Supplement to: "Automated detection of glaucoma with interpretable machine learning using clinical data and multi-modal retinal images"

Parmita Mehta^{1,*}, Christine A. Petersen², Joanne C. Wen², Michael R. Banitt², Philip P. Chen², Karine D. Bojikian², Catherine Egan⁴, Su-In Lee¹, Magdalena Balazinska^{1,3}, Aaron Y. Lee^{2,†}, Ariel Rokem^{3,†}, and The UK Biobank Eye and Vision Consortium⁵

¹Paul G. Allen School of Computer Science and Engineering, University of Washington

²Department of Ophthalmology, University of Washington ³University of Washington eScience Institute ⁴Moorfields Eye Hospital, NHS Trust, UK *parmita@cs.washington.edu †These authors contributed equally ⁵Full list of members provided in supplement

Supplementary Information

Members of the UK Biobank Eye and Vision Consortium include: Tariq Aslam, PhD, Manchester University, Sarah A. Barman, PhD, Kingston University, Jenny H. Barrett, PhD, University of Leeds, Paul Bishop, PhD, Manchester University, Peter Blows, BSc, NIHR Biomedical Research Centre, Catey Bunce, DSc, King's College London, Roxana O. Carare, PhD, University of Southampton, Usha Chakravarthy, FRCOphth, Queens University Belfast, Michelle Chan, FRCOphth, NIHR Biomedical Research Centre,Sharon Y.L. Chua, PhD, NIHR Biomedical Research Centre, David P. Crabb, PhD, UCL, Philippa M. Cumberland, MSc, UCL Great Ormond Street Institute of Child Health, Alexander Day, PhD, NIHR Biomedical Research Centre, Parul Desai, PhD, NIHR Biomedical Research Centre, Bal Dhillon, FRCOphth, University of Edinburgh, Andrew D. Dick, FRCOphth, University of Bristol, Cathy Egan, FRCOphth, NIHR Biomedical Research Centre, Sarah Ennis, PhD, University of Southampton, Paul Foster, PhD, NIHR Biomedical Research Centre, Marcus Fruttiger, PhD, NIHR Biomedical Research Centre, John E. J. Gallacher, PhD, University of Oxford, David F. Garway-Heath, FRCOphth, NIHR Biomedical Research Centre, Jane Gibson, PhD, University of Southampton, Dan Gore, FRCOphth, NIHR Biomedical Research Centre, Jeremy A. Guggenheim, PhD, Cardiff University, Chris J. Hammond, FRCOphth, King's College London, Alison Hardcastle, PhD, NIHR Biomedical Research Centre, Simon P. Harding, MD, University of Liverpool, Ruth E. Hogg, PhD, Queens University Belfast, Pirro Hysi, PhD, King's College London, Pearse A. Keane, MD, NIHR Biomedical Research Centre, Sir Peng T. Khaw, PhD, NIHR Biomedical Research Centre, Anthony P. Khawaja, DPhil, NIHR Biomedical Research Centre, Gerassimos Lascaratos, PhD, NIHR Biomedical Research Centre, Andrew J. Lotery, MD, University of Southampton, Tom Macgillivray, PhD, University of Edinburgh, Sarah Mackie, PhD, University of Leeds, Keith Martin, FRCOphth, University of Cambridge, Michelle McGaughey, Queen's University Belfast, Bernadette McGuinness, PhD, Queen's University Belfast, Gareth J. McKay, PhD, Queen's University Belfast, Martin McKibbin, FRCOphth, Leeds Teaching Hospitals NHS Trust, Danny Mitry, PhD, NIHR Biomedical Research Centre, Tony Moore, FRCOphth, NIHR Biomedical Research Centre, James E. Morgan, DPhil, Cardiff University, Zaynah A. Muthy, BSc, NIHR Biomedical Research Centre, Eoin O'Sullivan, MD, King's College Hospital NHS Foundation Trust, Chris G. Owen, PhD, University of London, Praveen Patel, FRCOphth, NIHR Biomedical Research Centre, Euan Paterson, BSc, Queens University Belfast, Tunde Peto, PhD, Queen's University Belfast, Axel Petzold, PhD, UCL, Jugnoo S. Rahi, PhD, UCL Great Ormond Street Institute of Child Health, Alicja R. Rudnikca, PhD, University of London, Jay Self, PhD, University of Southampton, Sobha Sivaprasad, FRCOphth, NIHR Biomedical Research Centre, David Steel, FRCOphth, Newcastle University, Irene Stratton, MSc, Gloucestershire Hospitals, NHS Foundation Trust, Nicholas Strouthidis, PhD, NIHR Biomedical Research Centre, Cathie Sudlow, DPhil, University of Edinburgh, Dhanes Thomas, FRCOphth, NIHR Biomedical Research Centre, Emanuele Trucco, PhD, University of Dundee, Adnan Tufail, FRCOphth, NIHR Biomedical Research Centre, Veronique Vitart, PhD, University of Edinburgh, Stephen A. Vernon, DM, Nottingham University Hospitals, NHS Trust, Ananth C. Viswanathan, FRCOphth, NIHR Biomedical Research Centre, Cathy Williams, PhD, University of Bristol, Katie Williams, PhD, King's College London, Jayne V. Woodside, MRCOphth, PhD, Queen's University Belfast, Max M. Yates, PhD, University of East Anglia, Jennifer Yip, PhD, University of Cambridge, and Yalin Zheng, PhD, University of Liverpool.

Supplementary Results and Methods

Reliability of self report

We ascertained the reliability of glaucoma labels. The labels that we used to determine whether a study participant did or did not have glaucoma were based on self-report. While previous studies have relied on self-report as labels[1], the reliability of these labels in the current cohort still needs to be ascertained. In particular, there is a concern of a high prevalence of false negatives as many cases of glaucoma go undetected [4, 5, 3, 2, 6]: there is a concern that there are many individuals who are labeled as healthy, while they already have glaucoma. In addition, there may be a concern of false positives: people who self-report that they have glaucoma but have other eye diseases. We relied on several different assessments: First, we looked at Test-retest reliability of self-reported labels. The total cohort had 96,020 participants. Of these, 12,397 (12.9%) had more than one visit (up to 3). Of these, only 49 (0.3%) had inconsistent self-reports of glaucoma across visits. That is, they reported glaucoma on one visit and reported that they did not have glaucoma on a subsequent visit. This means that 99.7% of individuals with repeated visits are consistent in their self-report of glaucoma, suggesting high test-retest reliability. Second, a subset of individuals in the UK Biobank dataset who were admitted as inpatients have clinical ICD-10 codes associated with them. These codes indicate a much higher level of confidence in a given label as they tell us that this individual was seen by a health-care provider and that the provider made a diagnostic decision to attach this diagnosis code to this individual. In the data we used for this study, only 38.46% of the subjects (45% of eyes) had an associated ICD-10 code. However, there are only 88 (11.61%) subjects (150 eyes) that have an ICD-10 code of glaucoma. All of these 88 subjects self-reported having glaucoma and none of the control (self-reported normal) had an ICD-10 code of glaucoma. This suggests a low prevalence of false negatives. Third, in addition to the self-report of glaucoma, participants could self-report on their use of medication. In our dataset, 296 (18.3%) subjects (515 (21%) eyes) reported that they take medication that is prescribed for glaucoma. Only one of these did not also self-report glaucoma. This is the one case we found using these three methods in which we might consider that we have in fact found a false negative (0.11%) . In addition, in our dataset 8 subjects (1.05%) / 15 eyes (1.17%) are self-reported to have glaucoma and also take other eye medication (not used for glaucoma). This is an indication of a low false positive rate.

Statistical analysis of group differences

Statistical analysis of differences between glaucoma, healthy, PTG and PTG post-diagnosis demonstrates that pulmonary capacity variables differ between the groups. This is true even when controlling for age: an ANCOVA of FVC values for the three groups, using age as a covariate found a statistically significant effect of group identity (Supplementary Table 1). To further confirm this, we followed this ANCOVA with a pairwise Dunn's test (Supplementary Table 2 and Supplementary Table 3). We found that FVC in PTG significantly differed from FVC in patients with glaucoma (Supplementary Table 2). Likewise, FVC in normal vs. patients with glaucoma differed significantly (Supplementary Table 2), but we did not find a difference between normal and PTG (Supplementary Table 2). Values for PEF were similar: the ANCOVA using age as a covariate showed a statistically significant effect of group (Supplementary Table 1); Dunn's tests showed similar differences: glaucoma differ significantly from both PTG and healthy, but healthy and PTG do not significantly differ (Supplementary Table 3). In addition, a pairwise Dunn's test confirmed that the IOP difference between glaucoma and PTG groups was statistically significant $(Z=10.74, p<0.05)$. Post-diagnosis, the difference between PTG and glaucoma in FVC and PEF was no longer statistically significant. This was also true for IOP, but this is based on only a limited sample of PTG $(n=21 \text{ retains})$, for which there is a measurement of IOP post-diagnosis (see Methods).

Shuffle Test Results

Area under ROC curve(%): 53.625

Dependent Variable	dF	F Value	p value
FVC	1.2403	64.51	$1.49e - 15$
PEF	1.2403	50.8	$1.35e - 12$

Supplementary Table 1: Analysis of covariance (ANCOVA) of pulmonary capacity variables (forced vital capacity (FVC), peak expiratory capacity(PEF)) for group identity amongst the three groups (Healthy, Glaucoma and PTG), controlling for age.

Supplementary Table 2: Dunn's test comparing FVC for Glaucoma, healthy, PTG and PTG (post-diagnosis) groups, with Bonferroni correction for p-values.

Supplementary Table 3: Dunn's test comparing PEF for Glaucoma, healthy, PTG and PTG (post-diagnosis) groups, with Bonferroni correction for p-values.

Supplementary Table 4: Dunn's test comparing Age for Glaucoma, healthy, PTG and PTG (post-diagnosis) groups, with Bonferroni correction for p-values.

Supplementary Table 5: Dunn's test comparing BMI for Glaucoma, healthy, PTG and PTG (post-diagnosis) groups groups, with Bonferroni correction for p-values.

Supplementary Table 6: Dunn's test comparing IOP for Glaucoma, healthy, PTG and PTG (post-diagnosis) groups, with Bonferroni correction for p-values.

Supplementary Table 7: Normalized confusion matrix for shuffle test.

Supplementary Table 8: All factors used for building the baseline models.

Supplementary Table 9: Other Eye medications.

Supplementary Table 10: Glaucoma medications.

Supplementary Figure 1: Selection criteria for glaucoma and progress to glaucoma subjects from UK Biobank data set.

Supplementary Figure 2: Sample of OCT volumes eliminated due to alignment errors, this shows first two OCT slices for four retinas where alignment failure occurred. All OCT slices were eliminated for such retinas.

Supplementary Figure 3: Data distribution: Age and gender distributions as well as number of subjects in the Normal (left), progress to glaucoma (middle) and glaucoma (right) subject groups.

Supplementary Figure 4: Sample of CFP eliminated from test set due to bad quality.

Supplementary Figure 5: Data processing for CFP images. Top row shows original images, bottom row shows cropped and pre-processed images.

Supplementary Figure 6: Interpreting model built on demographic, systemic and ocular data. A) The ten most important features from BM3 based on SHAP values. B-E) SHAP values vs feature values for age, IOP, BMI, and FVC respectively. Each point represents an individual subject and the color denotes whether or not they have been diagnosed with glaucoma. F-H) SHAP values vs feature values for features IOP, BMI, and FVC respectively, with each point colored based on age of subject.

Supplementary Figure 7: SHAP interaction values in BM3. Interactions between age and each of the other top variables in model BM3: A SHAP interaction values for age and IOP. B SHAP interaction values for age and BMI. C SHAP interaction values for.