

1 **Diabetic retinopathy environment-wide association study (EWAS) in NHANES 2005-8.**

2

3 Kevin Blighe, Ph.D.<sup>1</sup>, Sarega Gurudas, MSC.<sup>1</sup>, Ying Lee, Ph.D.<sup>1</sup>, Sobha Sivaprasad,  
4 FRCOphth.<sup>1,2,\*</sup>

5

6 <sup>1</sup>University College London, Institute of Ophthalmology, London, EC1V 9EL, United  
7 Kingdom.

8 <sup>2</sup>National Institute for Health Research, Moorfields Biomedical Research Centre, London,  
9 United Kingdom.

10

11 \* corresponding author:

12 Professor Sobha Sivaprasad

13 sobha.sivaprasad@nhs.net

14 +44 (0)20 7566 2039

15

16 **SUMMARY**

17 **Background:** Several circulating biomarkers are reported to be associated with diabetic  
18 retinopathy (DR). However, their relative contributions to DR compared to known risk  
19 factors, such as hyperglycemia, hypertension, and hyperlipidemia, remain unclear. In this  
20 data driven study, we used novel models to evaluate the associations of over 400 laboratory  
21 parameters with DR.

22 **Methods:** We performed an environment-wide association study (EWAS) of laboratory  
23 parameters available in National Health and Nutrition Examination Survey (NHANES)  
24 2007-8 in individuals with diabetes with DR as the outcome (test set). We employed  
25 independent variable ('feature') selection approaches, including parallelized univariate  
26 regression modeling, Principal Component Analysis (PCA), penalized regression, and  
27 RandomForest™. These models were replicated in NHANES 2005-6 (replication set).

28 **Findings:** The test and replication set consisted of 1025 and 637 individuals with available  
29 DR status and laboratory data respectively. Glycohemoglobin (HbA1c) was the strongest risk  
30 factor for DR. Our PCA-based approach produced a model that incorporated 18 principal  
31 components (PCs) that had AUC 0.796 (95% CI 0.761-0.832), while penalized regression  
32 identified a 9-feature model with 78.51% accuracy and AUC 0.74 (95% CI 0.72-0.77).  
33 RandomForest™ identified a 31-feature model with 78.4% accuracy and AUC 0.71 (95% CI  
34 0.65-0.77). On grouping the selected variables in our RandomForest™, hyperglycemia alone  
35 achieved AUC 0.72 (95% CI 0.68-0.76). The AUC increased to 0.84 (95% CI 0.78-0.9) when  
36 the model also included hypertension, hypercholesterolemia, hematocrit, renal and liver  
37 function tests.

38 **Interpretation:** All models showed that the contributions of established risk factors of DR  
39 especially hyperglycemia outweigh other laboratory parameters available in NHANES.

40

41 **RESEARCH IN CONTEXT**

42 What is already known about this subject?

- 43     ▪ There are >500 publications that report associations of candidate circulating  
44     biomarkers with diabetic retinopathy (DR).  
45     ▪ Although hyperglycemia, hypertension, and hyperlipidemia are established risk  
46     factors, they do not always explain the variance of this complication in people with  
47     diabetes; DR also shares risk factors with other diabetes complications including  
48     markers of renal and cardiovascular disease.  
49     ▪ ‘Holistic’ studies that quantify risk across all of these parameters combined are  
50     lacking.

51

52 What is the key question?

- 53     ▪ It is unclear whether risk models for DR may be improved by adding some of these  
54     reported biomarkers - there is an unmet need to systematically evaluate as many  
55     circulating biomarkers as possible to help rank their associations with DR.

56

57 What are the new findings?

- 58     ▪ We show that hyperglycemia is the strongest risk factor across all models.  
59     ▪ We stratified the rest of the highest ranked parameters into groups related to diabetes  
60     control, renal and liver function, and hematocrit changes.

61

62 How might this impact on clinical practice in the foreseeable future?

- 63     ▪ The importance of focusing on parameters beyond hyperglycemia control to reduce  
64     risk of progression from diabetes to DR is emphasized.

65

## 66 INTRODUCTION

67 Diabetes represents the most common cause of microvascular changes in the retina. The  
68 initial retinal lesions of diabetic retinopathy (DR) are microaneurysms but they can occur in  
69 eyes with and without diabetes (1-3). With increasing duration of diabetes, other lesions  
70 develop and co-exist in the retina such as retinal hemorrhages, exudates, intraretinal  
71 microvascular abnormalities and neovascularization of the retina or optic disc. Based on the  
72 presence of individual lesions or a constellation of them, DR severity level is graded from  
73 mild, moderate and severe non-proliferative diabetic retinopathy (NPDR) to proliferative  
74 diabetic retinopathy (PDR) (4, 5). Diabetic macular edema (DME) can occur in any stage of  
75 DR (5). In population-based studies, approximately a third of people with diabetes have DR  
76 (6, 7). The established systemic risk factors for DR are suboptimal control of hyperglycemia,  
77 hypertension and hyperlipidemia (8, 9). Hypertension can also cause some of these retinal  
78 lesions independent of diabetes (10).

79 There are several laboratory parameters that have been shown to be abnormal in people with  
80 DR such as hyperuricemia (11), low vitamin D levels (12), low thyroxine levels (13), anemia  
81 (14), oxidative stress and inflammatory markers (15). In addition, DR is also associated with  
82 markers of diabetic kidney disease including microalbuminuria and serum creatinine (16, 17)  
83 and cardiovascular disease markers such as raised C-reactive protein (CRP) (18). Most of  
84 these associations and risks of DR are reported based on analysis of candidate laboratory-  
85 based serum or urinary markers.

86 In addition to these risk factors, there are several other non-modifiable and modifiable  
87 risk factors that have been attributed to the development and progression of DR. Some  
88 of these include age of onset of diabetes, duration of diabetes, male sex, and ethnicity  
89 (19-21).

90 There is an unmet need to rank these reported retinal, systemic and laboratory risk  
91 factors to understand their relative contributions or associations with DR in people  
92 with diabetes. The National Health and Nutrition Examination Survey (NHANES -  
93 <https://wwwn.cdc.gov/Nchs/Nhanes/>) was initiated in the 1960s in order to examine the  
94 health and nutritional status of US citizens and has been surveying the population up to  
95 the present time. Since 1999, it has examined ~5000 citizens per year and includes  
96 various topics, including cardiovascular disease, diabetes, environmental exposures,

97 eye diseases, hearing loss, infectious diseases, kidney disease, nutrition, etc. The data  
98 also contains several laboratory markers including environmental toxins, allergens,  
99 pollutants.

100 In this study, we used an environment wide association study (EWAS) methodology  
101 (22-25) on NHANES 2007-8 to evaluate the rank order of systemic and laboratory  
102 risks of DR among individuals identified as having diabetes to evaluate their relative  
103 associations with DR. Our findings are then replicated in NHANES 2005/6. Our  
104 objective is not to only use previously reported risk factors but also provide new  
105 research avenues from this data driven agnostic modelling study.

## 106 **METHODS**

### 107 **Study data preparation**

108 We used National Health and Nutrition Examination Survey (NHANES) 2007-8 as our  
109 primary cohort and 2005-6 as a replication cohort. Both datasets were prepared in the  
110 same fashion, however, for ease of interpretation, the following methods describe  
111 2007-8. Specifically, three main categories of data were used: examination data  
112 (*Ophthalmology - Retinal Imaging data*; OPXRET\_E), demographics data (DEMO\_E),  
113 and laboratory data (**Figure 1** footnote). The main outcome of interest in the  
114 examination data was *4 levels retinopathy severity, worse eye* (OPDURL4) – this  
115 variable was recoded as binary with levels: no retinopathy; retinopathy (including mild  
116 NPR, moderate/severe NPR, and proliferative). All datasets were downloaded as SAS  
117 XPORT (xpt) format and read into R (v4.0.2) via the *Hmisc* package.

118 Individuals with a missing value in the main outcome variable were removed before  
119 aligning the examination, demographics, and laboratory data via each individual's  
120 respondent sequence number (SEQN). This dataset was then further filtered for only  
121 those individuals who had diabetes (**Figure 1**). Variables were removed from the data  
122 that had 0 variance (i.e. constant values) (**Supplementary Table 1**). Prior to any  
123 analysis, in addition, any variable that contained a single value occupying > 90% of  
124 total values was removed, as were variables that had > 90% missingness. Further  
125 specific filtering and encoding was then applied per dataset. [A] Examination data:  
126 variables that were different encodings of the main outcome were removed; variables

127 that related to the status of the examination appointment were removed; OPDUHMA  
128 was removed, as it is a combination of 2 other variables that were retained (OPDUMA  
129 and OPDUHEM); variables related to glaucoma, for which there is already a single  
130 variable, were removed; variables related to the left or right eye where there was  
131 already a variable for ‘worse’ eye were removed; values encoded as missing were  
132 recoded as NA; and all other remaining variables were encoded as binary, with 0  
133 representing the absence of the condition, and 1 representing any recorded presence (at  
134 any level) of the condition. [B] Demographic data: variables associated with  
135 interpreters and the language of the interview were removed; variables that were  
136 duplicates or different encoding of each were removed. [C] Laboratory data:  
137 categorical variables were removed and only continuous retained; duplicate variables  
138 related to the oral glucose tolerance test (OGTT\_E) were removed; variables related to  
139 time since domestic activities (‘pump gas’, ‘shower’, etc.) were removed; variables  
140 that were duplicates or different encodings of each were removed; variables measured  
141 on the imperial system of weights and measures were removed if they had a  
142 corresponding variable in SI units. We focused only on continuous laboratory variables  
143 for the following reasons: 1, in NHANES, the majority of categorical variables are  
144 derived from the continuous variables; 2, our PCA-based approach can only work on  
145 continuous variables; 3, for RandomForest™, having continuous variables increases  
146 the number of splitting points in the data, and metrics of importance such as Gini are  
147 known to exhibit less bias on such data (26).

#### 148 **Diabetes status**

149 To define the diabetes status for each individual, questionnaire data (DIQ\_E) was used  
150 in addition to variables already included in the laboratory data. Diabetes status was  
151 then defined as an individual satisfying any of the following: Self-reported diabetes  
152 (DIQ010); on anti-diabetes drugs (DIQ070); taking insulin (DIQ050); fasting blood sugar  
153 (FBS)  $\geq 6.1$  (110mg/dl) (LBDGLUSI); random blood sugar (RBS)  $\geq 11.1$  (200mg/dl)  
154 (LBDSGLSI); oral glucose tolerance test (OGTT)  $\geq 200$ mg/dl (LBDGLTSI);  
155 Glycohemoglobin (HbA1c)  $\geq 6.5\%$  (LBXGH).

#### 156 **Covariates**

157 Age, ethnicity, and diabetes duration were used as covariates. Diabetes duration was  
158 calculated as age at screening minus the age at which the individual was first informed  
159 that he/she had diabetes.

## 160 **Statistical analysis**

161 Prior to statistical analysis, continuous laboratory variables were logged ( $\log_e$ ) and then  
162 transformed into z-scores to ensure that these were on the same scale. In regression  
163 analysis, the complex sampling design of the NHANES dataset was accounted for  
164 through use of survey sampling weights via the *survey* package in R / CRAN. To do  
165 this, the following value-pairs were used with the *svydesign* function: (*id*, SDMVPSU;  
166 *strata*, SDMVSTRA; *weights*, WTMEC2YR; *nest*, TRUE).

167 Univariate analysis was performed on all candidate predictors using a survey-weighted  
168 compute-parallelized logistic regression model via the R / Bioconductor package  
169 *RegParallel*, adjusting for age, ethnicity, and duration of diabetes separately. The  
170 Benjamini-Hochberg (27) procedure was used to control the type I error false discovery  
171 rate (FDR). A customized Manhattan plot was generated using *ggplot2*, while pairwise  
172 scatter and correlation plots were generated via a customized pairs plot. Finally, a  
173 heatmap was generated via the R / Bioconductor package *ComplexHeatmap*.

174 As our study is also hypothesis-generating, multivariate approaches based on principal  
175 component analysis (PCA), penalized regression, and the RandomForest™  
176 classification algorithm were additionally used. Variables were pre-filtered and  
177 prepared as per univariate testing. Principal component analysis was performed via the  
178 R / Bioconductor package *PCAtools*. After conducting PCA, each eigenvector was then  
179 independently regressed against retinopathy outcome via binary logistic regression and  
180 those that passed  $p \leq 0.05$  were used to construct a multivariable model that was further  
181 tested in ROC analysis via the *pROC* package in R.

182 Separately, as model complexity and multi-collinearity can arise from a large number  
183 of predictors, elastic net regularization (penalized regression with L1 and L2 penalties  
184 of the Lasso and Ridge methods) was used to reduce the number of predictor variables  
185 using *glmnet* in R / CRAN. To fit the model, 100x cross-validation was used and alpha  
186 ( $\alpha$ ) set to 0.5. The final chosen variables were those whose coefficients were not

187 shrunk to zero – these were plot as violin plots with scatter overlays to show  
188 differences between non-DR and DR via *ggplot2*. To determine accuracy, model  
189 predictions were made on the data using the lambda ( $\lambda$ ) one-standard-error rule using  
190 the *predict* function from the *stats* package in R.

191 Finally, the RandomForest™ (RF) model was fitted via the *randomForest* R / CRAN  
192 package. For this, the dataset was divided randomly into 50% training and 50%  
193 validation. Prior to model fitting, the initial model was tuned using functionality  
194 provided by the *caret* package in R / CRAN, as follows: 1), a 10x cross-validation  
195 control function was defined via *trainControl* function; 2) the best value for ‘mtry’,  
196 i.e., the ideal number of variables to randomly sample, was determined using the *train*  
197 function across a search / tuning grid ranging between 1-40 and with Kappa as the  
198 metric; and 3) using the selected value of ‘mtry’, the ideal number of trees, ‘ntrees’,  
199 was determined also via the *train* function with selection metric based on Kappa. After  
200 the initial model was fit, variables with mean decrease in accuracy  $\leq 1\%$  were excluded  
201 and the model re-fit. This was then repeated in a recursive fashion until all variables  
202 with negative mean decrease accuracy were removed from the model.

203

#### 204 **Final risk models**

205 Variables selected from RandomForest™ were grouped based on similarity of function or  
206 clinical use. Each group was then used to create independent univariate or multivariable  
207 binary logistic regression models with DR as the end-point. A single Wald test p-value was  
208 derived for each model using *wald.test* from the *aod* package. ROC analysis was performed  
209 using *pROC*. McFadden’s and Nagelkerke’s pseudo- $R^2$  were derived via the *pscl* and *rms*  
210 packages, respectively.

211

#### 212 **Role of the funding source**

213 The funders had no role in study design, data collection, data analysis, data interpretation,  
214 writing, editing the report, or the decision to submit for publication.

215

## 216 **RESULTS**

### 217 **Study cohort**

218 In NHANES 2007-8, retinal imaging data is available for 3863 individuals, demographics  
219 data is available for 10149 individuals, and laboratory data is available for between 394 and



220 9307 individuals, depending on the individual laboratory dataset in NHANES (see **Figure 1**  
221 footnote). After aligning all data and filtering for those who had diabetes by our  
222 classification, 1025 individuals remained in our dataset. The selection process is illustrated in  
223 **Figure 1**, while **Table 1** provides an overview of the demographics of these individuals.

224

225 For our replication cohort, NHANES 2005-6, we prepared laboratory data following the same  
226 filter criteria as NHANES 2007-8 and produced a final dataset of 2459 individuals, among  
227 which 637 (with retinopathy, 176; no retinopathy, 461) had diabetes.

228

### 229 **Retinal lesions of diabetic retinopathy**

230 To help validate our methodology and cohort selection, we aimed to determine retinal lesions  
231 that define DR. To this end, we identified 9 retinal lesions in NHANES 2007-8 that were  
232 statistically significantly associated with DR and survived to p-value adjustment for false  
233 discovery (**Table 2**). The top lesions were retinal microaneurysms ( $p \leq 0.0001$ ), followed by  
234 retinal hard exudates (typically due to lipoprotein deposition in the retina and may be  
235 associated with macular edema) ( $p \leq 0.0001$ ). Other key lesions at  $p \leq 0.0001$  were retinal soft  
236 exudate (now termed cotton wool spots), retinal blot hemorrhages, intraretinal microvascular  
237 abnormalities (IRMA), and macular edema. In NHANES, retinal microaneurysms and retinal  
238 blot hemorrhages are encoded to be mutually exclusive, i.e., an individual is recorded as  
239 having retinal microaneurysms only when not accompanied with retinal blot hemorrhages,  
240 and *vice-versa* (**Table 2**).

241

### 242 **Univariate logistic regression analysis**

243 In total, 6 variables reached statistical significance in the unadjusted univariate analysis, 11  
244 after adjustment for age, 2 after adjustment for ethnicity, and 7 for diabetes duration  
245 (**Supplementary Figure 1; Table 3**). Glycohemoglobin (HbA1c) was the only variable that  
246 was statistically significant in both the unadjusted and adjusted analyses. Other risk variables  
247 of note that reached statistical significance in the unadjusted analysis included serum glucose  
248 (mmol/L) (i.e., RBS), osmolality (mmol/Kg), urinary albumin (mg/L), and fasting glucose  
249 (mmol/L) (i.e., FBS). The only protective variable, i.e., negatively associated, was  
250 hemoglobin (g/dL). These variables indicate suboptimal diabetes control, abnormal kidney  
251 function and presence of anemia as risk factors for DR. There was evidence of co-correlation  
252 among these statistically significant variables from the unadjusted analysis (**Supplementary**  
253 **Figure 2**).

254 Interestingly, after adjustment for diabetes duration, the following variables reached  
255 statistical significance: HbA1c (%), osmolality (mmol/Kg), urinary iodine ( $\mu\text{g/L}$ ), urinary  
256 cobalt ( $\mu\text{g/L}$ ), urinary triclosan (ng/mL), urinary creatinine ( $\mu\text{mol/L}$ ), and urinary barium  
257 ( $\mu\text{g/L}$ ).

258

### 259 **Principal component analysis**

260 Unsupervised PCA using all laboratory variables revealed that 59 PCs could account for 80%  
261 or more variation in the dataset. Eighteen PCs were statistically significantly associated with  
262 DR at  $p \leq 0.05$  via independent binomial regression models testing each PC (**Supplementary**  
263 **Table 2**). The top variables responsible for variation along these PCs included measures of  
264 blood glucose (HbA1c, random blood glucose and fasting blood glucose), kidney function  
265 markers (urinary albumin, blood urea nitrogen [BUN]), hematological markers (hematocrit,  
266 hemoglobin, red blood cell distribution width), inflammatory markers (CRP), white blood  
267 cell count, urinary nitrates, segmented neutrophil count), and toxic elements (urinary  
268 beryllium and cotinine) among others – these PCs were also statistically significantly  
269 correlated to microaneurysms, the previously-identified top retinal lesion, and the covariates  
270 used during univariate testing (**Supplementary Figure 3**). Through ROC analysis, these 18  
271 PCs achieved AUC 0.796 (95% CI: 0.761-0.832).

272

### 273 **Penalized regression model**

274 We fitted an unbiased elastic-net penalized regression model to the laboratory variables and  
275 cross-validated it 100x. The model selected 9 variables whose coefficients were not shrunk to  
276 zero: urinary albumin, BUN, urinary cobalt, CRP, HbA1c, blood osmolality, serum  
277 potassium, systolic blood pressure, and urinary nitrate (**Figure 2**). Of note, these  
278 measurements mainly represent diabetes and blood pressure control and kidney function. This  
279 model had an accuracy of 78.51% and AUC 0.74 (95% CI: 0.72-0.77) when predicted on the  
280 same dataset on which the model was produced.

281

### 282 **RandomForest™ classification model**

283 From our RandomForest™ model, HbA1c was the single best predictor of DR (mean  
284 decrease accuracy, 31.94%; Gini, 21.75) (**Table 4**). However, other notable variables of  
285 appreciable accuracy were markers of diabetes control (FBS, RBS) inflammation (CRP),  
286 kidney function (potassium, BUN, creatinine and urinary albumin), hematological markers

287 (hematocrit), systolic blood pressure, among others. The overall accuracy of the model on the  
288 validation cohort was 78.4% and AUC 0.71 (95% CI: 0.65-0.77).

289

### 290 **Replication cohort**

291 In the NHANES 2005-6 replication cohort, we performed penalized regression and  
292 RandomForest™ in the same way as per the 2007-8 cohort. Our penalized regression model  
293 identified urinary albumin (mg/L), cockroach IgE antibody (kU/L), HbA1c (%), hemoglobin  
294 (g/dL), and urinary nitrate (ng/mL), with a model accuracy of 73.16% and AUC 0.76 (95%  
295 CI: 0.73-0.78). RandomForest™ identified HbA1c (%) as the variable contributing most to  
296 accuracy (mean decrease 16.96%), with many other variables contributing appreciable  
297 accuracy to the overall model (**Supplementary Table 3**) - the overall model accuracy was  
298 72.98% and AUC 0.68 (95% CI: 0.61-0.75).

299

### 300 **Final clinical risk models**

301 The 31 features identified by RandomForest™ (**Table 4**) were grouped into different  
302 categories of blood tests according to diabetes status, hematocrit values, blood pressure (BP),  
303 immune markers, renal function, sterols, toxins and metals, and liver function. When  
304 modeled against DR outcome, each group varied in performance; diabetes tests alone  
305 achieved AUC 0.72 (95% CI: 0.68-0.76). A final clinical risk model comprising diabetes tests,  
306 BP, renal and liver function tests, hematocrit values, circulating sterols and immune markers  
307 achieved AUC 0.84 (95% CI: 0.78-0.9) ( $p=0.00013$ ) (Nagelkerke  $R^2$  0.36) (**Table 5; Figure**  
308 **3**).

309

310

### 311 **DISCUSSION**

312 This EWAS of NHANES 2007-8 data with DR outcomes in individuals with diabetes  
313 included an unbiased feature selection approach based on a rudimentary univariate regression  
314 enabled for compute parallelization, PCA, penalized regression, and RandomForest™ of a  
315 large number of laboratory parameters. In contrast, epidemiological studies are typically  
316 conducted based on pre-conceived hypotheses and involve a single or just a few variables.  
317 These methods can be scaled to datasets of any size and therefore provide ways of working  
318 with large clinical and epidemiological datasets for the purpose of searching for novel  
319 hypotheses that could then lead to further focused investigations.

320

321 In our rudimentary approach, which is ultimately running many univariate models in a  
322 parallelized fashion, HbA1c was the only variable to reach statistical significance after  
323 adjusted for age, ethnicity, and diabetes duration. The relationship between HbA1c and DR  
324 has been explored extensively and was selected as the strongest risk factor in every approach  
325 we undertook, with a mean decrease accuracy of 31.94% via RandomForest™.

326

327 Our penalized regression and RandomForest™ algorithms also identified an association  
328 between elevated systolic blood pressure—but not diastolic—and DR (mean decrease  
329 accuracy, 1.9%), again confirming literature (10, 28-30). Further risk variables identified by  
330 both penalized regression and RandomForest™ were renal function tests including BUN,  
331 urinary albumin, potassium, osmolality, and urinary nitrate. These confirm the strong  
332 association of DR with markers of impaired kidney function. Other known risk factors that  
333 contributed higher up in the ranking order include hematocrit (%) and cholesterol. Although  
334 HbA1c has the strongest association with DR, our study highlight how the addition of other  
335 clinical parameters, e.g., from renal and liver function and hematocrit can increase the  
336 sensitivity and specificity of predicting DR outcome, with our final clinical risk model  
337 achieving AUC 0.83 (95% CI: 0.77-0.89) ( $p=0.00012$ ) (Nagelkerke  $R^2$  0.33), higher than any  
338 traditional diabetes control parameter in isolation or in combination.

339

340 The EWAS methodology and our RandomForest™ approach of non-targeted recursive  
341 feature also indicates a small contribution from toxins and metals, including 3-  
342 hydroxyphenanthrene, 9-hydroxyfluorene, phthalates, blood o-Xylene, and blood  
343 nitromethane. Therefore, retina may be a target tissue for environmental contamination. Some  
344 of the associations provide directions to future mechanistic research in DR. For example, we  
345 found that retinal microaneurysms (FDR-adjusted  $p \leq 0.0001$ ), the most statistically significant  
346 retinal lesions in individuals with DR, is already correlated with some of the variables such as  
347 HbA1c, CRP, BUN, beryllium, and hematocrit, suggesting early effects. In contrast,  
348 increased urinary cobalt, triclosan and barium became significant only when adjusted for  
349 duration of diabetes. Most of these parameters are also linked to risk of allergies and lung  
350 disease, an association that has not been previously explored systematically.

351 As this is a cross-sectional study, a cause-effect relation cannot be established. Moreover, we  
352 are unable to rule out any confounding effects of any unmeasured factors. On the other hand,  
353 the main strength of the study is the use of the well characterised NHANES cohort in whom  
354 standardised protocols were used to measure laboratory parameters. We are not aware of any

355 other association studies in DR where over 400 laboratory parameters were analysed  
356 simultaneously to develop multiple models. As the top variables of all four data driven  
357 agnostic models were similar, we also believe our findings are generalisable.

358

### 359 **CONCLUSION**

360 We confirm that DR is a complex disease and that the already established risk factors  
361 contribute significantly to the risk models of DR, with HbA1c being the strongest risk factor.  
362 Although our model provides an accuracy of approximately 80%, it also provides  
363 mechanistic insights into future research on DR including interrogating the interaction of  
364 low-ranking risk factors with more established factors in the models and highlights need to  
365 explore epigenetic screens to gauge better how risk factors influence gene expression. Most  
366 importantly, the study reinforces the need to control known risk factors of DR, especially  
367 hyperglycemia.

368

369

### 370 **ACKNOWLEDGEMENTS**

371 We thank the many thousands of NHANES study participants who, over the course of  
372 decades, have provided valuable information for epidemiological studies.

373

### 374 **FUNDING**

375 This work was funded by Global Challenges Research Fund and UK Research and  
376 Innovation through the Medical Research Council grant number MR/P027881/1. The  
377 research was supported] by the National Institute for Health Research (NIHR) Biomedical  
378 Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute  
379 of Ophthalmology. The views expressed are those of the author(s) and not necessarily those  
380 of the NHS, the NIHR or the Department of Health.

381

### 382 **AUTHOR CONTRIBUTIONS**

383 KB, SG and SS conceived and designed the study. KB, SG and YL analysed the data. KB is  
384 the study data guarantor. All authors interpreted the results and reviewed the manuscript. KB  
385 had full access to all of the data and takes responsibility for the integrity of the data and the  
386 accuracy of data analysis. All authors read and approved the final manuscript.

387

388

389

390

391

392

393 **TABLES**

394 **Table 1. Demographic overview of study cohort.**

Characteristics mean (±SD) n (%)		Diabetes with no diabetic retinopathy (n=787)	Diabetic retinopathy (n=238)	p-value	β-coefficient	OR (95% CI)
Age	-	62.39 (±11.02)	63.53 (±10.54)	0.16	0.01	1.00 (1.00-1.02)
Sex	Male	420 (53.37)	126 (52.94)	-	-	-
	Female	367 (46.63)	112 (47.06)	0.91	0.02	1.02 (0.76-1.36)
Ethnicity	Non-hispanic white	369 (46.88)	91 (38.23)	-	-	-
	Non-hispanic black	166 (21.09)	74 (31.09)	0.0012	0.59	1.81 (1.27-2.58)
	Mexican-American	138 (17.54)	41 (17.23)	0.38	0.19	1.21 (0.79-1.83)
	Other Hispanic	89 (11.31)	28 (11.77)	0.32	0.24	1.28 (0.79-2.07)
	Other race - including multi-racial	25 (3.18)	4 (1.68)	0.43	-0.43	0.65 (0.22-1.91)
Education	Less than 9 <sup>th</sup> grade	148 (18.8)	54 (22.69)	-	-	-
	9-11 <sup>th</sup> grade <sup>±</sup>	148 (18.8)	51 (21.43)	0.8	-0.06	0.94 (0.61-1.47)
	High school graduate / GED or equivalent	199 (25.29)	56 (23.53)	0.24	-0.26	0.77 (0.5-1.19)
	Some college or AA degree	172 (21.86)	55 (23.11)	0.55	-0.13	0.88 (0.57-1.35)
	College graduate or above	120 (15.25)	22 (9.24)	0.014	-0.69	0.5 (0.29-0.87)
Marital status	Married	464 (58.96)	133 (55.88)	-	-	-
	Widowed	119 (15.12)	36 (15.13)	0.8	0.05	1.06 (0.69-1.61)
	Divorced	90 (11.44)	37 (15.55)	0.099	0.36	1.43 (0.94-2.2)
	Separated	32 (4.07)	7 (2.94)	0.53	-0.27	0.76 (0.33-1.77)
	Never married	53 (6.73)	18 (7.56)	0.56	0.17	1.19 (0.67-2.09)
Family income	Living with partner	29 (3.68)	7 (2.94)	0.69	-0.17	0.84 (0.36-1.97)
	\$0-\$4999	11 (1.4)	3 (1.26)	-	-	-
	\$5000-\$9999	47 (5.97)	12 (5.04)	0.93	-0.07	0.94 (0.23-3.89)
	\$10000-\$14999	82 (10.42)	22 (9.25)	0.98	-0.02	0.98 (0.25-3.84)
	\$15000-\$19999	68 (8.64)	22 (9.25)	0.81	0.17	1.19 (0.30-4.64)
	\$20000-\$24999	69 (8.77)	25 (10.5)	0.68	0.28	1.33 (0.34-5.16)
	\$25000-\$34999	103 (13.09)	42 (17.65)	0.55	0.40	1.50 (0.40-5.63)
	\$35000-\$44999	68 (8.64)	23 (9.66)	0.76	0.22	1.24 (0.32-4.84)
	\$45000-\$54999	55 (6.99)	12 (5.04)	0.76	-0.22	0.80 (0.19-3.31)
	\$55000-\$64999	41 (5.21)	15 (6.3)	0.68	0.29	1.34 (0.33-5.48)
	\$65000-\$74999	36 (4.57)	6 (2.52)	0.53	-0.49	0.61 (0.13-2.86)
	\$75000-\$99999	44 (5.59)	12 (5.04)	1.00	0.00	1.00 (0.24-4.17)
	≥\$100000	88 (11.18)	17 (7.14)	0.62	-0.35	0.71 (0.18-2.81)
	Over \$20000	31 (3.94)	9 (3.78)	0.93	0.06	1.06 (0.24-4.66)
	Under \$20000	17 (2.16)	2 (0.84)	0.4	-0.84	0.43 (0.06-3.01)
Missing	27 (3.43)	16 (6.73)	-	-	-	
Diabetes duration	-	9.05 (±11.05)	16.33 (±12.57)	≤ 0.0001	0.05	1.05 (1.04-1.07)

395

396

397

Notes: This table only relates to those individuals who have been determined as having diabetes by our selection criteria: 1, Self-reported diabetes (DIQ010); 2, On anti-diabetes drugs (DIQ070); 3, Taking insulin (DIQ050); 4, Fasting blood sugar (FBS) ≥ 6.1 (110mg/dl) (LBDGLUSI); 5, Random blood sugar (RBS) ≥ 11.1 (200mg/dl)

398 (LBDSGLSI); 6, Oral glucose tolerance test (OGTT)  $\geq$  200mg/dl (LBDGLTSI); 7, Glycohemoglobin (HbA1c)  
399  $\geq$  6.5% (LBXGH). Ethnicity, education, and diabetes duration contain at least one term that is statistically  
400 significant.  $\ddagger$  includes 12<sup>th</sup> grade with no diploma

401

402

403 **Table 2. Retinal lesions that constitute diabetic retinopathy outcome.**

Retinal co-morbidity	n (%)	$\beta$ -coefficient	OR (95% CI)	p-value	FDR-adjusted p-value
Retinal microaneurysms only, worse eye	129 (12.59)	5.67	288.87 (127.66-653.68)	$\leq$ 0.0001	$\leq$ 0.0001
Retinal hard exudate, worse eye	86 (8.39)	3.72	41.34 (20.31-84.17)	$\leq$ 0.0001	$\leq$ 0.0001
Retinal blot hemorrhages, worse eye	47 (4.59)	3.36	28.71 (14.11-58.42)	$\leq$ 0.0001	$\leq$ 0.0001
Retinal soft exudate, worse eye	76 (7.41)	4.2	66.49 (26.44-167.21)	$\leq$ 0.0001	$\leq$ 0.0001
IRMA, worse eye	62 (6.05)	2.85	17.36 (9.06-33.26)	$\leq$ 0.0001	$\leq$ 0.0001
Macular edema, worse eye	51 (4.98)	3.88	48.24 (17.18-135.44)	$\leq$ 0.0001	$\leq$ 0.0001
Retinal fibrous proliferation, worse eye	19 (1.85)	3.41	30.19 (6.92-131.67)	$\leq$ 0.0001	$\leq$ 0.0001
Macular edema in center, worse eye	26 (2.54)	4.53	92.33 (12.45-684.9)	$\leq$ 0.0001	$\leq$ 0.0001
Retinal new vessels elsewhere, worse eye	15 (1.46)	3.12	22.68 (5.08-101.23)	$\leq$ 0.0001	0.0003

404 Notes: Lesions are taken from the NHANES *ophthalmology - retinal imaging* (OPXRET\_E) dataset. Only  
405 lesions with FDR-adjusted  $p \leq 0.05$  are listed. Soft exudate is now termed cotton wool spots.

**Table 3. Laboratory variables associated with retinopathy in individuals with diabetes.**

Description	Unadjusted / Non-covariate adjusted				Age-adjusted				Ethnicity-adjusted				Diabetes duration-adjusted			
	$\beta$ -coefficient	OR (95% CI)	p-value	FDR-adjusted p-value	$\beta$ -coefficient	OR (95% CI)	p-value	FDR-adjusted p-value	$\beta$ -coefficient	OR (95% CI)	p-value	FDR-adjusted p-value	$\beta$ -coefficient	OR (95% CI)	p-value	FDR-adjusted p-value
Glycohemoglobin (%)	0.82	2.27 (1.84-2.8)	0.000	0.0003	0.85	2.34 (1.87-2.92)	0.000	0.0011	0.83	2.28 (1.82-2.87)	0.000	0.0036	0.72	2.05 (1.55-2.73)	0.002	0.0341
Glucose, serum (mmol/L)	0.54	1.72 (1.42-2.08)	0.001	0.0047	0.57	1.77 (1.45-2.15)	0.001	0.0115	0.54	1.71 (1.41-2.09)	0.002	0.0334	0.36	1.43 (1.1-1.86)	0.180	0.0520
Osmolality (mmol/Kg)	0.49	1.63 (1.34-1.99)	0.002	0.0143	0.45	1.57 (1.3-1.9)	0.004	0.0127	0.49	1.63 (1.34-1.99)	0.005	0.0707	0.39	1.48 (1.12-1.94)	0.140	0.0415
Albumin, urine (mg/L)	0.45	1.57 (1.28-1.93)	0.006	0.0288	0.43	1.53 (1.24-1.88)	0.012	0.0127	0.42	1.53 (1.25-1.87)	0.017	0.1418	0.25	1.28 (0.98-1.68)	0.905	0.1617
Hemoglobin (g/dL)	-0.33	0.72 (0.61-0.85)	0.013	0.0387	-0.30	0.74 (0.63-0.88)	0.042	0.0234	-0.29	0.75 (0.62-0.9)	0.094	0.2638	0.00	1 (0.73-1.37)	0.784	0.9883
Fasting Glucose (mmol/L)	0.49	1.63 (1.28-2.07)	0.011	0.0387	0.51	1.66 (1.3-2.13)	0.013	0.0127	0.46	1.59 (1.25-2.02)	0.031	0.2290	0.22	1.24 (0.93-1.66)	0.687	0.2731
4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL), urine (ng/mL)	-0.26	0.77 (0.67-0.88)	0.021	0.0539	-0.20	0.82 (0.71-0.94)	0.133	0.0552	-0.28	0.75 (0.65-0.87)	0.033	0.2290	-0.25	0.78 (0.52-1.19)	0.655	0.4061
Iodine, urine (ug/L)	-0.17	0.84 (0.76-0.93)	0.039	0.0873	-0.21	0.81 (0.73-0.9)	0.022	0.0156	-0.15	0.86 (0.77-0.96)	0.181	0.2638	-0.29	0.75 (0.63-0.89)	0.054	0.0341
Cobalt, urine (ug/L)	-0.51	0.6 (0.44-0.82)	0.052	0.1033	-0.50	0.6 (0.45-0.82)	0.058	0.0300	-0.51	0.6 (0.44-0.81)	0.073	0.2638	-0.52	0.59 (0.45-0.78)	0.021	0.0341
Hematocrit (%)	-0.31	0.73 (0.6-0.89)	0.061	0.1089	-0.28	0.76 (0.62-0.92)	0.148	0.0595	-0.28	0.75 (0.62-0.92)	0.202	0.2638	-0.03	0.97 (0.7-1.35)	0.787	0.9329
Blood urea nitrogen (mmol/L)	0.33	1.4 (1.13-1.73)	0.075	0.1225	0.27	1.31 (1.01-1.71)	0.645	0.1584	0.35	1.43 (1.16-1.75)	0.066	0.2638	0.15	1.17 (0.87-1.57)	0.233	0.4793
Albumin (g/L)	-0.22	0.8 (0.69-0.93)	0.116	0.1736	-0.21	0.81 (0.69-0.94)	0.183	0.0687	-0.19	0.83 (0.7-0.99)	0.575	0.3225	-0.09	0.92 (0.74-1.14)	0.412	0.5899
Urinary Triclosan (ng/mL)	-0.42	0.65 (0.49-0.88)	0.126	0.1743	-0.40	0.67 (0.5-0.89)	0.165	0.0639	-0.42	0.66 (0.49-0.89)	0.195	0.2638	-0.60	0.55 (0.36-0.83)	0.135	0.0412
Mean cell hemoglobin (pg)	-0.24	0.79 (0.66-	0.159	0.2046	-0.27	0.77 (0.65-	0.076	0.0368	-0.17	0.84 (0.7-	0.069	0.4855	-0.01	0.99 (0.76-	0.174	0.9489



		0.93)				0.91)				1.02)				1.28)		
Lead, urine (µg/L)	-0.40	0.67 (0.49-0.91)	0.0 222	0.2668	-0.40	0.67 (0.49-0.91)	0.0 230	0.0831	-0.41	0.66 (0.49-0.9)	0.0 228	0.2638	-0.43	0.65 (0.43-0.99)	0.0 658	0.1312
Creatinine, urine (µmol/L)	-0.22	0.8 (0.67-0.97)	0.0 348	0.2723	-0.19	0.83 (0.68-1)	0.0 710	0.1664	-0.27	0.77 (0.64-0.92)	0.0 156	0.2638	-0.28	0.76 (0.64-0.9)	0.0 070	0.0341
Alanine aminotransferase (ALT) (U/L)	-0.25	0.78 (0.62-0.97)	0.0 380	0.2723	-0.20	0.82 (0.66-1.02)	0.0 965	0.2003	-0.22	0.81 (0.64-1.01)	0.0 869	0.4259	-0.09	0.92 (0.68-1.23)	0.5 656	0.6993
Creatinine (µmol/L)	0.24	1.27 (1.04-1.54)	0.0 335	0.2723	0.18	1.19 (0.96-1.48)	0.1 291	0.2529	0.20	1.23 (0.99-1.53)	0.0 933	0.4362	0.13	1.14 (0.89-1.45)	0.3 111	0.4699
Red blood cell count (million cells/µL)	-0.19	0.82 (0.7-0.97)	0.0 372	0.2723	-0.13	0.88 (0.74-1.04)	0.1 439	0.2740	-0.19	0.83 (0.7-0.98)	0.0 540	0.3172	0.01	1.01 (0.74-1.37)	0.9 623	0.9769
Mean cell volume (fL)	-0.21	0.81 (0.68-0.98)	0.0 463	0.2723	-0.25	0.78 (0.65-0.93)	0.0 159	0.0623	-0.15	0.86 (0.71-1.06)	0.1 865	0.6799	-0.05	0.95 (0.72-1.26)	0.7 361	0.8368
Platelet count (1000 cells/µL)	-0.21	0.81 (0.68-0.98)	0.0 435	0.2723	-0.18	0.83 (0.68-1.02)	0.1 005	0.2049	-0.23	0.79 (0.67-0.94)	0.0 216	0.2638	-0.29	0.75 (0.57-0.99)	0.0 616	0.1250
Mean platelet volume (fL)	0.24	1.27 (1.04-1.55)	0.0 352	0.2723	0.26	1.3 (1.06-1.59)	0.0 243	0.0835	0.22	1.25 (1.02-1.53)	0.0 564	0.3225	0.09	1.09 (0.83-1.43)	0.5 336	0.6642
Cotinine (ng/mL)	-0.15	0.86 (0.76-0.99)	0.0 451	0.2723	-0.09	0.91 (0.79-1.04)	0.1 994	0.3443	-0.18	0.84 (0.74-0.95)	0.0 221	0.2638	-0.06	0.94 (0.66-1.34)	0.7 352	0.8368
Insulin (pmol/L)	-0.32	0.73 (0.55-0.97)	0.0 469	0.2723	-0.30	0.74 (0.55-1)	0.0 699	0.1660	-0.29	0.75 (0.56-1)	0.0 734	0.3816	-0.22	0.8 (0.51-1.26)	0.3 520	0.5062
Blood cadmium (nmol/L)	-0.23	0.8 (0.65-0.97)	0.0 429	0.2723	-0.25	0.78 (0.64-0.95)	0.0 273	0.0866	-0.23	0.79 (0.64-0.98)	0.0 589	0.3225	-0.28	0.76 (0.5-1.16)	0.2 224	0.3528
Urinary perchlorate (ng/mL)	-0.25	0.78 (0.63-0.95)	0.0 294	0.2723	-0.24	0.78 (0.63-0.98)	0.0 495	0.1267	-0.24	0.78 (0.64-0.96)	0.0 399	0.2668	-0.31	0.73 (0.53-1.01)	0.0 818	0.1525
Urinary nitrate (ng/mL)	-0.28	0.75 (0.6-0.94)	0.0 264	0.2723	-0.23	0.79 (0.63-1)	0.0 675	0.1632	-0.27	0.76 (0.6-0.96)	0.0 422	0.2688	-0.34	0.71 (0.55-0.91)	0.0 186	0.0528
Cesium, urine (µg/L)	-0.33	0.72 (0.54-0.95)	0.0 339	0.2723	-0.32	0.72 (0.55-0.96)	0.0 390	0.1059	-0.33	0.72 (0.54-0.95)	0.0 389	0.2665	-0.34	0.71 (0.47-1.06)	0.1 161	0.1950
Thallium, urine (µg/L)	-0.40	0.67 (0.48-0.95)	0.0 417	0.2723	-0.39	0.68 (0.48-0.96)	0.0 488	0.1266	-0.42	0.66 (0.47-0.92)	0.0 307	0.2638	-0.48	0.62 (0.4-0.94)	0.0 414	0.0923
25OHD2+25OHD3 (nmol/L)	-0.24	0.79	0.0	0.2723	-0.25	0.78	0.0	0.0831	-0.18	0.83	0.1	0.5685	-0.20	0.82	0.2	0.3401

		(0.65-0.95)	279			(0.65-0.94)	230			(0.66-1.04)	390			(0.61-1.1)	129	
Blood Toluene (ng/mL)	-0.21	0.81 (0.68-0.96)	0.0 304	0.2723	-0.20	0.82 (0.69-0.97)	0.0 366	0.1029	-0.22	0.8 (0.65-0.98)	0.0 578	0.3225	-0.30	0.74 (0.45-1.21)	0.2 536	0.3896
C-reactive protein (mg/dL)	-0.19	0.82 (0.69-0.99)	0.0 551	0.3102	-0.18	0.83 (0.69-1.01)	0.0 827	0.1820	-0.25	0.78 (0.66-0.92)	0.0 153	0.2638	-0.27	0.77 (0.6-0.98)	0.0 489	0.1063
Barium, urine (µg/L)	-0.39	0.68 (0.47-0.99)	0.0 600	0.3114	-0.38	0.69 (0.48-0.99)	0.0 607	0.1525	-0.38	0.68 (0.47-1)	0.0 756	0.3885	-0.44	0.64 (0.47-0.87)	0.0 140	0.0415
Urinary 4-tert-octylphenol (ng/mL)	-0.34	0.71 (0.45-1.12)	0.1 607	0.5259	-0.33	0.72 (0.46-1.12)	0.1 682	0.3007	-0.42	0.66 (0.43-1.01)	0.0 846	0.4203	-0.82	0.44 (0.23-0.87)	0.0 331	0.0794
Dimethyldithiophosphate (µg/L)	-0.31	0.74 (0.51-1.06)	0.1 173	0.4800	-0.35	0.71 (0.49-1.01)	0.0 783	0.1782	-0.21	0.81 (0.59-1.12)	0.2 319	0.7382	-0.67	0.51 (0.29-0.91)	0.0 388	0.0887

Variables were first tested in an unadjusted / non-covariate adjusted analysis, and then again adjusting for age, ethnicity, and diabetes duration. To provide a broad overview, any variable passing nominal (i.e. prior to FDR-correction)  $p \leq 0.05$  from either the non-covariate adjusted or any of the covariate-adjusted analyses are listed.

**Table 4. RandomForest™-selected variables (features).**

Marker	Group	Mean decrease accuracy	Mean decrease Gini
Glycohemoglobin (%)	Diabetes status	31.93719324	21.75225602
C-reactive protein (mg/dL)	Immune markers	11.51047187	10.1975135
Potassium (mmol/L)	Renal function	11.44142126	8.198116194
Albumin, urine (mg/L)	Renal function	8.220187056	10.21346621
Monocyte number (1000 cells/uL)	Immune markers	7.663589246	3.957711466
Osmolality (mmol/Kg)	Renal function	7.510989556	5.33445026
White blood cell count (1000 cells/uL)	Immune markers	7.440157644	4.405168072
Blood urea nitrogen (mmol/L)	Renal function	7.224789174	5.073947519
Segmented neutrophils num (1000 cell/uL)	Immune markers	7.020397225	4.067402271
Fasting Glucose (mmol/L)	Diabetes status	6.563694988	2.826827797
Red cell distribution width (%)	Hematocrit	6.138538515	4.947185792
Urinary nitrate (ng/mL)	Renal function	5.899174386	5.844258103
Glucose, serum (mmol/L)	Diabetes status	5.85339684	4.77357191
2-hydroxyphenanthrene (ng/L)	Toxins / Metals	4.028082549	2.307374348
MCHC (g/dL)	Hematocrit	3.936015913	3.896804592
Creatinine (µmol/L)	Renal function	3.530916132	3.747500758
Mono-2-ethyl-5-carboxypentyl phthalate	Toxins / Metals	2.996581682	2.212151134
Blood Nitromethane (pg/mL)	Toxins / Metals	2.936160917	3.31634662
Phosphorus (mmol/L)	Toxins / Metals	2.819084205	3.768671211
Total Cholesterol (mmol/L)	Sterols	2.448578413	3.833301666
Enterodiol (ng/mL)	Sterols	2.401651721	2.722762228
Hematocrit (%)	Hematocrit	2.364741874	4.841281382
Mono-n-octyl phthalate (ng/mL)	Toxins / Metals	2.215508752	0.231647322
Mean cell hemoglobin (pg)	Hematocrit	2.183482824	4.292282119
Gamma glutamyl transferase (U/L)	Liver Function	1.989695745	4.015070221
Systolic blood pressure	Blood pressure	1.892815461	8.85430752
Blood o-Xylene (ng/mL)	Toxins / Metals	1.670964308	3.708248225
Lactate dehydrogenase LDH (U/L)	Liver Function	1.593869272	4.231659273
9-hydroxyfluorene (ng/L)	Toxins / Metals	1.430814333	2.206778568
Cholesterol (mmol/L)	Sterols	1.201366974	4.025207689
3-hydroxyphenanthrene (ng/L)	Toxins / Metals	1.00569817	1.63630888

Notes: The model was initially trained on all laboratory variables in an unsupervised fashion, with Kappa-based model tuning to select the optimum values for ‘mtry’ (the ideal number of variables to randomly sample) and ‘ntrees’ (the ideal number of trees). Only variables contributing >1% mean decrease in accuracy from the initial model were retained, followed by recursive steps to remove low-informative variables. Variables are manually assigned to groups based on similar organ function or other characteristic. The Gini importance measure relates to the ‘splitting’ criterion that is employed in classification trees, and it is known to be less biased for continuous variables (26), which naturally have more splitting points compared to categorical variables. ‘Group’ is manually curated.

**Table 5. Final clinical risk models.**

Model	Wald test p-value	McFadden R <sup>2</sup>	Nagelkerke R <sup>2</sup>	AUC (95% CI)
Diabetes Status	≤ 0.0001	0.102	0.142	0.72 (0.677-0.763)
Hematocrit	0.0023	0.007	0.009	0.57 (0.539-0.601)
Blood Pressure (BP)	≤ 0.0001	0.017	0.024	0.586 (0.554-0.618)
Immune Markers	0.42	0.002	0.002	0.527 (0.495-0.559)
Renal function tests (renal)	≤ 0.0001	0.039	0.054	0.636 (0.604-0.669)
Sterols (include cholesterol)	0.055	0.012	0.017	0.582 (0.522-0.642)

Toxins / Metals	0.94	0.029	0.041	0.589 (0.478-0.7)
Liver function tests	0.07	0.002	0.003	0.525 (0.494-0.557)
Diabetes control + BP + Renal function	$\leq 0.0001$	0.13	0.18	0.73 (0.686-0.774)
Diabetes control + BP + renal function + Hematocrit	$\leq 0.0001$	0.135	0.184	0.737 (0.694-0.78)
Diabetes control + BP + renal function + Hematocrit + Sterols + Liver function	$\leq 0.0001$	0.238	0.315	0.823 (0.765-0.881)
All groups $\pm$	0.00013	0.272	0.355	0.84 (0.783-0.897)

Features from RandomForest™ were grouped logically based on similar function or clinical use (**Table 4**) and then tested independently in a univariate or multivariate regression model against DR outcome.

$\pm$  The only toxins / metal included was Phosphorus (mmol/L) – others filtered out due to high missingness (>50%), resulting in difficulty fitting model.

## FIGURE LEGENDS

**Figure 1. Cohort selection process for NHANES 2007-8.**

**Figure 2.**

**Penalized regression-selected variables.**

Variables were selected from a 100x cross-validated model with  $\alpha=0.5$ . Final variable selection was based on coefficients not shrunk to 0. Model accuracy was determined to be 78.4% accuracy and AUC 0.71 (95% CI: 0.65-0.77).

**Figure 3. Final clinical risk models.**

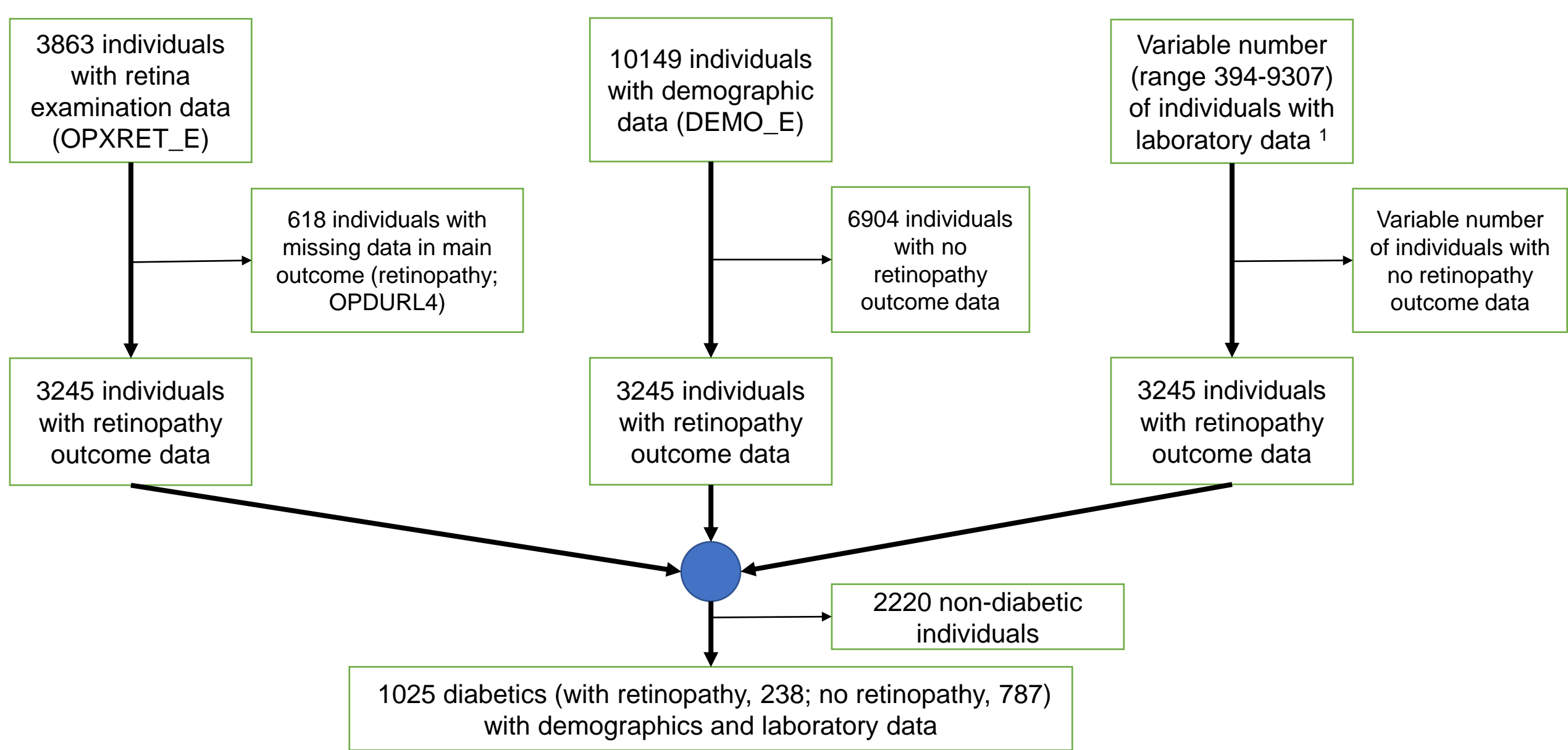
Features from RandomForest™ were grouped logically based on similar function or clinical use (**Table 4**) and then tested independently in a univariate or multivariate regression model against DR outcome. A final risk model including markers of hypertension, hypercholesterolemia, renal and liver function tests, and hematocrit achieved AUC 0.84 (0.78-0.9).

## REFERENCES

1. Feman SS. The natural history of the first clinically visible features of diabetic retinopathy. *Trans Am Ophthalmol Soc.* 1994;92:745-73.
2. Chao JR, Lai M-Y, Azen SP, Klein R, Varma R, Group tLALES. Retinopathy in Persons without Diabetes: The Los Angeles Latino Eye Study. *Investigative Ophthalmology & Visual Science.* 2007;48(9):4019-25.
3. Venkatramani J, Mitchell P. Ocular and systemic causes of retinopathy in patients without diabetes mellitus. *BMJ (Clinical research ed).* 2004;328(7440):625-9.
4. Singh R, Ramasamy K, Abraham C, Gupta V, Gupta A. Diabetic retinopathy: an update. *Indian J Ophthalmol.* 2008;56(3):178-88.
5. Wang W, Lo ACY. Diabetic Retinopathy: Pathophysiology and Treatments. *Int J Mol Sci.* 2018;19(6):1816.

6. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35(3):556-64.
7. Heintz E, Wiréhn AB, Peebo BB, Rosenqvist U, Levin LA. Prevalence and healthcare costs of diabetic retinopathy: a population-based register study in Sweden. *Diabetologia*. 2010;53(10):2147-54.
8. Group UPDSU. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. *Lancet*. 1998;352(9131):837-53.
9. Chatziralli IP. The Role of Dyslipidemia Control in the Progression of Diabetic Retinopathy in Patients with Type 2 Diabetes Mellitus. *Diabetes Ther*. 2017;8(2):209-12.
10. Liu L, Quang ND, Banu R, Kumar H, Tham Y-C, Cheng C-Y, et al. Hypertension, blood pressure control and diabetic retinopathy in a large population-based study. *PLoS One*. 2020;15(3):e0229665-e.
11. Chen X, Meng Y, Li J, She H, Zhao L, Zhang J, et al. Serum uric acid concentration is associated with hypertensive retinopathy in hypertensive chinese adults. *BMC Ophthalmol*. 2017;17(1):83-.
12. Luo B-A, Gao F, Qin L-L. The Association between Vitamin D Deficiency and Diabetic Retinopathy in Type 2 Diabetes: A Meta-Analysis of Observational Studies. *Nutrients*. 2017;9(3):307.
13. Kong X, Wang J, Gao G, Tan M, Ding B, Li H, et al. Association between Free Thyroxine Levels and Diabetic Retinopathy in Euthyroid Patients with Type 2 Diabetes Mellitus. *Endocr Res*. 2020;45(2):111-8.
14. Merin S, Freund M. Retinopathy in severe anemia. *Am J Ophthalmol*. 1968;66(6):1102-6.
15. Khan AA, Rahmani AH, Aldebasi YH. Diabetic Retinopathy: Recent Updates on Different Biomarkers and Some Therapeutic Agents. *Curr Diabetes Rev*. 2018;14(6):523-33.
16. Moriya T, Tanaka S, Kawasaki R, Ohashi Y, Akanuma Y, Yamada N, et al. Diabetic retinopathy and microalbuminuria can predict macroalbuminuria and renal function decline in Japanese type 2 diabetic patients: Japan Diabetes Complications Study. *Diabetes Care*. 2013;36(9):2803-9.
17. Chen YH, Chen HS, Tarng DC. More impact of microalbuminuria on retinopathy than moderately reduced GFR among type 2 diabetic patients. *Diabetes Care*. 2012;35(4):803-8.
18. Lim LS, Tai ES, Mitchell P, Wang JJ, Tay WT, Lamoureux E, et al. C-reactive Protein, Body Mass Index, and Diabetic Retinopathy. *Investigative Ophthalmology & Visual Science*. 2010;51(9):4458-63.
19. Raymond NT, Varadhan L, Reynold DR, Bush K, Sankaranarayanan S, Bellary S, et al. Higher prevalence of retinopathy in diabetic patients of South Asian ethnicity compared with white Europeans in the community: a cross-sectional study. *Diabetes care*. 2009;32(3):410-5.
20. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Current diabetes reports*. 2013;13(6):814-23.
21. Wong TY, Klein R, Islam FMA, Cotch MF, Folsom AR, Klein BEK, et al. Diabetic retinopathy in a multi-ethnic cohort in the United States. *American journal of ophthalmology*. 2006;141(3):446-55.

22. McGinnis DP, Brownstein JS, Patel CJ. Environment-Wide Association Study of Blood Pressure in the National Health and Nutrition Examination Survey (1999–2012). *Scientific Reports*. 2016;6(1):30373.
23. Zhuang X, Guo Y, Ni A, Yang D, Liao L, Zhang S, et al. Toward a panoramic perspective of the association between environmental factors and cardiovascular disease: An environment-wide association study from National Health and Nutrition Examination Survey 1999-2014. *Environ Int*. 2018;118:146-53.
24. Hall MA, Dudek SM, Goodloe R, Crawford DC, Pendergrass SA, Peissig P, et al. Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank. *Pac Symp Biocomput*. 2014:200-11.
25. Patel CJ, Rehkopf DH, Leppert JT, Bortz WM, Cullen MR, Chertow GM, et al. Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States national health and nutrition examination survey. *Int J Epidemiol*. 2013;42(6):1795-810.
26. Nembrini S, König IR, Wright MN. The revival of the Gini importance? *Bioinformatics*. 2018;34(21):3711-8.
27. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.
28. Stratton IM, Kohner EM, Aldington SJ, Turner RC, Holman RR, Manley SE, et al. UKPDS 50: risk factors for incidence and progression of retinopathy in Type II diabetes over 6 years from diagnosis. *Diabetologia*. 2001;44(2):156-63.
29. Klein R, Klein BEK. Blood pressure control and diabetic retinopathy. *Br J Ophthalmol*. 2002;86(4):365-7.
30. Zheng Y, Lamoureaux EL, Lavanya R, Wu R, Ikram MK, Wang JJ, et al. Prevalence and risk factors of diabetic retinopathy in migrant Indians in an urbanized society in Asia: the Singapore Indian eye study. *Ophthalmology*. 2012;119(10):2119-24.



Footnotes:

<sup>1</sup> Laboratory datasets (2007-8): ALB\_CR\_E; APOB\_E; BIOPRO\_E; BPX\_E; CARB\_E; CBC\_E; COTNAL\_E; CRP\_E; DEET\_E; ENX\_E; EPH\_E; FASTQX\_E; FERTIN\_E; FOLATE\_E; FOLFMS\_E; GHB\_E; GLU\_E; HDL\_E; HEPA\_E; HEPBD\_E; HEPB\_S\_E; HEPC\_E; HIV\_E; HPVSER\_E; HPVSWR\_E; HSV\_E; IHG\_E; OGGT\_E; OPD\_E; PAH\_E; PBCD\_E; PERNT\_E; PFC\_E; PHTHTE\_E; PHYTO\_E; POOLTF\_E; PP\_E; PSA\_E; SSHCV\_E; SSUSG\_E; TCHOL\_E; TFR\_E; THYROD\_E; TRIGLY\_E; UAM\_E; UAS\_E; UCPREG\_E; UHG\_E; UHM\_E; UIO\_E; UPHOPM\_E; UPP\_E; VID\_E; VIT\_B6\_E; VOC\_E; VOCMWB\_E; VOCWB\_E.

Model accuracy: 78.51%

OPDURL4 ■ 0 ■ 1

