olvera-Barrios, Christopher G Owen, Paolo Remagnino, Alicja R Rudnicka, Adnan Tufail, Charlotte Wahlich, Roshan Welikala, Kathryn Willis

Acknowledgements: We wish to thank Sean Devine and Asif Mirza from the Homerton Healthcare NHS Foundation Trust IT department for their support throughout this project. We thank our Study Advisory Group members: Rosalind Given-Wilson, Alastair Denniston, Kevin Dunbar, Samantha Mann, Fiona Martin.

Funding This work was funded by NHS Transformation Directorate and The Health Foundation and managed by the National Institute for Health and Social Care Research (AI_HI200008). "*Ethnic differences in performance and perceptions of Artificial Intelligence retinal image analysis systems for the detection of diabetic retinopathy in the NHS Diabetic Eye Screening Programme"* https://fundingawards.nihr.ac.uk/award/AI_HI200008].

Disclaimer The views expressed in this publication are those of the author(s) and not necessarily those of the NHS Transformation Directorate, The Health Foundation, National Institute for Health Research, or the Department of Health and Social Care

Competing interests None to declare. None of the authors have any CoI to declare in relation to any ARIAS provider.

Data sharing statement: Data will be shared in grouped format in online supplements by population subgroups for each ARIAS by level of diabetic eye disease. In compliance with our Data Protection Impact Assessment, sharing of individual patient data is not possible and all data is stored by the data controller, the Homerton Healthcare NHS Foundation Trust.

References

1. Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022;183:109119. doi: 10.1016/j.diabres.2021.109119 [published Online First: 2021/12/10]
2. Hex N, Bartlett C, Wright D, et al. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet Med* 2012;29(7):855-62. doi: 10.1111/j.1464-5491.2012.03698.x
3. Bellemo V, Lim G, Rim TH, et al. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. *Curr Diab Rep* 2019;19(9):72. doi: 10.1007/s11892-019-1189-3
4. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet* 2010;376(9735):124-36. doi: 10.1016/S0140-6736(09)62124-3
5. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16–64 years), 1999–2000 with 2009–2010. *BMJ Open* 2014;4(2):e004015. doi: 10.1136/bmjopen-2013-004015
6. Pham TM, Carpenter JR, Morris TP, et al. Ethnic Differences in the Prevalence of Type 2 Diabetes Diagnoses in the UK: Cross-Sectional Analysis of the Health Improvement Network Primary Care Database. *Clin Epidemiol* 2019;11:1081-88. doi: 10.2147/CLEP.S227621
7. Sivaprasad S, Gupta B, Gulliford MC, et al. Ethnic variations in the prevalence of diabetic retinopathy in people with diabetes attending screening in the United Kingdom (DRIVE UK). *PLoS One* 2012;7(3):e32182. doi: 10.1371/journal.pone.0032182
8. Tufail A, Kapetanakis VV, Salas-Vega S, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. *Health Technol Assess* 2016;20(92):1-72. doi: 10.3310/hta20920
9. Tufail A, Rudisill C, Egan C, et al. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology* 2017;124(3):343-51. doi: 10.1016/j.ophtha.2016.11.014

10. Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *British Journal of Ophthalmology* 2021;105(5):723-28. doi: 10.1136/bjophthalmol-2020-316594
11. Lee AY, Yanagihara RT, Lee CS, et al. Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. *Diabetes Care* 2021;44:1168-75. doi: 10.2337/dc20-1877
12. Bhaskaranand M, Ramachandra C, Bhat S, et al. The Value of Automated Diabetic Retinopathy Screening with the EyeArt System: A Study of More Than 100,000 Consecutive Encounters from People with Diabetes. *Diabetes Technol Ther* 2019;21:635-43. doi: 10.1089/dia.2019.0164
13. Zhelev Z, Peters J, Rogers M, et al. Automated grading in the Diabetic Eye Screening Programme: External review against programme appraisal criteria for the UK National Screening Committee In: Committee UNS, ed., 2021.
14. Noor P. Can we trust AI not to further embed racial bias and prejudice? *BMJ* 2020;368:m363. doi: 10.1136/bmj.m363
15. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447-53. doi: 10.1126/science.aax2342
16. Department of Health and Social Care. Equity in medical devices: independent review - final report 2024 [Available from: https://www.gov.uk/government/publications/equity-in-medical-devices-independent-review-final-report.
17. Rajesh AE, Davidson OQ, Lee CS, et al. Artificial Intelligence and Diabetic Retinopathy: AI Framework, Prospective Studies, Head-to-head Validation, and Cost-effectiveness. *Diabetes Care* 2023;46(10):1728-39. doi: 10.2337/dci23-0032
18. Lee A, Taylor P, Kalpathy-Cramer J, et al. Machine Learning Has Arrived! *Ophthalmology* 2017;124(12):1726-28. doi: 10.1016/j.ophtha.2017.08.046
19. Safety D-GfHaF. Medical Devices - EUDAMED: European Commission - Public Health; 2022 [Available from: https://health.ec.europa.eu/medical-devices-eudamed_en accessed last accessed 20th June 2022 2022.
20. Administration USFD. Medical Device Databases: USA.gov; 2022 [Available from: https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/medical-device-databases accessed 25th June 2022 2022.
21. Olvera-Barrios A, Owen CG, Anderson J, et al. Ethnic disparities in progression rates for sight-threatening diabetic retinopathy in diabetic eye screening: a population-based retrospective cohort study. *BMJ Open Diabetes Res Care* 2023;11(6) doi: 10.1136/bmjdrc-2023-003683 [published Online First: 2023/11/11]
22. NHS England. Diabetic eye screening: guidance on camera approval 2023 [Available from: https://www.gov.uk/government/publications/diabetic-eye-screening-approved-cameras-and-settings/diabetic-eye-screening-guidance-on-camera-approval accessed 20th November 2023 2023.
23. Public Health England. Diabetic eye screening: guidance when adequate images cannot be taken 2021 [Available from: https://www.gov.uk/government/publications/diabetic-eye-screening-pathway-for-images-and-where-images-cannot-be-taken/diabetic-eye-screening-guidance-when-adequate-images-cannot-be-taken.
24. Public Health England. NHS Diabetic Eye Screening Programme: grading definitions for referable disease 2021 [Available from: https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria/nhs-diabetic-eye-screening-programme-grading-definitions-for-referable-disease accessed 20th February 2024.
25. Public Health England. Diabetic eye screening pathways: patient, grading, referral, surveillance 2023 [Available from: https://www.gov.uk/government/publications/diabetic-eye-screening-pathways-patient-grading-referral-surveillance.
26. Public Health England. NHS Screening Programmes in England 1 April 2017 to 31 March 2018: PHE publications gateway number: GW-243; 2019 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/783537/NHS_Screening_Programmes_in_England_2017_to_2018_final.pdf [Accessed February 2023].

27. Ministry of Housing CLG. English indices of deprivation 2019: Crown copyright, 2019; 2019 [Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.
28. Public Health England. Diabetic eye screening: retinal image grading criteria 2021 [Available from: https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria accessed 23 November 2023.
29. Group ETDRSR. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* 1991;98(5 Suppl):823-33.
30. Lee AY, Campbell JP, Hwang TS, et al. Recommendations for Standardization of Images in Ophthalmology. *Ophthalmology* 2021;128(7):969-70. doi: 10.1016/j.ophtha.2021.03.003 [published Online First: 20210405]
31. Shweikh Y, Sekimitsu S, Boland MV, et al. The Growing Need for Ophthalmic Data Standardization. *Ophthalmol Sci* 2023;3(1):100262. doi: 10.1016/j.xops.2022.100262 [published Online First: 20221220]
32. Willis K, Chaudhry UAR, Chandrasekaran L, et al. What are the perceptions and concerns of people living with diabetes and National Health Service staff around the potential implementation of AI-assisted screening for diabetic eye disease? Development and validation of a survey for use in a secondary care screening setting. *BMJ Open* 2023;13(11):e075558. doi: 10.1136/bmjopen-2023-075558 [published Online First: 2023/11/16]

Table 1: Precision for detection rates between 90% to 100% according to sample size for number of proliferative diabetic retinopathy (R3)

Desired detection rate / Sensitivity	**95% confidence interval based on number with disease outcome of interest (R3)**		
N	**600**	**300**	**100**
90%	87.3% to 92.3%	86.0% to 93.2%	82.4% to 95.1%
92%	89.5% to 94.0%	88.3% to 94.8%	84.8% to 96.4%
94%	91.8% to 95.8%	90.7% to 96.4%	87.4% to 97.8%
96%	94.1% to 97.4%	93.1% to 97.9%	90.0% to 98.9%
98%	96.5% to 99.0%	95.7% to 99.3%	93.0% to 99.8%
100%	99.4% to 100%	98.8% to 100%	96.4% to 100%

N=number of active R3 DR cases. Binomial exact 95% confidence intervals. DR diabetic retinopathy

Table 2: Characteristics of screening encounters between 1st January 2021 to 31st Dec 2022

Characteristics	**Ethnic Group (column %)**				
	White N = 64446	**Black N = 34416**	**South Asian N = 78885**	**Other/Unknown N = 25100**	**Total N = 202847§**
Age at screening visit (years)	64.2(14.9)	61.4(13.7)	57.6(13.5)	59.5(13.7)	60.6(14.3)
Age Group n (%)					
<30 years	1752(2.7)	635(1.8)	1247(1.6)	592(2.4)	4226(2.1)
30 to < 45 years	4393(6.8)	2534(7.4)	12279(15.6)	2714(10.8)	21920(10.8)
45 to <60 years	15659(24.3)	12687(36.9)	29964(38.0)	8910(35.5)	67220(33.1)
60 to <75 years	26110(40.5)	12282(35.7)	26742(33.9)	9670(38.5)	74804(36.9)
75+ years	16532(25.7)	6278(18.2)	8653(11.0)	3214(12.8)	34677(17.1)
Female n (%)	28202(43.8)	18348(53.3)	37280(47.3)	11832(47.1)	95662(47.2)
Type of Diabetes n (%)					
Type 1	4845(7.5)	1033(3.0)	1296(1.6)	788(3.1)	7962(3.9)
Type 2	57658(89.5)	32343(94.0)	75963(96.3)	23533(93.8)	189497(93.4)
Other/MODY	184(0.3)	89(0.3)	211(0.3)	83(0.3)	567(0.3)
Not specified/Missing	1759(2.7)	951(2.8)	1415(1.8)	696(2.8)	4821(2.4)
Duration of Diabetes n (%)					
< 10 years	37210(57.7)	19844(57.7)	44857(56.9)	15201(60.6)	117112(57.7)
10 to < 20 years	20082(31.2)	10524(30.6)	24296(30.8)	7228(28.8)	62130(30.6)
20+ years	7146(11.1)	4043(11.7)	9732(12.3)	2659(10.6)	23580(11.6)
Missing	8(0.01)	5(0.01)	0(0.0)	12(0.05)	25(0.01)
Visual acuity in worst eye n (%)					
At least 6/6	28475(44.2)	15526(45.1)	38517(48.8)	12261(48.8)	94779(46.7)
<6/6 to 6/9	22443(34.8)	12393(36.0)	27202(34.5)	8750(34.9)	70788(34.9)
<6/9 to 6/18	8402(13.0)	3835(11.1)	8621(10.9)	2578(10.3)	23436(11.6)
Worse than 6/18	2811(4.4)	1327(3.9)	2530(3.2)	787(3.1)	7455(3.7)
CF	402(0.6)	177(0.5)	285(0.4)	126(0.5)	990(0.5)
HM	902(1.4)	461(1.3)	729(0.9)	258(1.0)	2350(1.2)
PL	403(0.6)	243(0.7)	374(0.5)	128(0.5)	1148(0.6)
NPL	316(0.5)	254(0.7)	281(0.4)	109(0.4)	960(0.5)
Missing eye	61(0.1)	48(0.1)	55(0.1)	18(0.1)	182(0.1)
Not measured/missing	231(0.4)	152(0.4)	291(0.4)	85(0.3)	759(0.4)
IMD Quintile n (%)					
1 (most deprived)	19600(30.4)	16401(47.7)	30142(38.2)	8064(32.1)	74207(36.6)
2	10514(16.3)	7255(21.1)	13727(17.4)	4611(18.4)	36107(17.8)
3	9191(14.3)	4773(13.9)	11614(14.7)	3545(14.1)	29123(14.4)
4	6788(10.5)	2683(7.8)	9720(12.3)	3289(13.1)	22480(11.1)
5 (least deprived)	18103(28.1)	3114(9.0)	13144(16.7)	5414(21.6)	39775(19.6)
missing	250(0.4)	190(0.6)	538(0.7)	177(0.7)	1155(0.6)
Total number of images processed*	368011	211715	448950	146372	1175048
Median no. of per encounter images (IQR)	6(4,6)	6(4,7)	6(4,6)	6(4,7)	6(4,6)
Final human DR grade in worst eye					
R0M0	44129(68.5)	21616(62.8)	50056(63.5)	16535(65.9)	132336(65.2)
R1M0	12172(18.9)	6596(19.2)	16205(20.5)	4848(19.3)	39821(19.6)
R1M1	4280(6.6)	3938(11.4)	8472(10.7)	2288(9.1)	18978(9.4)
R2	1104(1.7)	693(2.0)	1621(2.0)	537(2.2)	3955(1.9)
R2M0	304(0.5)	167(0.5)	408(0.5)	147(0.6)	1026(0.5)
R2M1	800(1.2)	526(1.5)	1213(1.5)	390(1.6)	2929(1.4)
R3	700(1.0)	397(1.2)	710(0.9)	262(1.0)	2069(1.0)
R3M0	353(0.5)	138(0.4)	273(0.3)	100(0.4)	864(0.4)
R3M1	347(0.5)	259(0.8)	437(0.6)	162(0.6)	1205(0.6)
Ungradable	2054(3.2)	1173(3.4)	1819(2.3)	630(2.5)	5676(2.8)
Missing**	7(<0.01)	3(<0.01)	2(<0.01)	0(0.0)	12(0.006%)

*Data given as mean (SD) or count (percent). IQR = interquartile range. 39 encounters identified to be camera test encounters were removed. DR=Diabetic retinopathy. *This includes both retinal and non-retinal images captured as part of routine screening. **Missing DR grades are for incomplete screening encounters where a subsequent completed screening encounter with DR grade is available. UK National Screening Committee human grading classifications hierarchy are, no observable retinopathy (R0), mild non-proliferative retinopathy (R1), no observable maculopathy or non-referable maculopathy (M0), ungradable images (U), moderate-severe non-proliferative retinopathy (R2), referable maculopathy (M1) and proliferative retinopathy (R3). § 39 encounters were fundus camera checks and are excluded from the table*

List of Figure legends

Figure 1. Sequence of ARIAS evaluation steps. Items shaded in purple define activities undertaken by vendors. Items shaded in blue define activities undertaken by the research team. Items shaded in green define activities undertaken by the research team within the Trust Research Environment.

Footnote: API = Application Programming Interface

ARIAS = Automated Retinal Image Analysis System

CRI = Curated Retinal Image dataset (the evaluation dataset)

NDA = Non-Disclosure Agreement

NEL DESP = North East London Diabetic Eye Screening Programme

PT = project team server housed in Trusted Research Environment

TRE = Trusted Research Environment at Homerton Healthcare NHS Foundation Trust

SRI = Sample Retinal Image dataset (the test dataset)

Note – 'Results for data analysis' were the pseudonymised encounter ID ARIAS outputs only, and were exported to the data analysis team at St George's University of London.

Figure 2. Enrolment process of ARIAS vendors from identification of potential candidate ARIAS to the final ARIAS that proceeded with the clinical evaluation.

*Reasons given for withdrawal: not confident their ARIAS would perform well on our data, not able to deliver ARIAS within 6-month time window, Algorithm does not produce diabetic retinopathy specific outcomes. NHS = National health Service; DESP = Diabetic Eye Screening Programme

Figure 3. a) time taken (in days) for vendors to enrol in the study from initial invitation letter (blue), prepare their software for evaluation (orange) and rectify software verification issues on the test dataset (grey). b) Histogram of images per encounter in the evaluation dataset. c) Histogram of image resolutions in the evaluation dataset (the lowest image resolutions are infrequent, <0.01%, and represent thumbnails duplicates of higher resolution images for the same encounter)

Figure 1

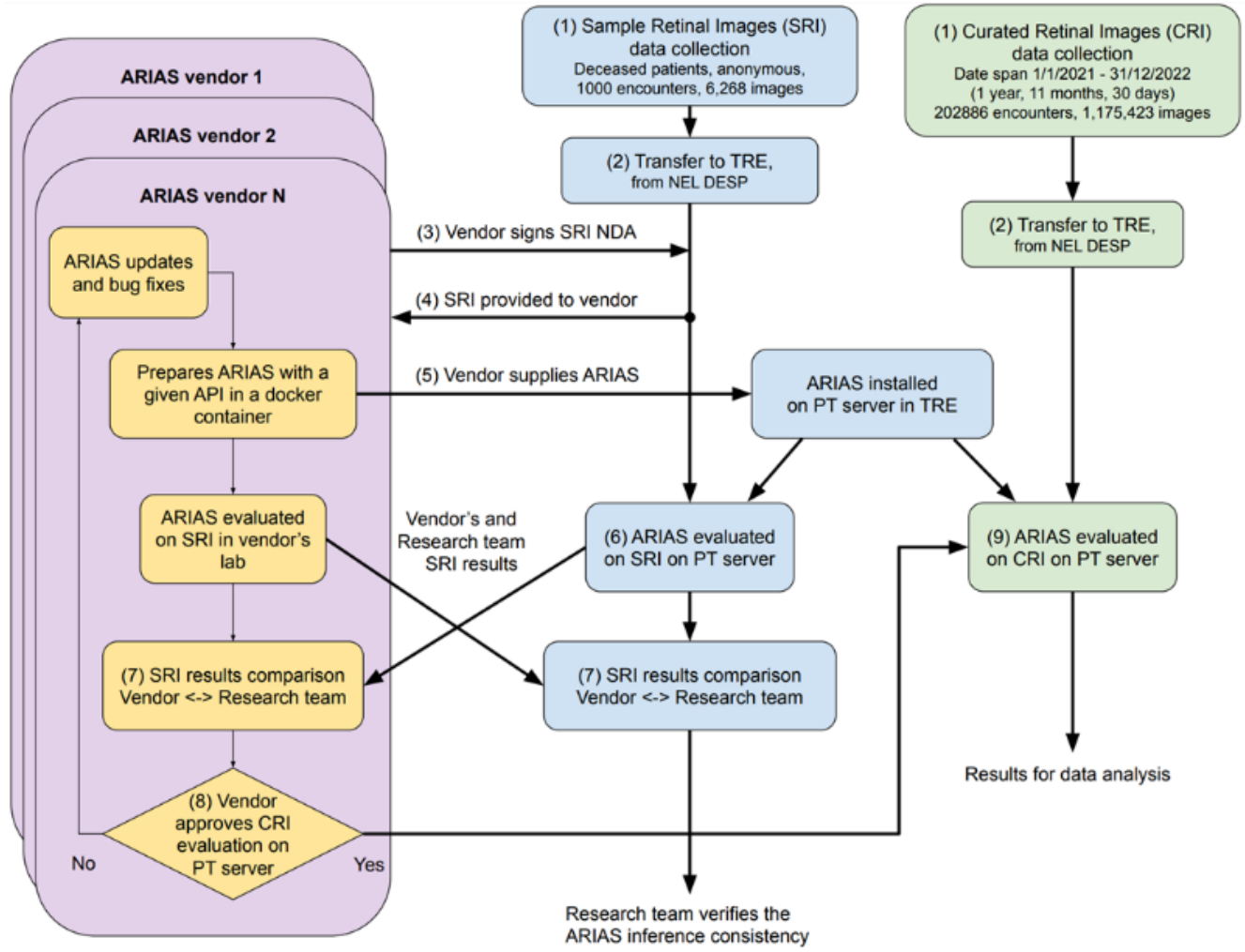
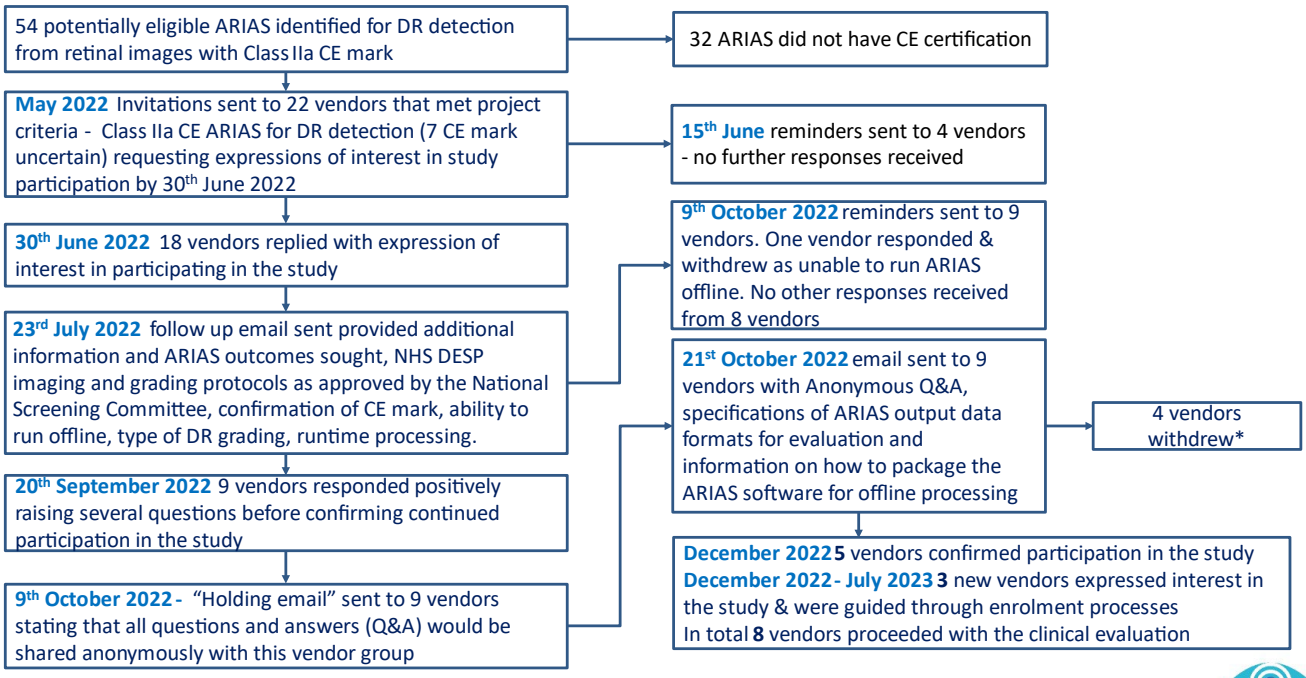
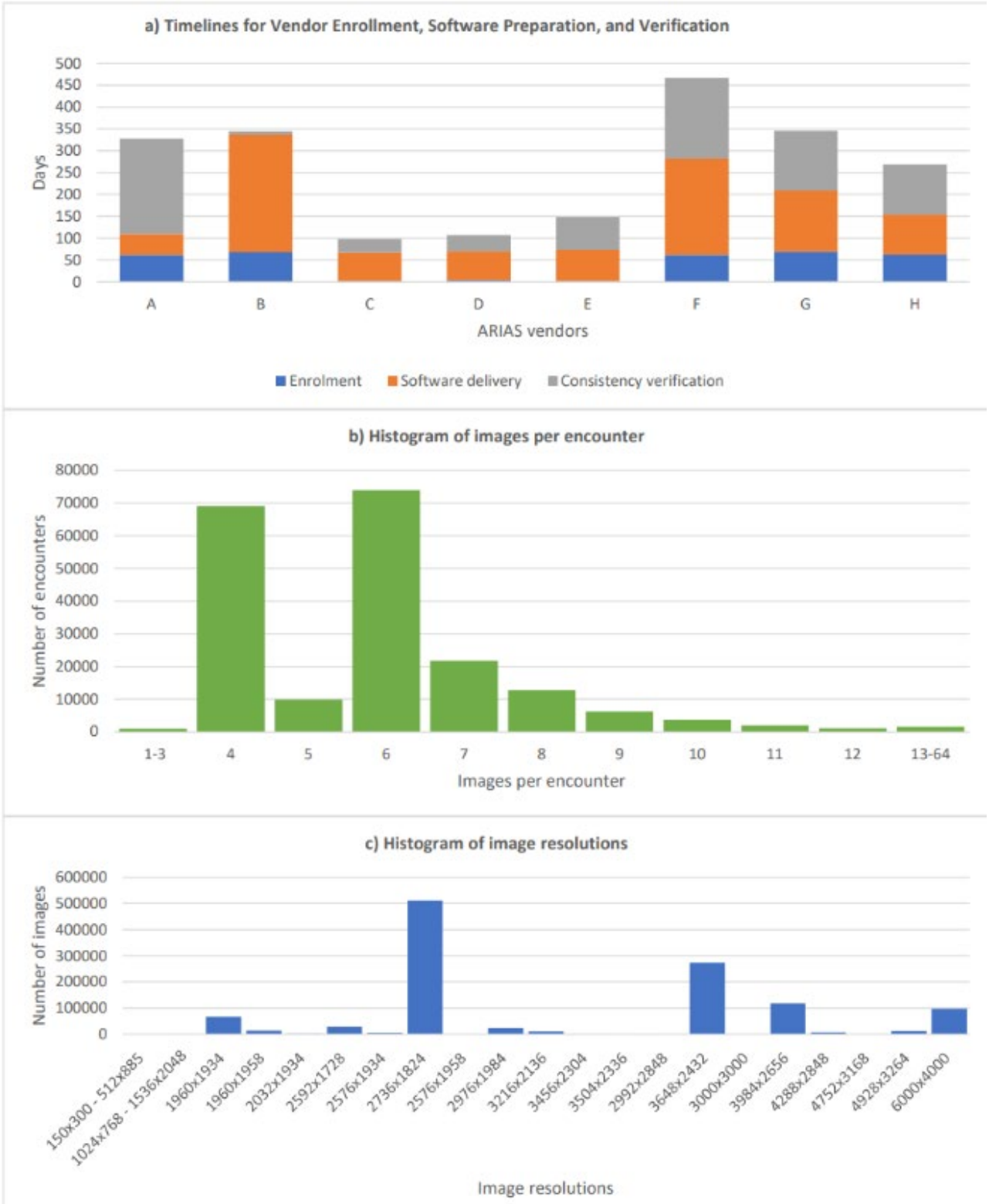

ARIAS vendor 1
ARIAS vendor 2
ARIAS vendor N
ARIAS updates and bug fixes
Prepares ARIAS with a given API in a docker container
ARIAS evaluated on SRI in vendor's lab
(7) SRI results comparison Vendor <-> Research team
(8) Vendor approves CRI evaluation on PT server
No
Yes
(1) Sample Retinal Images (SRI) data collection
Deceased patients, anonymous, 1000 encounters, 6,268 images
(2) Transfer to TRE, from NEL DESP
(3) Vendor signs SRI NDA
(4) SRI provided to vendor
(5) Vendor supplies ARIAS
ARIAS installed on PT server in TRE
Vendor's and Research team SRI results
(6) ARIAS evaluated on SRI on PT server
(7) SRI results comparison Vendor <-> Research team
Research team verifies the ARIAS inference consistency
(1) Curated Retinal Images (CRI) data collection
Date span 1/1/2021 - 31/12/2022 (1 year, 11 months, 30 days)
202886 encounters, 1,175,423 images
(2) Transfer to TRE, from NEL DESP
(9) ARIAS evaluated on CRI on PT server
Results for data analysis

Figure 2



*Reasons given for withdrawal: not confident their ARIAS would perform well on our data, not able to deliver ARIAS within 6 month time window, Algorithm does not produce diabetic retinopathy specific outcomes

Figure 3


a) Timelines for Vendor Enrollment, Software Preparation, and Verification
Days
500
450
400
350
300
250
200
150
100
50
0
A
B
C
D
E
F
G
H
ARIAS vendors
Enrolment
Software delivery
Consistency verification
b) Histogram of images per encounter
Number of encounters
80000
70000
60000
50000
40000
30000
20000
10000
0
1-3
4
5
6
7
8
9
10
11
12
13-64
Images per encounter
c) Histogram of image resolutions
Number of images
600000
500000
400000
300000
200000
100000
0
150x300 - 512x885
1024x768 - 1536x2048
1960x1934
1960x1958
2032x1934
2592x1728
2576x1934
2736x1824
2576x1958
2976x1984
3216x2136
3456x2304
3504x2336
2992x2848
3648x2432
3000x3000
3984x2656
4288x2848
4752x3168
4928x3264
6000x4000
Image resolutions