Published in final edited form as:
Lancet Digit Health. 2023 May ; 5(5): e257–e264. doi:10.1016/S2589-7500(23)00022-5.

A deep learning model for novel systemic biomarkers in photographs of the external eye: a retrospective study

Boris Babenko[*],
Ilana Traynis,
Christina Chen,
Preeti Singh,
Akib Uddin,
Jorge Cuadros,
Lauren P Daskivich,
April Y Maa,
Ramasamy Kim,
Eugene Yu-Chuan Kang,
Yossi Matias,
Greg S Corrado,
Lily Peng,
Dale R Webster,
Christopher Semturs,
Jonathan Krause,
Avinash V Varadarajan,
Naama Hammel[*],
Yun Liu[*]

(B Babenko PhD, C Chen MD, P Singh MS, A Uddin MHSc, Y Matias PhD, G S Corrado PhD, L Peng MD, D R Webster PhD, C Semturs MS, J Krause PhD, A V Varadarajan MS, N Hammel MD, Y Liu PhD); **Advanced Clinical, Deerfield, IL, USA** (I Traynis MD); **EyePACS, Santa Cruz, CA, USA** (J Cuadros OD); **Ophthalmic Services and Eye Health Programs, Los Angeles County Department of Health Services, Los Angeles, CA, USA** (L P Daskivich MD); **Department of Ophthalmology, University of Southern California Keck School of Medicine/ Roski Eye Institute, Los Angeles, CA USA** (L P Daskivich); **Department of Ophthalmology,**

Correspondence to: Dr Naama Hammel, Google Health, Palo Alto, CA 94303, USA nhammel@google.com, or, Dr Yun Liu, Google Health, Palo Alto, CA 94304, USA liuyun@google.com.

[*] Contributed equally

Contributors
BB, CC, NH, and YL conceived and designed the study. BB and YL analysed the data. IT conducted the literature review. JC, LPD, AYM, AVV, PS, and NH were involved in data collection and curation. AU, RK, EY-CK, YM, GSC, LP, DRW, CS, JK, and AVV provided strategic guidance and oversight. GSC, LP, DRW, CS, JK, and AVV obtained funding. BB, IT, NH, and YL drafted the manuscript with feedback from all authors. JC, LPD, and AYM verified the raw data; BB and YL verified the results of this study. All Google authors had full access to all the data in the study. All authors contributed to data interpretation, read and approved the final manuscript, and had final responsibility for the decision to submit for publication.

See Online for appendix

Emory University School of Medicine, Atlanta, GA, USA (A Y Maa MD); **Regional Telehealth Services, Technology-based Eye Care Services (TECS) division, Veterans Integrated Service Network (VISN) 7, Decatur, GA, USA** (A Y Maa); **Aravind Eye Hospital, Madurai, Tamil Nadu, India** (R Kim MD); **Department of Ophthalmology, Linkou Medical Center, Chang Gung Memorial Hospital, Taoyuan, Taiwan** (E Y-C Kang MD)

Summary

Background—Photographs of the external eye were recently shown to reveal signs of diabetic retinal disease and elevated glycated haemoglobin. This study aimed to test the hypothesis that external eye photographs contain information about additional systemic medical conditions.

Methods—We developed a deep learning system (DLS) that takes external eye photographs as input and predicts systemic parameters, such as those related to the liver (albumin, aspartate aminotransferase [AST]); kidney (estimated glomerular filtration rate [eGFR], urine albumin-to-creatinine ratio [ACR]); bone or mineral (calcium); thyroid (thyroid stimulating hormone); and blood (haemoglobin, white blood cells [WBC], platelets). This DLS was trained using 123 130 images from 38 398 patients with diabetes undergoing diabetic eye screening in 11 sites across Los Angeles county, CA, USA. Evaluation focused on nine prespecified systemic parameters and leveraged three validation sets (A, B, C) spanning 25 510 patients with and without diabetes undergoing eye screening in three independent sites in Los Angeles county, CA, and the greater Atlanta area, GA, USA. We compared performance against baseline models incorporating available clinicodemographic variables (eg, age, sex, race and ethnicity, years with diabetes).

Findings—Relative to the baseline, the DLS achieved statistically significant superior performance at detecting AST >36·0 U/L, calcium <8·6 mg/dL, eGFR <60·0 mL/min/1·73 m^2, haemoglobin <11·0 g/dL, platelets <150·0 × 10^3/μL, ACR ≥300 mg/g, and WBC <4·0 × 10^3/μL on validation set A (a population resembling the development datasets), with the area under the receiver operating characteristic curve (AUC) of the DLS exceeding that of the baseline by 5·3–19·9% (absolute differences in AUC). On validation sets B and C, with substantial patient population differences compared with the development datasets, the DLS outperformed the baseline for ACR ≥300·0 mg/g and haemoglobin <11·0 g/dL by 7·3–13·2%.

Interpretation—We found further evidence that external eye photographs contain biomarkers spanning multiple organ systems. Such biomarkers could enable accessible and non-invasive screening of disease. Further work is needed to understand the translational implications.

Funding—Google.

Introduction

Ocular sequelae resulting from systemic disease have been well documented[1] and are the basis for globally established screening programmes that identify diabetic retinal disease from fundus photography.[2] Work from the past 5 years has shown that a number of systemic biomarkers, such as blood pressure, glycated haemoglobin (HbA_{1c}),[3] and estimated glomerular filtration rate (eGFR),[4] can also be detected from fundus photographs.[5] Although this suggests an opportunity for non-invasive detection of systemic disease, the need for a trained photographer and a specialised fundus camera to capture retinal photographs presents a barrier to clinical implementation.

By contrast, our previous work[6] examined whether a few markers of systemic and ocular disease related to diabetes could be identified from photographs of the external eye, which in principle do not require specialised cameras to obtain. We found that our deep learning system (DLS) could identify poor blood glucose control, diabetic retinopathy, and diabetic macular oedema from external eye photographs. Since clinical literature has indicated that external eye structures manifest many signs of systemic disease beyond diabetes, we considered whether machine learning could glean known or novel signals of other systemic disease from photographs of the external eye.

In this work we tested the hypothesis that machine learning could predict several additional biomarkers of systemic diseases, such as ones related to kidney, liver, thyroid, bone or mineral, and blood count, from external eye photographs. To test this hypothesis, we used data where patients had both external eye photographs taken and clinical parameters or laboratory measurements (eg, serum chemistry panels or complete blood counts) to develop and validate a DLS. We investigated measurements spanning a variety of systems: albumin, albumin-to-creatinine ratio (ACR), calcium, eGFR, aspartate aminotransferase (AST), thyroid stimulating hormone (TSH), haemoglobin, platelets, white blood cells (WBC), blood pressure, BMI, and more.

Methods

Datasets

To develop and evaluate the DLS, we used data from three sources: (1) clinics in the Los Angeles County Department of Health Services (LACDHS) using the EyePACS teleretinal diabetic screening programme (56 579 patients seen in 2013–21); (2) Veterans Affairs (VA) primary care clinics in the greater Atlanta area participating in the Technology-based Eye Care Services (TECS, an initiative developed at the Atlanta VA Healthcare System) programme (10 030 patients seen in 2017–20); and (3) community-based outpatient clinics in the Atlanta VA Healthcare System’s diabetic retinopathy screening programme called TeleRetinal Imaging (TRI; 5552 patients seen in 2009–17; see appendix p 16).

For EyePACS/LACDHS, 11 sites were used for training, four for tuning, and three for external validation (validation dataset A). This yielded an approximate 65:15:20 patient ratio, which was empirically expected to be a reasonable tradeoff in sample size for training versus evaluation. Patients who had visits at sites across these splits were excluded. The validation sets from the Atlanta area (B and C) served as additional external validation. Patients across all datasets had their pupils dilated in preparation for fundus photography (ie, not dilated specifically for external eye photography) as part of standard of care. Ethics review and Institutional Review Board exemption for this retrospective study on deidentified data were obtained via Advarra, County of Los Angeles Public Health, and Emory University Institutional Review Boards. TRIPOD reporting guidelines were followed (appendix pp 13–15).

At all sites, external eye photographs were taken as part of the standard imaging protocol for teleretinal eye screening to assess the anterior segment. The imaging protocols were largely similar at the LACDHS and VA sites, with the eyes imaged using the Topcon (NW8 and NW400) and Zeiss Cirrus Photo 600 fundus cameras, and the patient positioned slightly further from the camera relative to the position used for photographing the fundus. More details can be found in the appendix (p 12). External eye images were not filtered based on image quality.

Baseline characteristics and parameters predicted

Parameters predicted by the DLS included baseline characteristics and clinical or laboratory measurements. Baseline clinicodemographic characteristics (ie, age, sex, race and ethnicity, and years with diabetes) were self-reported by patients and recorded at each site. For systolic and diastolic blood pressure and weight, we averaged measurements within 90 days of photograph acquisition, whereas height was averaged over 365 days (for sensitivity analyses, see appendix p 7).

Serum and urine laboratory measurements, as well as clinical measurements, were extracted from the medical records. For serum and urine tests, we took the measurement closest in time to photograph acquisition, and excluded laboratory results more than 30 days away for international normalised ratio (INR), exceeding 90 days for HbA_{1c}, and exceeding 180 days for all other measurements.

In all datasets, we computed the eGFR using the race free 2021 CKD-EPI Creatinine equation,[7] as recommended by the National Kidney Foundation and the American Society of Nephrology’s Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease, in the absence of cystatin C in retrospective data.[8] This also ensured consistency of eGFR estimation across datasets and over time.

Additional details about labels are discussed in the appendix (p 2).

DLS development

We trained a convolutional neural network to take an external eye photograph as input, and predict all clinical and laboratory measurements in a multitask fashion (ie, one prediction “head” per task). Specifically, each clinical and laboratory measurement was thresholded (cutoffs are listed in appendix pp 36–40) to facilitate uniform treatment as classification tasks and use of the standard cross entropy loss during training. The cutoffs were selected during model development by consulting the American Board of Internal Medicine laboratory reference ranges[9] and taking into account dataset statistics (ie, we eschewed cutoffs yielding too few positive cases to train and evaluate on reliably). Clinical and laboratory measurements with multiple cutoffs were configured as a single multiclass head (eg, a laboratory measurement with two cutoffs [X and Y] was configured to have three classes, corresponding to: ≤X, >X and ≤Y, >Y).

From the list of prediction targets, we selected the nine most promising ones based on DLS performance versus the baseline model on the tuning set, potential clinical utility, and representation of multiple physiological systems: albumin <3·5 g/dL (liver and kidney), AST

>36·0 U/L (liver), calcium <8·6 mg/dL (kidney and endocrine), eGFR <60·0 mL/min/1·73 m^2 (kidney), haemoglobin <11·0 g/dL (blood count), platelets <150·0 × 10^3/μL (blood count), TSH >4·0 mU/L (thyroid), urine ACR ≥300·0 mg/g (kidney), and WBC <4·0 × 10^3/μL (blood count). We prespecified these primary analyses and associated multiple testing corrections (see Statistical analysis) before running the model on the validation sets. Ensembling and additional modelling details are discussed in the appendix (p 1).

Additional supplementary analysis included training a DLS that takes both an external eye image and baseline characteristics as input (appendix p 4). For this model, we one-hot-encoded categorical metadata variables and normalised scalar values to unit variance and zero mean using the training set statistics, and then concatenated these to the prelogits (ie, penultimate layer of the convolutional neural network).

DLS evaluation

To evaluate our approach, we computed the difference between the area under receiver operator characteristic curve (AUC) of the DLS and that of a baseline model that only takes baseline characteristics as input. To avoid bias towards patients with many visits, for each prediction target, we filtered for visits where the target measurement was available, and then randomly selected one visit per patient (appendix p 16).

Baseline models for comparison

Baseline models to contextualise DLS performance were logistic regression models trained with the scikit-learn Python library (version 1.0.2) in Python version 3.9.15, using class-balanced weighting and the default L2 regularisation setting (C=1·0). Each validation set had different variables available, with EyePACS/LACDHS being most comprehensive. Baseline models were trained only on datapoints for which all input variables and the target variable were available. To avoid having to drop a large fraction of data in our primary analysis, we used different baselines for each dataset, such that each baseline model used only input variables that were available for at least 85% of the datapoints in the respective dataset: age, sex, race and ethnicity, and years with diabetes for the EyePACS/LACDHS testing dataset; age, sex, and race and ethnicity for VA TECS; age and sex for VA TRI. All baseline models were trained on the EyePACS/LACDHS training dataset (same as the DLS). Secondary analysis includes comparison to baseline models with additional variables (eg, BMI and blood pressure; appendix p 10).

Statistical analysis

We measured the AUC and corresponding 95% CIs for both the DLS and the baseline models using the DeLong method.[10] To evaluate the statistical significance of superiority of the DLS over the baseline models we used the DeLong paired AUC comparison test.[10] The superiority analyses on the nine prediction tasks listed in the DLS development section were prespecified and documented as primary analyses before analysing performance on the validation sets. Alpha was adjusted using Bonferroni correction for multiple hypotheses testing (α=0·05 for one-sided superiority test, divided by nine tasks=0·0056).

Prespecified secondary analyses included AUC comparison for the complete set of clinical and laboratory measurements and corresponding cutoffs, subgroup analysis, additional metrics (eg, positive predictive value), explainability analysis, and sensitivity analysis with respect to the time gap between the clinical or laboratory measurement and the photograph.

Role of the funding source

Google was involved in the design and conduct of the study; management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; and decision to submit the manuscript for publication.

Results

The three data sources spanned a variety of patient populations and imaging protocols (table). The EyePACS/LACDHS dataset (including validation set A) and VA teleretinal screening dataset (validation set C) were exclusively patients with diabetes presenting for diabetic retinopathy screening, while the TECS dataset (validation set B) included patients both with and without diabetes. With respect to demographic variables, patients in validation set A had a mean age of about 57 years, slightly skewed towards female (54·6%), and the majority were Hispanic (80·4%). Validation sets B and C represent a veteran population, both with a mean age above 60 years, predominantly male (84·8–95·4%), and (where this information was available) majority Black (47–56%) or White (43–52%).

For the set of primary analyses, we first compared the AUC of the DLS against the AUC of a baseline model that takes clinicodemographic variables (without the external eye image) as input. Figure 1 summarises these results for all three validation sets (see appendix p 26 for more complete data including p values and confidence intervals, and appendix p 19 for full receiver operating characteristic curves for validation set A). We found that for ACR ≥300·0 mg/g (severely increased albuminuria) and haemoglobin <11·0 g/dL (moderate anaemia), the external eye model AUCs were consistently statistically significantly higher than the baseline models across all validation sets: absolute AUC improvements of 8·7%, 13·2%, and 11·7% for ACR ≥300·0 mg/g in the three validation sets, respectively (p<0·0001 for all), and absolute AUC improvements of 19·9% and 7·3% for haemoglobin <11·0 g/dL in validation sets A and B (p<0·005 for both; this label was not available in validation set C). Performance of other prediction tasks is discussed in the appendix (p 3).

We also report positive predictive values with thresholds based on the 5% of patients with the highest predicted likelihood for each prediction target, and in each respective validation set (appendix p 27). Overall, we observed similar trends here, although p values were generally higher due to the smaller denominator (only 5% are predicted as positive).

Results for all variables and cutoffs that we included in the model development are summarised in the appendix (pp 36–40). A few additional tasks performed well and merit future validation. For example, for the alanine aminotransferase >29·0 U/L, blood urea nitrogen >20·0 mg/dL, potassium >5·0 mEq/L, and sodium <136·0 mEq/L tasks, the external eye model outperformed the baseline by 4·5%, 6·7%, 7·1%, and 5·0%, respectively, in validation set A. Unfortunately, these variables were not available in the other validation sets, either because they were not collected or the partner institution was not able to export them due to technical reasons. Subgroup analysis for the primary analyses as well as analysis adjusting for multiple baseline variables, to account for possible correlation between variables, are discussed in the appendix (p 8). The image resolution requirements for the primary analyses are further examined in the appendix (p 9). Finally, sensitivity of results to the time gap between photograph and clinical or laboratory measurement are discussed in appendix (p 7).

To get insight into what parts of the image are most important to the DLS performance, we conducted the following explainability experiments: we masked different regions of the image during both training and evaluation, resulting in the DLS only seeing certain parts of the external eye. Pupil or iris masking used location information from an iris and pupil detector (see appendix p 5). In addition, we included a DLS for greyscale images to assess the importance of colour. Cataracts served as a positive control in these experiments, as most of the signal should be in the pupil (ie, performance should deteriorate if the pupil is masked). Results suggested that the information is generally not isolated to only the pupil or iris, and that colour information is at least somewhat important for most prediction targets (figure 2 and appendix p 6).

Discussion

Our study demonstrates that a DLS can detect biomarkers of systemic disease from external eye photographs. We developed and evaluated this DLS using datasets from diverse populations across the USA. The results showed that our DLS performed significantly better than baseline clinicodemographic models at predicting kidney function and blood count abnormalities across all three validation sets and at predicting abnormalities in liver and multiorgan parameters in validation set A. Results for several other systemic parameters appeared promising, and are discussed in the appendix (p 11). We next discuss our findings and their potential implications on an organ system by organ system basis.

For the detection of abnormalities in kidney function, a detailed review of the performance outcomes showed that the DLS performed best at more severe disease thresholds. For ACR, the model improvement compared with the baseline was more pronounced in increased severity of albuminuria (ie, DLS performance improvement above the baseline was greatest for ACR >1500 mg/g and least for ACR >30 mg/g). Similarly, eGFR performance improvement was most substantial in more severe kidney disease. In datasets with a predominantly older population, more severe declines in kidney function were needed for the model improvement to become statistically significant as compared with the baseline. We theorise that since kidney function gradually declines with age,[11] a markedly steeper drop in glomerular filtration rate might be required to manifest as an abnormality in an ageing population.[12] Consistent with this hypothesis, we found that in younger subgroups, in which eGFR is usually highest, the model could detect more subtle deterioration of kidney function. These observations suggest that tools based on this DLS would thus be best used in screening settings, such as for detection of moderate kidney disease in a young, healthy population or for severe kidney dysfunction in an older population (both of which can be asymptomatic).[13]

With respect to low haemoglobin levels, non-invasive screening for anaemia has long been considered with physical examination, such as findings of pale oral mucosa, conjunctiva, or nailbed.[14–16] Numerous studies have shown that conjunctiva examination can help detect severe anaemia;[14,17,18] thus, it is unsurprising that our DLS had a statistically significant superior performance as compared with the baseline for detection of low haemoglobin. Our study is unique, however, in that external eye photographs were captured without any physical manipulation, such as pulling the eyelid down to expose the lower palpebral conjunctiva, which has been the more common site for evaluating anaemia.[16,19] In fact, our ablation experiments (figure 2) suggest that DLS is detecting other novel signals for anaemia, not only from conjunctiva colour, since our model still outperforms the baseline even when the image is greyscale and when all eye structures are masked except for the iris.

An unexpected study outcome was that despite exceeding baseline performance, the absolute performance for liver (AST) and thyroid (TSH) abnormalities was lacklustre, with AUCs in the low 60s (61·7–62·5%). In a recent study, Xiao and colleagues[20] developed a DLS for detecting hepatobiliary disease from slit lamp and fundus photographs. They focused on chronic liver conditions and found that relative to milder disease, their DLS had improved performance in cases of advanced liver disease (cirrhosis, liver cancer). Additional studies on chronic liver disease and more severe disease will be needed to assess our DLS's performance. The performance of Xiao and colleagues' model was also better for slit lamp photographs, which capture images of the external eye and anterior segment, as compared with fundus photographs, which capture images of the retina. This is consistent with previous literature that, in addition to icterus, other external findings are known to be associated with liver disease. Conjunctival and corneal xerosis, Bitot's spot, keratomalacia, dry eyes, and corneal ulcers are examples of such findings,[21] and are due in part to vitamin A and D deficiencies.[22] Similarly, thyroid disease is known to be associated with numerous external eye findings, such as conjunctival hyperaemia and chemosis, lid retraction, and proptosis.[23] We suspect that the performance for thyroid disease was largely limited by the low number of cases for the selected thresholds in our datasets, and larger studies using more severe thresholds might yield better results.

On the topic of the DLS's potential uses, our image resolution sensitivity analysis showed that even with low resolution images (75 pixels across, corresponding to less than 1% of the pixel count of modern smartphone cameras), the DLS still outperformed the baseline for several systemic parameters. This promising observation is in line with previous diabetes-related external eye research.[6] Taking into consideration both the low image resolution requirement and lack of need for extensive clinical training, external eye photography (such as via smartphones) might be easier to use by the general user population compared with clinician-focused tools such as a slit lamp. Across the different predictions where the DLS outperformed the baseline, the AUCs ranged from 61·5% to 87·7%, which might not be accurate enough for a diagnostic device, but is in line with other scenarios like cardiovascular risk assessment,[24] mammography,[25] and prescreening for diabetes.[26]

Several limitations apply to our study. First, regarding population characteristics, our datasets were primarily from diabetic retinopathy screening populations. The exception was validation set B, which was from a general eye screening programme, albeit the subset of this dataset with systemic laboratory measurements available was again (and perhaps unsurprisingly) predominantly diabetic. Second, regarding cameras, all images were collected on fundus cameras and it is yet unknown if images collected via alternative camera types, such as smartphone cameras, would result in comparable DLS performance. Third, patients in all datasets had dilated pupils; requiring pupil dilation would limit possible use cases so further work is needed to determine to what extent dilation affects performance. Fourth, several subgroup analyses showed a trend towards poorer performance that was difficult to interpret in light of limited sizes of subgroups (eg, age, sex, BMI, race; see appendix p 8). More focused data collection for DLS refinement and evaluation across subgroups will be needed before considering clinical use. In addition, further development and evaluation on future datasets where cystatin C is available consistently will be important to ensure more accurate eGFR measurements, particularly for Black patients.[8] Thus, further studies are needed to determine generalisability to broader patient populations and other devices. Similarly, more work is needed to understand the downstream impact of using this type of tool in a broad screening capacity, such as availability and accessibility of follow-up care and costs. Fifth, years with diabetes appeared to be predictive of some targets but was unavailable in validation sets B and C, potentially underestimating baseline performance. Furthermore, variables such as medication, comorbidities, and presence of eyelid abnormalities were not consistently available and therefore could not be incorporated into the baseline models. Lastly, insights from our explainability experiments were limited to the location of the signal (eg, cornea or iris, pupil). Additional explainability research could improve our understanding of the specific features learned by the DLS and whether these features are visible to the human eye.

Leveraging deep learning tools to detect systemic disease might be useful in several ways: individuals identified as having early or mild disease could receive early intervention to prevent progression, whereas individuals detected to have more severe disease could be prioritised for immediate care. Previous work has explored other modalities for detection of systemic disease by developing DLS that use retinal fundus images,[27] echocardiograms,[28] and electrocardiograms[29,30] to identify systemic disease. These studies have required specialised equipment or medication, such as dilating drops, which are not readily available outside of medical facilities. Our study suggests that non-invasive imaging of the external eye can provide information about systemic disease without the use of specialised equipment or medication. Removing these barriers to clinical assessment could have substantial health impact. Further studies are needed to determine if this DLS could identify systemic disease markers from external eye images captured by other camera types and how external eye screening could be effectively adapted and used in both clinical and non-clinical settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was funded by Google. We would like to acknowledge Dave Steiner, Yuan Liu, and Michael Howell for manuscript feedback; Elvia Figueroa and the LACDHS teleretinal diabetic retinopathy screening programme staff for data collection and programme support; Andrea Limon and Nikhil Kookkiri for EyePACS data collection and support; Charles Demosthenes for extracting the VA data; and Peter Kuzmak for getting images for the VA data. Part of the data processing team at LACDHS was supported by grants UL1TR001855 and UL1TR000130 from the National Center for Advancing Translational Science (NCATS) of the US National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the US Department of Veteran Affairs, or the US Government.

Declaration of interests

BB, CC, PS, AU, YM, GSC, LP, DRW, CS, JK, AVV, NH, and YL are Google employees and own Alphabet stock, and have patents related to this work planned and/or pending. IT is a paid consultant to Google. JC is the Chief Executive Officer of EyePACS, which has several contracts with Google. LPD consults for and has studies funded by EyePACS. All other authors declare no competing interests.

Data sharing

This study used deidentified data from EyePACS and the TECS and TRI programmes at the Atlanta VA Healthcare System. Interested researchers should contact JC (jcuadros@eyepacs.com) to inquire about access to EyePACS data and approach the Office of Research and Development at https://www.research.va.gov/resources/ORD_Admin/ord_contacts.cfm to inquire about access to VA data. Those interested in retraining a model can find the pretrained (BiT-M) architecture at https://github.com/google-research/big_transfer/blob/master/README.md.

References

1. Pavan-Langston D Manual of ocular diagnosis and therapy, 5th edn. Philadelphia, PA: Lippincott Williams and Wilkins, 2002.
2. Solomon SD, Chew E, Duh EJ, et al. Diabetic retinopathy: a position statement by the American Diabetes Association. Diabetes Care 2017; 40: 412–18. [PubMed: 28223445]
3. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018; 2: 158–64. [PubMed: 31015713]
4. Sabanayagam C, Xu D, Ting DSW, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. Lancet Digit Health 2020; 2: e295–302. [PubMed: 33328123]
5. Wagner SK, Fu DJ, Faes L, et al. Insights into systemic disease through retinal imaging-based oculomics. Transl Vis Sci Technol 2020; 9: 6.
6. Babenko B, Mitani A, Traynis I, et al. Detection of signs of disease in external photographs of the eyes via deep learning. Nat Biomed Eng 2022; 6: 1370–83. [PubMed: 35352000]
7. Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med 2021; 385: 1737–49. [PubMed: 34554658]
8. Delgado C, Baweja M, Crews DC, et al. A unifying approach for GFR estimation: recommendations of the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. Am J Kidney Dis 2022; 79: 268–88.e1. [PubMed: 34563581]
9. American Board of Internal Medicine. ABIM laboratory test reference ranges—January 2022. https://www.abim.org/Media/bfijryql/laboratory-reference-ranges.pdf (accessed Feb 2, 2022).
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837–45. [PubMed: 3203132]
11. Weinstein JR, Anderson S. The aging kidney: physiological changes. Adv Chronic Kidney Dis 2010; 17: 302–07. [PubMed: 20610357]
12. Waas T, Schulz A, Lotz J, et al. Distribution of estimated glomerular filtration rate and determinants of its age dependent loss in a German population-based study. Sci Rep 2021; 11: 10165. [PubMed: 33986324]

13. Shlipak MG, Tummalapalli SL, Boulware LE, et al. The case for early identification and intervention of chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int 2021; 99: 34–47. [PubMed: 33127436]
14. Sheth TN, Choudhry NK, Bowes M, Detsky AS. The relation of conjunctival pallor to the presence of anemia. J Gen Intern Med 1997; 12: 102–06. [PubMed: 9051559]
15. Mannino RG, Myers DR, Tyburski EA, et al. Smartphone app for non-invasive detection of anemia using only patient-sourced photos. Nat Commun 2018; 9: 4924. [PubMed: 30514831]
16. Collings S, Thompson O, Hirst E, Goossens L, George A, Weinkove R. Non-invasive detection of anaemia using digital photographs of the conjunctiva. PLoS One 2016; 11: e0153286. [PubMed: 27070544]
17. Strobach RS, Anderson SK, Doll DC, Ringenberg QS. The value of the physical examination in the diagnosis of anemia. Correlation of the physical findings and the hemoglobin concentration. Arch Intern Med 1988; 148: 831–32. [PubMed: 3355303]
18. McMurdy JW, Jay GD, Suner S, Trespalacios FM, Crawford GP. Diffuse reflectance spectra of the palpebral conjunctiva and its utility as a noninvasive indicator of total hemoglobin. J Biomed Opt 2006; 11: 014019. [PubMed: 16526896]
19. Park SM, Visbal-Onufrak MA, Haque MM, et al. mHealth spectroscopy of blood hemoglobin with spectral super-resolution. Optica 2020; 7: 563–73. [PubMed: 33365364]
20. Xiao W, Huang X, Wang JH, et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. Lancet Digit Health 2021; 3: e88–97. [PubMed: 33509389]
21. Prasad D, Bhriguvanshi A. Ocular manifestations of liver disease in children: clinical aspects and implications. Ann Hepatol 2020;19: 608–13. [PubMed: 31901314]
22. Venu M, Martin E, Saeian K, Gawrieh S. High prevalence of vitamin A deficiency and vitamin D deficiency in patients evaluated for liver transplantation. Liver Transpl 2013; 19: 627–33. [PubMed: 23495130]
23. Scott IU, Siatkowski MR. Thyroid eye disease. Semin Ophthalmol 1999; 14: 52–61. [PubMed: 10758212]
24. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014;129 (suppl 2): S49–73. [PubMed: 24222018]
25. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005; 353: 1773–83. [PubMed: 16169887]
26. Bang H, Edwards AM, Bomback AS, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med 2009; 151: 775–83. [PubMed: 19949143]
27. Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deeplearning algorithms. Lancet Digit Health 2020; 2: e526–36. [PubMed: 33328047]
28. Hughes JW, Yuan N, He B, et al. Deep learning evaluation of biomarkers from echocardiogram videos. EBioMedicine 2021; 73: 103613. [PubMed: 34656880]
29. Kwon JM, Cho Y, Jeon KH, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digit Health 2020; 2: e358–67. [PubMed: 33328095]
30. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. Circ Arrhythm Electrophysiol 2019; 12: e007284. [PubMed: 31450977]

Research in context

Evidence before this study

We searched Google Scholar and PubMed for published articles available in English pertaining to external eye manifestations of systemic disease up to July 1, 2022. Our specific search terms were “external eye”, “ocular manifestations”, “ophthalmic”, ”conjunctiva”, “pupil”, “systemic disease”, “diabetes”, “cardiovascular”, “hypertension”, “kidney”, “renal”, “liver”, “thyroid”, and “hemoglobin”. Previous studies identified external eye findings, via clinical examination or photography of the eyelids, conjunctiva, and cornea, which were associated with or predictive of various systemic diseases. Our previous work showed that a deep learning system could identify poorly controlled blood glucose levels, diabetic retinal disease, and lipid abnormalities from photographs of the external eye. However, there are no studies to date that have used machine learning to predict markers of other systemic disease from external eye photographs.

Added value of this study

In this study we developed a deep learning system that predicted nine biomarkers of systemic disease from external eye photographs that were obtained from multiple screening sites in the USA. The deep learning system could predict biomarkers of kidney and haematological disease in datasets that had both similar and markedly different characteristics than the developmental dataset. Additionally, it could identify markers of liver disease in the dataset that was most similar to the developmental dataset. Our study shows that markers of multiple, diverse systemic diseases could be predicted from external eye photographs.

Implications of all the available evidence

Our study shows that deep learning could be leveraged to predict biomarkers of multiple systemic diseases from photographs of the external eye alone. The ability to identify individuals at risk for systemic disease with an accessible, non-invasive imaging modality could enable prescreening tools that identify high-risk patients for confirmatory testing, helping to reduce barriers to care for individuals living in resource-limited areas and providing earlier detection of disease. The images in this work were captured using a fundus camera, so further studies are needed to assess the generalisability of this deep learning system to images captured by more accessible devices such as smartphone cameras.

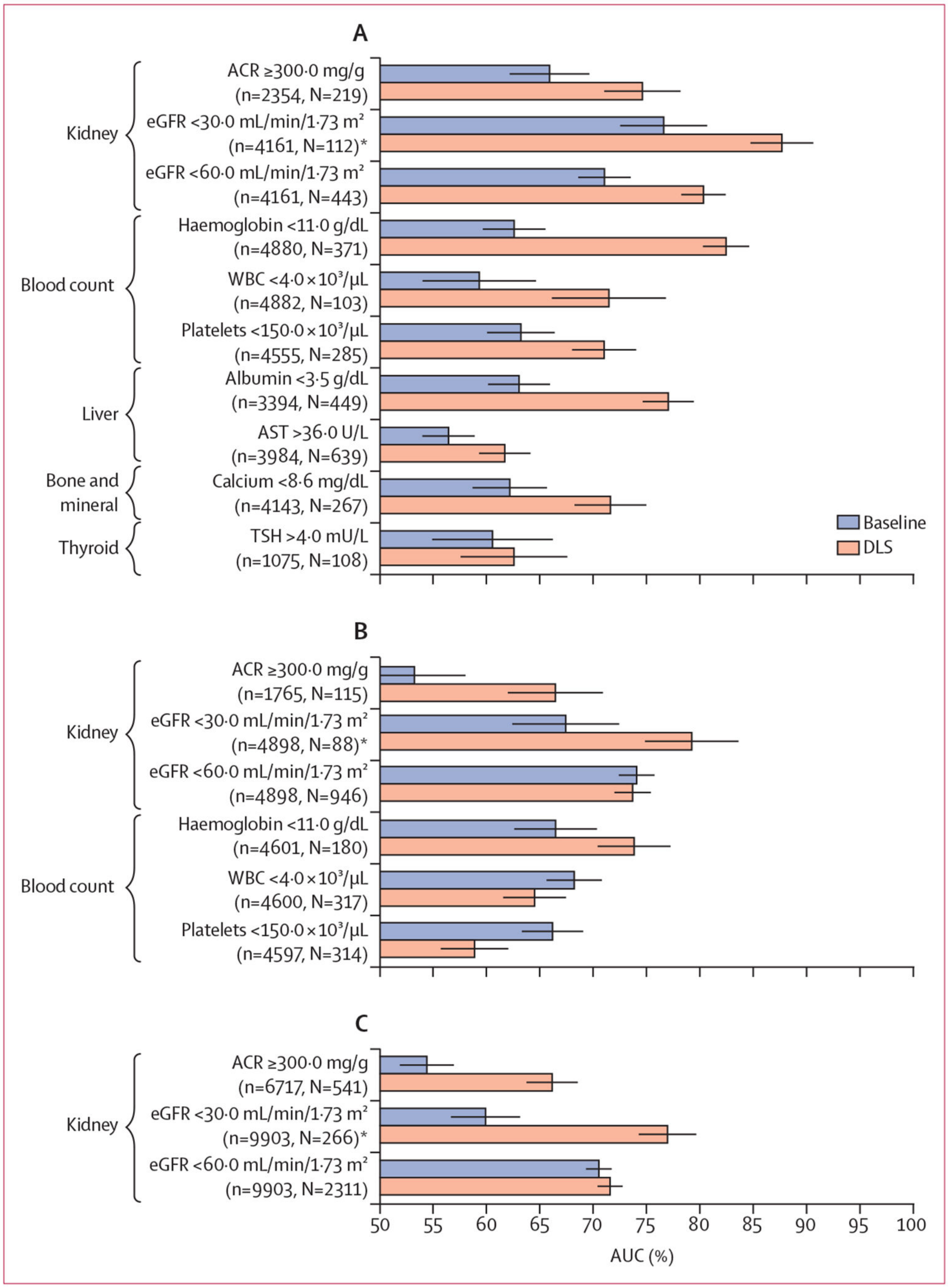

Figure 1: Comparison of AUC of the baseline model and the DLS

Results are presented for validation sets A, B, and C. n refers to the total number of datapoints, and N refers to the number of positives. Error bars show 95% CIs computed using the DeLong method. Appendix p 26 contains these results in tabular format, and includes p values. ACR=albumin-to-creatinine ratio. AST=aspartate aminotransferase. AUC=area under the receiver operating characteristic curve. DLS=deep learning system. eGFR=estimated glomerular filtration rate. TSH=thyroid stimulating hormone. WBC=white blood cells. *Indicates that the target was prespecified as secondary analysis; all others were prespecified as primary analysis.

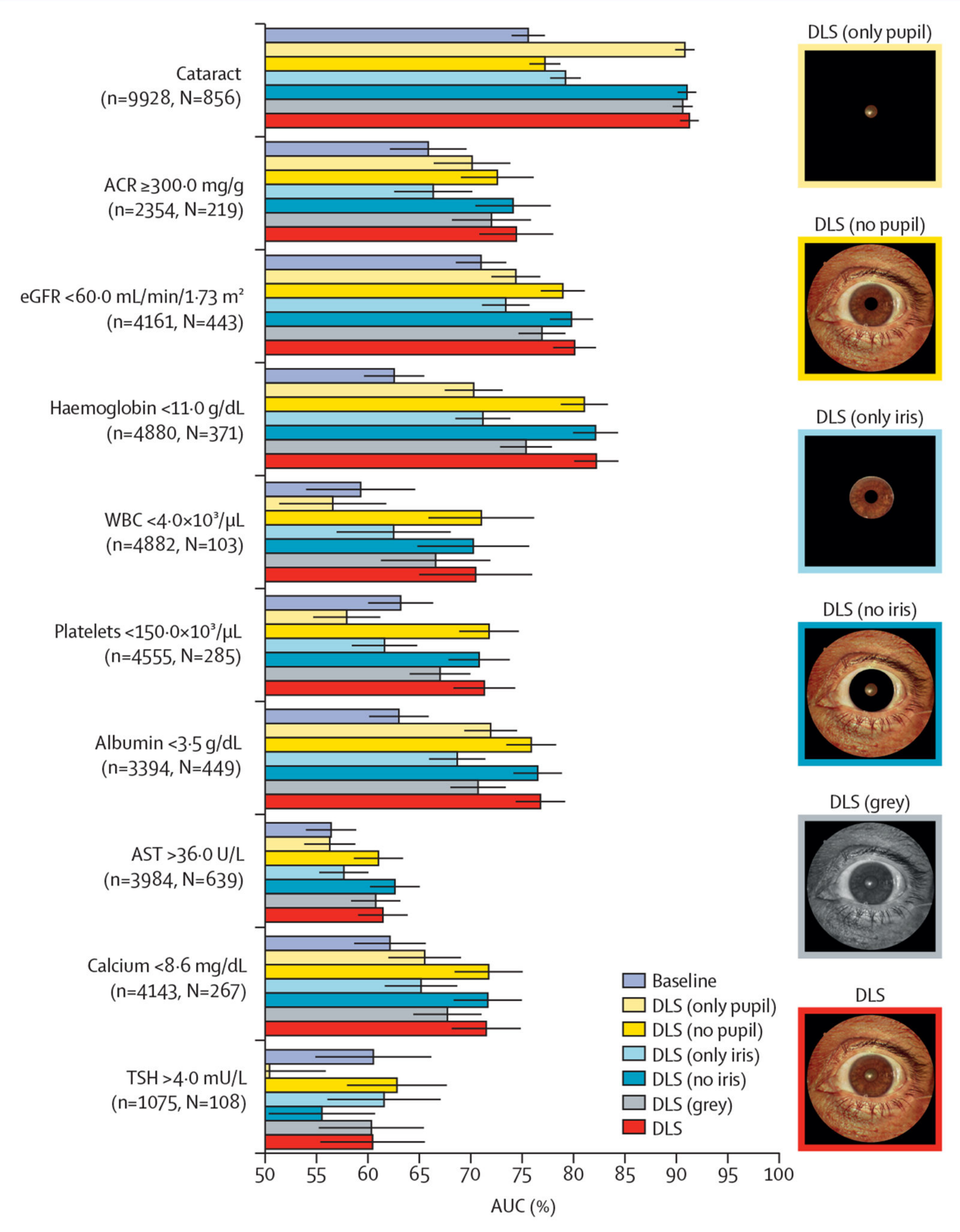

Figure 2: Experiments masking different image regions or removing colour

Results are presented for validation set A. n refers to the total number of datapoints, and N refers to the number of positives. The reported performances are for models both trained and evaluated on images with the specified manipulations. Manipulations described as “only X” refer to images with only that part of the anatomy visible; “no X” refers to images with that region masked, but every other area visible; “grey” refers to images converted to greyscale to remove colour. Error bars show 95% CIs computed using the DeLong method. For details, see appendix p 6. Note that all DLS results here are single models rather than ensembles. ACR=albumin-to-creatinine ratio. AST=aspartate aminotransferase. AUC=area under the receiver operating characteristic curve. DLS=deep learning system. eGFR=estimated glomerular filtration rate. TSH=thyroid stimulating hormone. WBC=white blood cells.

Table:

Dataset characteristics

	Development datasets		Validation datasets		
	Training dataset	**Tuning dataset**	**A**	**B**	**C**
Source	EyePACS/LACDHS (training)	EyePACS/LACDHS (tuning)	EyePACS/LACDHS (testing)	VA TECS	VA TRI
Geographical location	California *	California (independent sites from training and testing) *	California (independent sites from training and tuning) *	Georgia	Georgia
Number of patients	38 398	8253	9928	5552	10 030
Number of visits	62 213	14 245	15 023	5737	12 810
Number of images	123 130	28 137	29 566	11 161	23 671
Mean age, years (SD)	56·4 (10·3)	56·7 (9·7)	56·4 (9·9)	61·8 (12·6)	63·1 (9·8)
Sex					
Female	22 985 (59·9%)	5171 (62·7%)	5423 (54·6%)	846 (15·2%)	466 (4·6%)
Male	15 403 (40·1%)	3081 (37·3%)	4505 (45·4%)	4706 (84·8%)	9564 (95·4%)
Unknown	10 (<0·1%)	1 (<0·1%)	0	0	0
Race and ethnicity					
Hispanic	25 478 (66·4%)	4997 (60·5%)	7986 (80·4%)	0	0
White	1777 (4·6%)	501 (6·1%)	363 (3·7%)	2384 (42·9%)	2244 (22·4%)
Black	3680 (9·6%)	642 (7·8%)	602 (6·1%)	3098 (55·8%)	2046 (20·4%)
Asian/Pacific islander	2437 (6·3%)	915 (11·1%)	438 (4·4%)	35 (0·6%)	29 (0·3%)
Native American	34 (<0·1%)	26 (0·3%)	4 (<0·1%)	35 (0·6%)	20 (0·2%)
Other	502 (1·3%)	76 (0·9%)	535 (5·4%)	0	0
Unknown	4490 (11·7%)	1096 (13·3%)	0	0	5691 (56·7%)
Has diabetes	38 398 (100%)	8253 (100%)	9928 (100%)	2349 (42·3%)	10 030 (100%)
Median time with diabetes, years (IQR)	8·0 (2·0–13·0)	8·0 (3·0–13·0)	8·0 (2·0–13·0)	NA	NA
Time with diabetes unknown	694 (1·8%)	197 (2·4%)	0	5552 (100%)	10 030 (100%)

Baseline variables (age, sex, race and ethnicity, and years with diabetes) were not available for all patients. For categorical variables, we specify the number of patients per variable value and the percentage of available data. See appendix p 16 for a detailed flowchart of how data were processed. LACDHS=Los Angeles County Department of Health Services. VA=Veterans Affairs. TECS=Technology-based Eye Care Services. TRI=TeleRetinal Imaging. NA=not available.

[*] Splitting was done by site to ensure the setup constituted external validation (splitting by location).