A Global Review of Publicly Available Datasets for Ophthalmic Imaging: Recognising Barriers to Access, Usability and Generalisability

Saad M Khan, MBChB[1]*, Xiaoxuan Liu, MBChB[1,2,3,4]*, Siddharth Nath, PhD[5], Edward Korot, MD[3,6], Livia Faes, MD[3,7], Siegfried K Wagner, BMBCh[8], Pearse A Keane, MD[4,8], Neil J Sebire[4], MD, Matthew J Burton, FRCOphth[9], Alastair K Denniston, PhD[1,2,4,8]

[1]Academic Unit of Ophthalmology, Institute of Inflammation & Ageing, College of Medical and Dental Sciences, University of Birmingham, United Kingdom; [2]Ophthalmology Department, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom; [3]Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom; [4]Health Data Research UK, London, United Kingdom; [5] Department of Ophthalmology and Visual Sciences, McGill University, Montreal, Quebec, Canada; [6]Stanford University Byers Eye Institute, Palo Alto, CA, USA; [7]Eye Clinic, Cantonal Hospital of Lucerne, Lucerne, Switzerland; [8]NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom; [9]International Centre for Eye Health, Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, United Kingdom

This review is now published in the Lancet Digital Health Journal and forms part of The Lancet Global Health Commission on Global Eye Health, which examines some of the central issues in global eye health

UNIVERSITY OF BIRMINGHAM

Moorfields Eye Hospital NHS Foundation Trust

HDRUK Health Data Research UK

NHS University Hospitals Birmingham NHS Foundation Trust

Background: Publicly available health data are valuable resources for digital health research (Rajpurkar et al., 2017; Kermany et al., 2018). There is a growing demand for large, readily available imaging datasets to serve as training data for machine learning models, particularly in the field of ophthalmology (De Fauw et al., 2018; Gulshan et al., 2016). However, it is currently unknown how many datasets exist. This study aims to identify all public ophthalmic imaging datasets, detailing their accessibility and describing which diseases and populations are represented.

Methods: A search was conducted on Medline, Google, and Google Datasets Search in December 2019 to identify publicly available ophthalmic imaging datasets based on pre-specified inclusion & exclusion criteria (*Figure 1*).

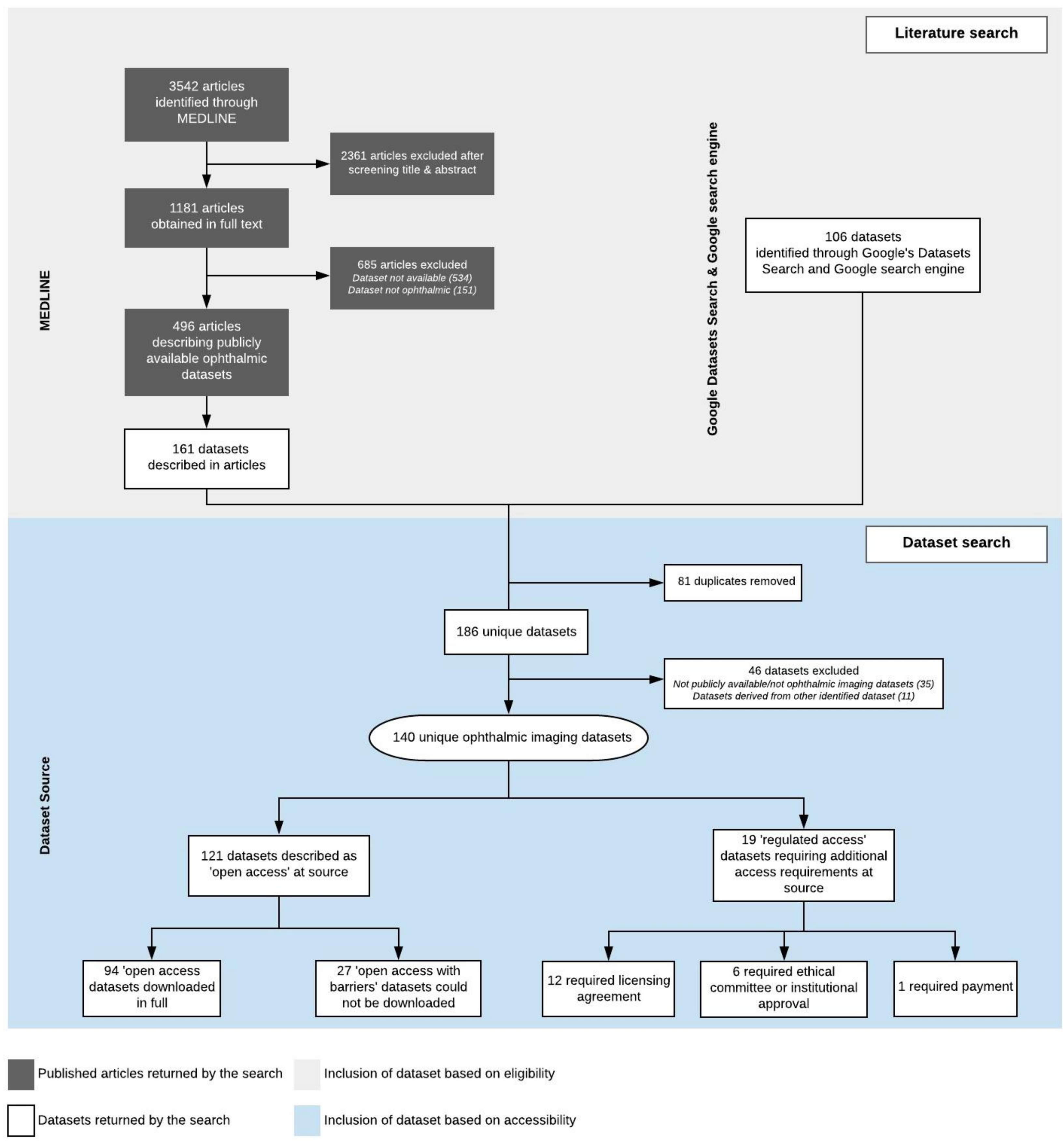
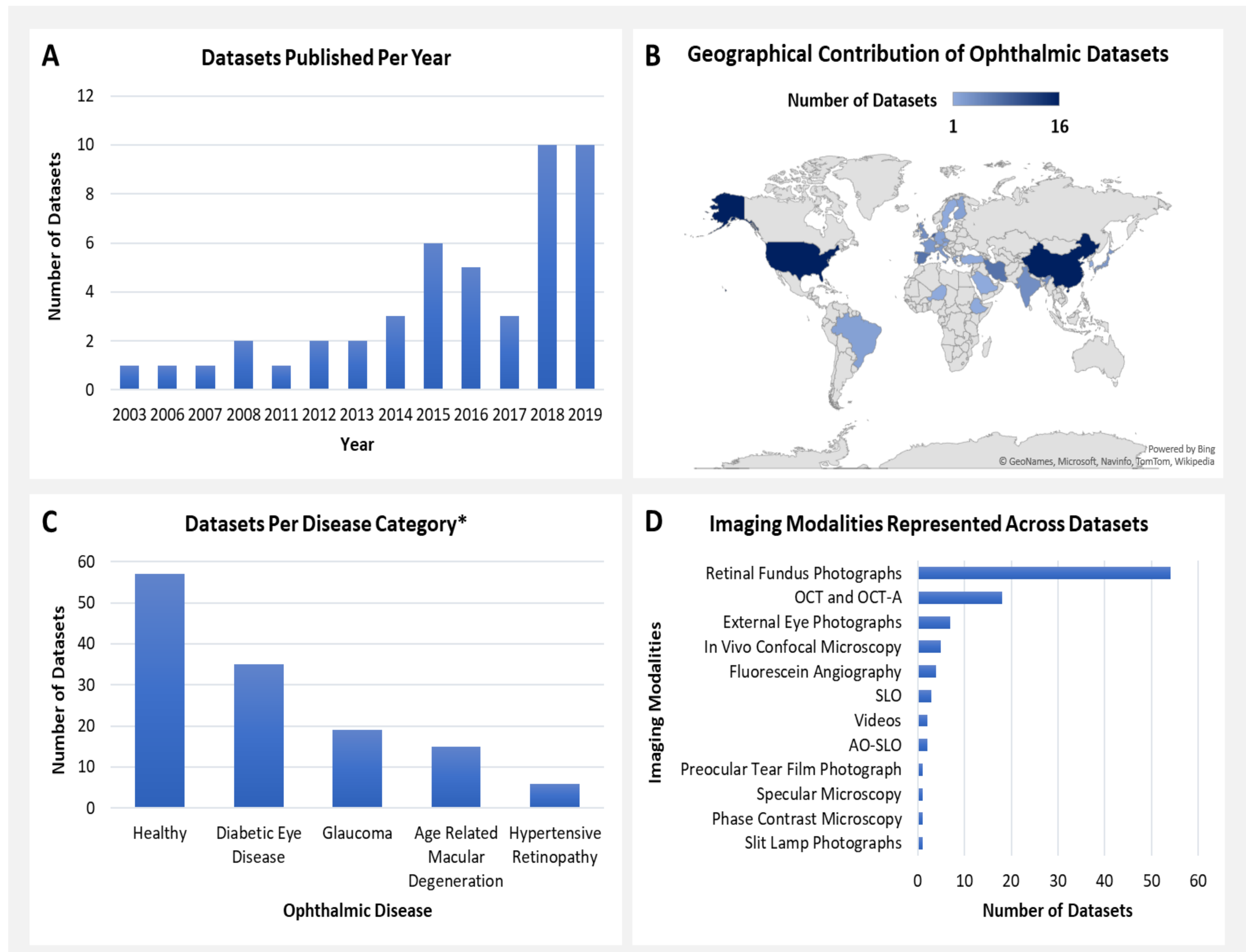


Figure 1. Dataset identification (MEDLINE articles, Google's Datasets Search and the Google Search Engine), dataset selection and details of accessibility

We defined dataset accessibility as "open access": no or minimum requirements (e.g. submission of personal information, email request) for access; "open access with barriers": fulfilling the theoretical criteria for "open access" but inaccessible due to unpredictable reasons (e.g. no response to requests or broken hyperlinks); and "regulated access": required fulfilment of formal agreements, approvals, or payment. Data from each dataset was downloaded and information relating to the data accessibility, images, ophthalmic diseases, population demographics, and ground truth were extracted.

Results: We identified 94 "open access" datasets containing 507,724 images and 125 videos from at least 122,364 subjects. Database inception was reported by 47/94 (50%) datasets and ranged from 2003 to 2019 (*Figure 2A*).



*Only diseases represented in ≥ 5 datasets have been included. Where datasets include multiple diseases, they are counted multiple times.
Abbreviations – AO-SLO: Adaptive Optics-Scanning Laser Ophthalmoscopy, **SLO:** Scanning Laser Ophthalmoscopy, **OCT:** Optical Coherence Tomography, **OCT-A:** Optical Coherence Tomography-Angiography.

Figure 2. The information related to the (A) publication date, (B) geographical distribution, (C) represented diseases, and (D) image types

Most datasets originated from Asia and Europe (34 datasets each). We found no known publicly available datasets for ophthalmic images in 172 countries equating to nearly 45% of the global population (*Figure 2B*). Besides healthy eyes, the most common diseases included were diabetic eye disease (35 datasets) and glaucoma (19 datasets) (*Figure 2C*). Images included fundus photographs (54 datasets), optical coherence tomography (18 datasets), and external eye photographs (7 datasets) among others (*Figure 2D*). Only 55/94 (59%) datasets included ground truth labels, manual segmentation, or both. We also assessed completeness of metadata reporting for each dataset (*Figure 3*). Technical information relating to image numbers, imaging devices, and image file formats were well-reported (100% of datasets), however clinical characteristics such as patient demographics and inclusion criteria were poorly reported (missing in >20% of datasets).

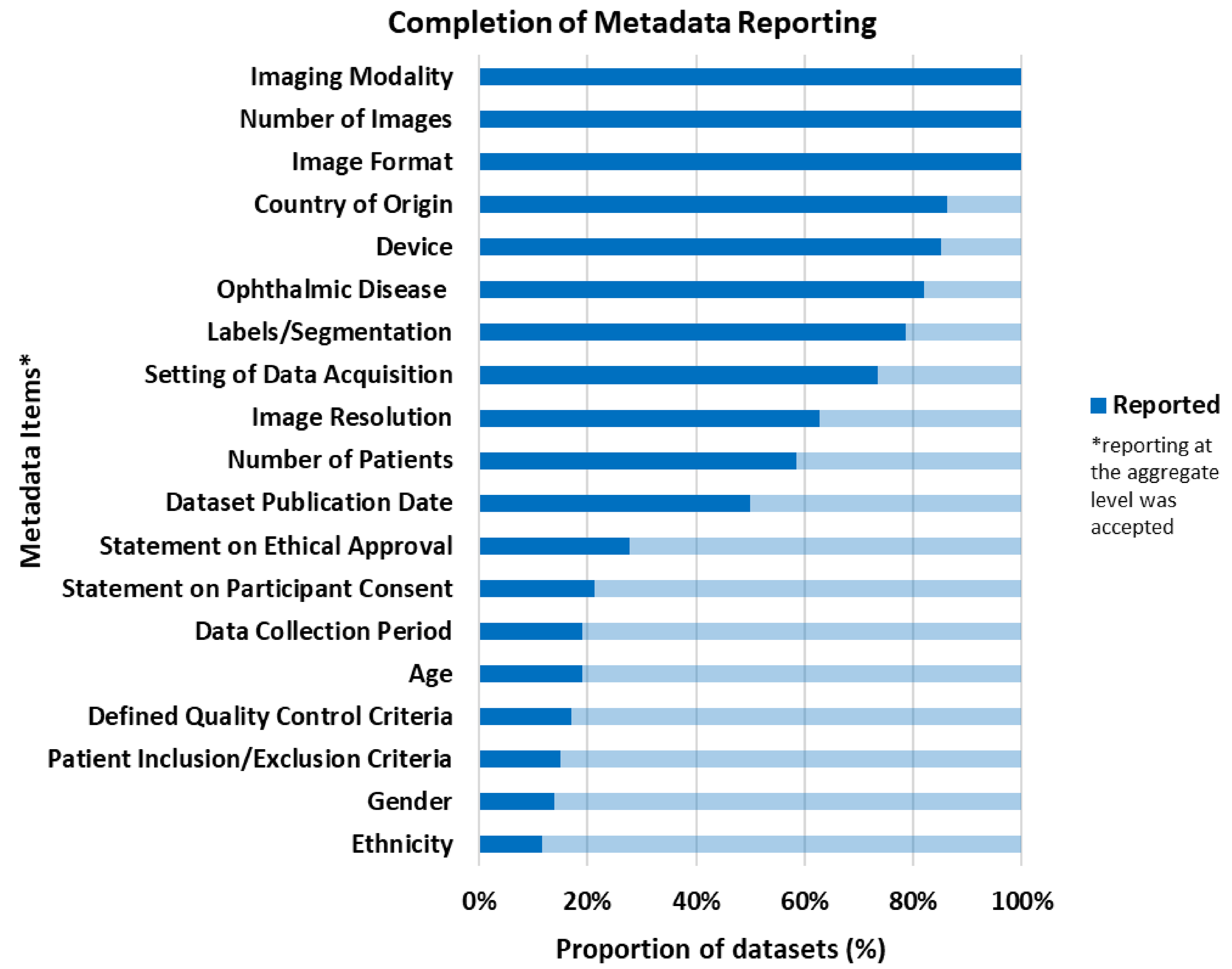


Figure 3. Completion of reporting of metadata items across all 94 datasets

Conclusions: This study provides greater visibility of a diverse collection of publicly available ophthalmic datasets. However, diseases are unequally represented and accompanying clinical metadata is often poorly reported. Future datasets covering underrepresented eye diseases and populations, with adequate reporting of clinical metadata, will help advance the application of artificial intelligence in ophthalmology.

References:

De Fauw, J. et al. (2018) 'Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease', Nature Medicine, 24(9), pp.1342-1350.

Gulshan, V. et al. (2016) 'Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs', JAMA: the journal of the American Medical Association, 316(22), pp. 2402–2410.

Kermany, D. S. et al. (2018) 'Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning', Cell, 172(5), pp. 1122–1131.e9.

Rajpurkar, P. et al. (2017) 'CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning'. arXiv.