

1 **Chances and challenges of machine learning based disease**  
2 **classification in genetic association studies illustrated on age-related**  
3 **macular degeneration**

4

5 Felix Günther,<sup>1,2</sup> Caroline Brandl,<sup>1,3</sup> Thomas W. Winkler,<sup>1</sup> Veronika Wanner,<sup>1</sup> Klaus Stark,<sup>1</sup>  
6 Helmut Küchenhoff,<sup>2†\*</sup> Iris M. Heid<sup>1†\*</sup>

7

8 <sup>1</sup>Department of Genetic Epidemiology, University of Regensburg, Regensburg, Bavaria, 93053,  
9 Germany

10 <sup>2</sup>Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of  
11 Munich, Munich, Bavaria, 80539, Germany

12 <sup>3</sup>Department of Ophthalmology, University Hospital Regensburg, Regensburg, Bavaria, 93053,  
13 Germany

14 † These authors jointly supervised this work

15 \*Correspondence to: [iris.heid@klinik.uni-regensburg.de](mailto:iris.heid@klinik.uni-regensburg.de) (I.M.H.), and [helmut.kuechenhoff@stat.uni-](mailto:helmut.kuechenhoff@stat.uni-muenchen.de)  
16 [muenchen.de](mailto:helmut.kuechenhoff@stat.uni-muenchen.de) (H.K.)

17

18 **Abstract**

19 **Imaging technology and machine learning algorithms for disease classification set the**  
20 **stage for high-throughput phenotyping and promising new avenues for genome-wide**  
21 **association studies (GWAS). Despite emerging algorithms, there has been no successful**  
22 **application in GWAS so far. We established machine learning based disease classification**  
23 **in genetic association analysis as a misclassification problem. To evaluate chances and**  
24 **challenges, we performed a GWAS based on automated classification of age-related**  
25 **macular degeneration (AMD) in UK Biobank (images from 135,500 eyes; 68,400 persons).**  
26 **We quantified misclassification of automatically derived AMD in internal validation data**  
27 **(images from 4,001 eyes; 2,013 persons) and developed a maximum likelihood approach**  
28 **(MLA) to account for it when estimating genetic association. We demonstrate that our MLA**  
29 **guards against bias and artefacts in simulation studies. By combining a GWAS on**  
30 **automatically derived AMD classification and our MLA in UK Biobank data, we were able**  
31 **to dissect true association (*ARMS2/HTRA1*, *CFH*) from artefacts (near *HERC2*) and to**  
32 **identify eye color as relevant source of misclassification. On this example of AMD, we are**

33 **able to provide a proof-of-concept that a GWAS using machine learning derived disease**  
34 **classification yields relevant results and that misclassification needs to be considered in**  
35 **the analysis. These findings generalize to other phenotypes and also emphasize the utility**  
36 **of genetic data for understanding misclassification structure of machine learning**  
37 **algorithms.**

38

## 39 **INTRODUCTION**

40 Imaging technology allows for non-invasive access to detailed disease features in large studies  
41 and genome-wide association studies (GWAS) on such disease phenotypes can be expected to  
42 accelerate knowledge gain. However, image-based disease classification can be challenging for  
43 large sample sizes due to time-intensive, tiresome manual inspection. This limitation can be  
44 overcome by automated disease classification via machine learning and particularly deep  
45 learning algorithms. Such emerging approaches<sup>1</sup> can classify diseases effortlessly also for huge  
46 sample sizes as needed for GWAS or other -omics approaches.

47 Deep learning algorithms require enormous input data with available gold standard  
48 classification, in order to “learn” classification reliably. Once trained and tested, the algorithms  
49 can be applied to external image data, but they cannot critically reflect unusual findings or  
50 incorporate unforeseen aspects, for which the human eye and brain has un-met capability. At the  
51 current time, the input data to train algorithms is limited and often specific to a certain setting  
52 (e.g. patients from a clinic). Some characteristics that appear useful for disease classification in  
53 one setting might be misinterpreted in another, which can hamper transferability of trained  
54 models; a topic discussed as dataset shift or domain shift<sup>2-4</sup>. Most predictions of deep learning  
55 algorithms for image-based disease classification will be error-prone and the structure of  
56 misclassification will generally be unknown. When using automated disease classification as  
57 outcome for association analyses and GWAS, the underlying response misclassification is  
58 usually unaccounted for, giving rise to biased effect estimates and potentially false-positive  
59 associations<sup>5-7</sup>. Extent and structure of the misclassification process can be assessed by *internal*

60 *validation data*, i.e. a subset of participants with both automated and gold standard classification,  
61 which can also be utilized to account for response misclassification in statistical models<sup>7,8</sup>.

62 At present, it is unclear whether machine learning based disease classification is of any  
63 utility for association analyses, particularly for detecting disease signals in GWAS. We thus set  
64 out to evaluate machine learning derived disease classification in GWAS on the example of age-  
65 related macular degeneration (AMD) and we developed a statistical approach accounting for the  
66 implied response misclassification. AMD is an ideal role model, as a common disease  
67 ascertained via imaging of the central retina<sup>9</sup> and with particularly strong known genetic effects<sup>10</sup>.  
68 The manual grading of images for AMD requires a substantial effort by trained staff and is  
69 currently an obstacle for homogeneous disease classification within and across large studies.  
70 For example, in UK Biobank<sup>11</sup>, >135,000 color fundus images are available for >68,000 study  
71 participants, but there is no manually classified AMD available so far. Several machine learning  
72 algorithms have been emerging to classify AMD: some show promising performance, but still  
73 yield misclassified predictions, have acknowledged issues due to domain shift or insufficient  
74 sample size for training, or they lack validation in external studies<sup>12–15</sup>. So far, there is no GWAS  
75 on fundus image ascertained AMD available in UK Biobank, manually classified or machine  
76 learning based.

77

## 78 **MATERIALS AND METHODS**

### 79 **Machine learning based disease classification in GWAS as misclassification problem**

80 We consider a binary disease  $Y$ , for which each individual has a true status of disease (disease  
81 yes/no). A *gold standard* classification often involves manual grading of medical images via  
82 trained medical staff, which is considered here to correspond to the true disease classification.  
83 When applying a trained machine learning algorithm on medical images, we yield an automated  
84 disease classification  $Y^*$  for each individual. For an individual  $i$  with true disease status  $Y_i = y_i$ ,  
85 the classification  $Y_i^* = y_i^*$  can either be correct or error-prone ( $y_i^* = y_i$ , or  $y_i^* \neq y_i$ ). If a gold  
86 standard classification is available (for at least a subset of study participants, internal validation  
87 data), the performance of the algorithm can be quantified by cross-tabulation of the observed

88 error-prone  $y^*$  and the gold-standard classification  $y$  across all participants in the validation sub-  
89 study (confusion matrix); the (mis-)classification process can be characterized by classification  
90 probabilities  $P(Y^* = k|Y = l)$ , for  $l, k \in \{0,1\}$ . For  $l = k = 1$  and  $l = k = 0$ , these probabilities  
91 correspond to the sensitivity and specificity of the algorithm, respectively.

92 In the following, we focus on *bilateral diseases* due to our motivating example of an eye  
93 disease (AMD): for each individual  $i$ , two entity-specific binary disease variables  $Z_{1i}, Z_{2i} \in \{0,1\}$   
94 (here: AMD per eye) are used to define the binary person-specific disease status as the “worse-  
95 entity disease status”  $Y_i := \max(Z_{1i}, Z_{2i})$ , corresponding to “AMD in at least one eye” versus “AMD  
96 in none of the two eyes” in our example. The error-prone machine learning based classification  
97 of entity-specific disease  $Z_{1i}^*, Z_{2i}^*$ , will propagate to an error-prone person-specific disease status,  
98  $Y_i^* = \max(Z_{1i}^*, Z_{2i}^*)$ , when compared to the manually graded; “true”  $Y_i$ .

99 We were interested in evaluating the potential and consequences of such automatically  
100 classified disease in GWAS. The standard approach in GWAS is logistic regression for modelling  
101 the association of a genetic variant (observed as genotypes  $\in \{0,1,2\}$  or imputed allelic dosages  
102  $\in [0,2]$ ) with a binary disease status, usually adjusted for other covariates like age, sex, and  
103 genetic principal components; Wald-tests are used to test for genetic association, accounting for  
104 multiple testing by judging at a Bonferroni-corrected significance level of  $p < 5 \times 10^{-8}$ . When the  
105 association of the genetic variant with the true disease status  $Y$  (here: manually classified  
106 persons-specific AMD) follows a logistic regression model, the usage of the error-prone disease  
107 status  $Y^*$  (here: automatically derived person-specific AMD) in the logistic regression will lead to  
108 a mis-specified model (*naïve association analysis*) with known consequences of decreased  
109 power, biased genetic association estimates, and potentially false-positive associations<sup>5-7</sup>.

110

### 111 **MLA to adjust for response misclassification in bilateral disease**

112 While there are methods available to account for response misclassification for classic diseases  
113 in standard logistic regression<sup>5-7</sup>, there is currently no methodology readily available for bilateral  
114 disease. As described previously<sup>16</sup>, the conceptual challenge here is to account for two types of  
115 misclassification: (i) the entity-specific misclassification that propagates to an error-prone person-

116 specific disease status, where the person-specific disease status is used in the association  
117 analysis, and (ii) a person-specific misclassification from a missing disease status in one of the  
118 two entities. We thus developed an MLA to account for the fact that we are using an error-prone  
119 response  $Y_i^* := \max(Z_{1i}^*, Z_{2i}^*)$ ,  $Z_{1i}^*, Z_{2i}^* \in \{0,1\}$ , in the association analysis, while the true disease  
120  $Y_i := \max(Z_{1i}, Z_{2i})$ ,  $Z_{1i}, Z_{2i} \in \{0,1\}$ , is assumed to follow a logistic regression model.

121 Details are provided in **Appendix A**. The general idea of the MLA is to factorize the  
122 likelihood of the observed, error-prone response data into two parts, the model for the association  
123 between risk factor and true (but in general unobserved) response (*true association model*) and  
124 a model for the misclassification process (*misclassification model*). We adapted this well-  
125 established methodology for analyzing misclassified binary response data<sup>7,8</sup> to the scenario of  
126 bilateral disease with a “worse-entity” disease definition (i.e. the person-specific disease status  
127 is defined as the status of the worse entity). Under the assumption of independent  
128 misclassification for the observed disease in the two entities  $\mathbf{z}_{1i}^*, \mathbf{z}_{2i}^*$  of an individual  $i$ , we derive

$$129 \quad \mathbf{P}(\mathbf{z}_{1i}^*, \mathbf{z}_{2i}^* | \mathbf{x}_i) = \sum_{\mathbf{z}_{1i}, \mathbf{z}_{2i} \in \{0,1\}} \underbrace{\mathbf{P}(\mathbf{z}_{1i}^* | \mathbf{z}_{1i}, \mathbf{x}_i) \times \mathbf{P}(\mathbf{z}_{2i}^* | \mathbf{z}_{2i}, \mathbf{x}_i)}_{\text{misclassification model}} \times \underbrace{\mathbf{P}(\mathbf{z}_{1i}, \mathbf{z}_{2i}, | \mathbf{x}_i)}_{\text{true association model}}.$$

130 The *misclassification model* is characterized by the sensitivity and specificity of the entity-  
131 specific classification process; the *true association model* is the assumed logistic regression  
132 model for the person-specific disease status. When internal validation data is available, the  
133 parameters of both models can be estimated jointly by optimizing a likelihood with different  
134 contributions of participants with only the error-prone response and participants in the validation  
135 data with true and error-prone response available.

136 Our developed approach allows us to adjust for both the entity-specific misclassification  
137 from an automated classification and the misclassification of the person-specific status when one  
138 entity is ungradable. Altogether, we model four parameters in the MLA: (i) the conditional  
139 probability of worse-entity disease given the covariate of interest, (ii) the probability of disease in  
140 both entities conditional on the disease in at least one entity (to adjust for missing information of  
141 one of two entities), as well as (iii) the sensitivity and (iv) the specificity of the entity-specific  
142 misclassification process. For each parameter, the conditional probabilities are modeled using

143 the logistic function (as in standard logistic regression) allowing for a dependency on a  
144 parameter-specific set of person-specific covariates. An open source R implementation is  
145 available (**Web Resources**).

146

#### 147 **Simulation study to investigate the performance of the MLA**

148 We repeatedly simulated association data for a standard normal covariate  $X$  and a (true and  
149 error-prone) binary outcome of a *bilateral* disease. To do this, we (1) sampled the true, person-  
150 specific disease status associated with  $X$ , (2) derived the true entity-specific disease status  
151 (e.g. manual eye-specific AMD classification) given assumptions, (3) sampled the entity-specific  
152 error-prone disease status (e.g. automated AMD classification), and (4) derived an error-prone,  
153 person-specific disease status. Afterwards, we removed the true disease status for most  
154 individuals, yielding only a subset with both true and error-prone disease status available  
155 (validation data). In different simulation scenarios, we varied sensitivity and specificity of the  
156 entity-specific classification. Classification probabilities were either constant for all individuals  
157 (non-differential misclassification) or varying with  $X$  (differential misclassification). We also varied  
158 the fraction of individuals with missing classification in one of two entities. Data was sampled with  
159 or without an effect of  $X$  on the true person-specific response  $Y$  ( $\beta_Y \in \{0,1\}$ , log OR) and on the  
160 probability  $\delta$  of having disease in both entities given disease in at least one entity ( $\beta_\delta \in \{0,1\}$ , log  
161 OR). We estimated the covariate effect using the naive analysis (logistic regression, which  
162 ignores misclassification) and the developed MLA1 and MLA2 accounting for response  
163 misclassification without (MLA1) and with allowing (MLA2) for differential misclassification,  
164 respectively. To compare the performance of the naïve analysis and the derived MLA, we  
165 investigated the distribution of effect estimates  $\hat{\beta}_Y$  across simulation runs, computed the mean  
166 squared error of estimates relative to true effects, frequencies of rejected tests for no association,  
167 and coverage frequencies of 95%-confidence intervals. A detailed description of the simulation  
168 study, data sampling, and estimated models is given in **APPENDIX B**.

169

#### 170 **UK Biobank study information and data**

171 UK Biobank recruited ~500,000 individuals aged 40-69 years from across the United Kingdom.  
172 Genetic data is available from the Affymetrix UK Biobank Axiom Array imputed to the Haplotype  
173 Reference Consortium<sup>17</sup> and the UK10K haplotype resource<sup>18</sup> (details described elsewhere<sup>11</sup>).  
174 The UK Biobank baseline data contains 135,500 fundus images of 68,400 individuals. The  
175 images are taken with the Topcon 3D OCT-1000 Mark II system with a field angle of 45° without  
176 application of mydriasis<sup>19</sup>. The images can be utilized for automated or manual AMD  
177 classification, however, there is no image-based AMD classification publicly available so far.

178

### 179 **AMD classification in UK Biobank derived from a machine learning algorithm and** 180 **manually**

181 We performed an automated AMD classification for 68,400 individuals with available fundus  
182 images in UK Biobank with additional manual classification in a subset of 2,013 participants as  
183 described in the following.

184 In epidemiological studies, AMD is usually classified per eye via manual grading of color  
185 fundus images by trained graders using established classification systems. One such system is  
186 the Age-Related Eye Disease Study (AREDS) 9-step Severity Scale<sup>20</sup>, which defines early AMD  
187 combining a 6-step drusen area scale with a 5-step pigmentary abnormality scale and is therefore  
188 particularly detailed and time-consuming when applied manually. Another more recent system is  
189 the Three Continent AMD Consortium Severity Scale (3CC)<sup>9</sup>, which defines early AMD based on  
190 drusen size, drusen area and presence of pigmentary abnormalities and is thus more practical  
191 to apply manually. While the definition of “advanced AMD” is fairly robust across systems, each  
192 system defines “early” or “intermediate” AMD differently, but provides a clear assignment strategy  
193 to “no”, “early/intermediate” or “advanced AMD” (or “no” and “any AMD”).

194 To obtain an eye-specific AMD status for the 135,500 images of the UK Biobank ( $\leq 1$  image  
195 per eye; 67,100 individuals with images for both eyes, 1,300 with image for only one eye), we  
196 applied a published convolutional neural network ensemble<sup>14</sup> to the fundus images following  
197 recommendations of the authors (**Web Resources**). The ensemble was trained to classify each  
198 image into the AREDS 9-step severity scale or three additional categories for advanced AMD

199 (GA, NV, mixed GA+NV, “AREDS 9+3 steps”) or “ungradable”. From this, we derived the person-  
200 specific automated AMD status as the AMD status of the worse eye (i.e. the higher score of the  
201 ARED9+3) or as the status of the only eye, if applicable. We collapsed AREDS AMD severity  
202 steps 2-9 or any of the 3 advanced AMD categories to “any AMD”.

203 To generate internal validation data, we selected a subset of UK Biobank individuals for  
204 additional manual grading. When randomly sampling participants, one would expect to catch only  
205 a few AMD individuals; we thus enriched the validation sample with persons likely to be affected  
206 by AMD or likely to be unaffected: (i) persons with high genetic risk score for AMD based on the  
207 known 52 variants for advanced AMD<sup>10</sup> (> 99<sup>th</sup> percentile, n=829), (ii) persons with low genetic  
208 risk score (<1<sup>st</sup> Percentile, n=828), and (iii) persons with self-reported AMD not already selected  
209 (n=356). The machine learning based AMD classification was not used to select individuals into  
210 the validation subset. The selected 2,013 individuals were manually classified for AMD according  
211 to the 3CC<sup>9</sup> system by a trained ophthalmologist (five AMD categories, 1 for no AMD, 3 for early,  
212 1 for advanced AMD, and 1 “ungradable”). We collapsed the five AMD categories to “any AMD”,  
213 “no AMD”, or “ungradable” and derived eye-specific as well as person-specific confusion matrices  
214 based on the detailed (AREDS 9+3 and 5-category 3CC) and collapsed classifications. To  
215 conduct the GWAS with automatically derived “any AMD”, we restricted the data with available  
216 automated AMD classification to unrelated individuals of European ancestry with valid GWAS  
217 data (see below), and derived the confusion matrices also for the restricted validation data.

218

### 219 **Genetic association analyses for AMD without and with accounting for misclassification**

220 We performed a GWAS on the automatically derived “any AMD” versus “no AMD” in unrelated  
221 UK Biobank participants (relatedness status > 3<sup>rd</sup> degree) of European ancestry (self-report  
222 “White”, “British”, “Irish” or “Any other white background”) as recommended<sup>21</sup>. For each variant,  
223 we applied a standard logistic regression model (i.e. the naïve analysis ignoring misclassification  
224 in the automatically derived AMD status) under the additive genotype model and applied a Wald-  
225 test as implemented in QUICKTEST<sup>22</sup>. We included age and the first two genetic principal  
226 components as covariates. We excluded variants with low minor allele count (MAC<400,



227 calculated as  $MAC = 2 \times N \times MAF$ , sample size  $N$ , minor allele frequency  $MAF$ ) or with low  
228 imputation quality ( $rsq < 0.4$ ) yielding 11,567,158 analyzed variants. To correct for potential  
229 population stratification, we applied a Genomic Control correction ( $\lambda = 1.01$  based on the  
230 analyzed variants excluding the 34 known AMD loci)<sup>23</sup>.

231 We selected genome-wide significant variants ( $P_{GC} < 5.0 \times 10^{-8}$ ), clumped them into  
232 independent regions ( $\geq 500$ kb between independent regions) and selected the variant with lowest  
233 P-value in each region ("lead variant"). We also selected the 21 of the 34 reported lead variants  
234 from the established advanced AMD loci, for which we had  $\geq 80\%$  power to detect them in a UK  
235 Biobank sample size of 3,544 cases and 44,521 controls with nominally significance - under the  
236 assumption that the reported effect sizes for advanced AMD were the true effect sizes and  
237 ignoring any misclassification in the AMD classification (**APPENDIX C**). Information on linkage  
238 disequilibrium in Europeans was obtained from LDLink<sup>24</sup>. Enrichment of directionally consistent  
239 or enrichment of nominally significant association for the 21 reported lead variants (when  
240 compared to the reported direction literature) was tested based on the Exact Binomial test for  
241  $H_0: Prob = 0.5$  or  $H_0: Prob = 0.05$ , respectively.

242 To evaluate the robustness of the genetic association upon accounting for the  
243 misclassification, we applied the derived MLAs for the selected variants. For this, we modelled  
244 the conditional probability of AMD depending on age, genetic variant and two genetic principal  
245 components (as in the naïve analysis). The MLAs accounted for the misclassification of the eye-  
246 specific automated classification and for the person-specific misclassification from missing AMD  
247 status in one of two eyes. For the misclassification process of the eye-specific automated  
248 classification (quantified by sensitivity and specificity), we allowed for a linear association with  
249 age and modelled two scenarios for the association with the genetic variant: (i) no dependency  
250 (non-differential, MLA1) or (ii) linear dependency (differential misclassification, MLA2). We  
251 compared association estimates of the naïve analysis with MLA1- and MLA2- analysis and  
252 judged significance at Bonferroni-corrected significance levels for a family-wise error rate of 0.05.  
253 To allow for comparisons across different models, we did not apply Genomic control correction  
254 for these comparative analyses. Additionally, we evaluated robustness of findings from the naïve

255 analysis for the selected lead variants upon adjusting for 20 instead of 2 genetic principal  
256 components.

257 To follow-up on the *HERC2* lead variant finding (see Results), we quantified lightness of  
258 fundus images by calculating gray levels for the “RGB” fundus images (weighted sum of R, G  
259 and B values,  $0.30*R+0.59*G+0.11*B$ , as implemented in IrfanView).

260

## 261 **RESULTS**

### 262 **Linking misclassification theory to machine learning disease classification**

263 We here establish the usage of machine learning derived disease classification in genetic  
264 association analyses as a response misclassification problem in logistic regression (**Methods**).  
265 We present a newly developed maximum likelihood approach (MLA) for *bilateral diseases* like  
266 AMD (**Methods**). This includes two versions: (1) assuming *non-differential misclassification*  
267 (MLA1, i.e. no dependency of misclassification probabilities on the covariate of interest, here the  
268 genetic variant) and (2) allowing for *differential misclassification* (MLA2, i.e. dependency on the  
269 covariate of interest). There are existing MLAs for considering response misclassification in  
270 logistic regression using internal validation data<sup>7,8</sup>: these MLAs refer to *classic diseases* where  
271 the misclassification is on the person-specific disease status. Our developed approach provides  
272 a general framework for bilateral diseases with entity-specific misclassification that propagates  
273 to person-specific disease misclassification. Our approach also allows for missing classification  
274 in one of two entities, which is a second source of bias in association analyses for bilateral  
275 diseases as reported previously<sup>16</sup>. We exemplify our approach on machine learning derived AMD  
276 compared to manually graded AMD. Since machine learning algorithms for AMD are trained on  
277 images with human manual AMD grading as benchmark, we assume the manual classification  
278 to be gold standard.

279 We evaluated the performance of our developed MLA1 and MLA2 in a simulation study.  
280 By this, we documented substantial bias and lack of type-I error control when the naïve analysis  
281 was applied, which was comparable to theory for classic (non-bilateral) diseases<sup>5,7</sup>. We also  
282 showed our MLA1 and MLA2 to effectively remove bias and keep type-I error when specified  
283 correctly (**Table 1, APPENDIX D, Supplementary Table 1**).

284

### 285 **AMD in UK Biobank based on automated classification and validation data**

286 We applied a published convolutional neural network ensemble<sup>14</sup> to automatically derive eye-  
287 and person-specific AMD classifications for 68,400 UK Biobank participants with fundus images  
288 at baseline (135,000 eyes) (**Supplemental Table 2a**). From this, we derived eye-specific “any

289 AMD” status (i.e. any early AMD stage or advanced AMD versus AMD-free) and person-specific  
290 “any AMD” status based on the worse eye (**Methods**). Among the 68,400 participants, 10,128  
291 were ungradable for AMD in both eyes (i.e. missing person-specific AMD status, 14.8%), 4,870  
292 were classified as “any AMD” and 53,402 as AMD-free (**Supplemental Table 2b**). Among the  
293 58,272 gradable participants (of these: 20.2% gradable only in one eye), 8.4% had AMD and  
294 91.6% were AMD-free. This included 48,065 unrelated individuals of European ancestry with  
295 GWAS data (3,544 “any AMD” cases, 44,521 AMD-free controls; 19.8% with only one eye  
296 gradable; **Supplemental Table 2b**).

297 To quantify the performance of automated AMD classification, we manually classified  
298 AMD in a subset as internal validation data (4,001 images,  $\leq 1$  image per eye, 2,013 individuals).  
299 When comparing automated to manual (true) “any AMD” status, we found an eye-specific  
300 sensitivity of 73% and specificity of 90% in the full validation data and a person-specific sensitivity  
301 of 77% and specificity of 91% among the participants in the GWAS (**Table 2a/b**). We found no  
302 structural differences between the full validation data and when restricting to the GWAS data  
303 (1,327 individuals, **Supplemental Table 3a/b**). Both, the manual and automated classification  
304 included the category “ungradable”. Among the 4,001 eyes, 1,101 were manually ungradable, of  
305 which the automatic classification yielded 74% as ungradable as well, but classified 9% as AMD  
306 and 17% as AMD-free, which raises concerns about these classifications. In summary, we found  
307 the automated classification to yield reasonable, but error-prone results.

308

### 309 **GWAS on automated AMD classification in naïve analysis identifies two loci**

310 While we have some idea about the extent of the misclassification from validation data and about  
311 its impact on genetic association estimates from simulations, it is unclear whether the automated  
312 any AMD classification is “good enough” for GWAS. We conducted a GWAS for person-specific  
313 automatically derived “any AMD” in UK Biobank (3,544 “any AMD” cases; 44,521 controls)  
314 applying logistic regression as usual, which is without accounting for misclassification (naïve  
315 analysis). We found 53 variants with genome-wide significance ( $P_{GC} < 5.0 \times 10^{-8}$ ) spread across two  
316 distinct loci (defined as lead variant and proxies  $\pm 500$ kb, **Figure 1a/b; Supplemental Table**

317 **4a**): the known *ARMS2/HTRA1* locus (lead variant here rs370974631,  $P_{GC}=3.1 \times 10^{-20}$ , effect allele  
318 frequency EAF=0.23) and an unknown locus for AMD near *HERC2* (lead variant rs12913832,  
319  $P_{GC}=4.7 \times 10^{-16}$ , EAF=0.23). This *ARMS2/HTRA1* lead variant was highly correlated to the  
320 reported lead variant for advanced AMD, rs3750846, and effect estimates were directionally  
321 consistent ( $r^2=0.93$ ; **Supplemental Table 4b**). The next best known locus is the *CFH* locus,  
322 which showed close to genome-wide significance here (smallest  $P$ -value  $P_{GC}=7.0 \times 10^{-7}$ ,  
323 rs6695321, EAF=0.62): rs6695321 is in linkage disequilibrium with two reported *CFH* variants  
324 (rs61818925, rs570618:  $r^2=0.63$  or  $0.40$ ,  $D'=0.81$  or  $1.00$ , EAF=0.58 or  $0.36$ , respectively;  
325 **Supplemental Table 4b**) suggesting that rs6695321 captures the signals of these two reported  
326 variants.

327 Among the reported lead variants of the 34 advanced AMD loci<sup>10</sup>, we had  $\geq 80\%$  power to  
328 detect 21 of these with nominal significance (**Supplemental Table 5**). When comparing effect  
329 sizes of these 21 variants from this analysis on “any AMD” in UK Biobank with reported effect  
330 sizes for advanced AMD, we found 15 with directional consistency ( $P_{Bin}=0.078$ ) and 7 with  
331 directionally consistent nominal significance ( $P_{Bin}=4.9 \times 10^{-5}$ ; **Figure 3a, Supplemental Table 4c**).  
332 The overall smaller effect sizes for automated “any AMD” compared to reported effect sizes for  
333 advanced AMD can be explained by a bias from misclassified automated AMD and by smaller  
334 effect sizes for early AMD merged into the definition of “any AMD”. For the other 13 of the 34  
335 variants, we refrained from interpreting results due to lack of power in this analysis  
336 (**Supplemental Table 4c**). Results were similar when adjusting for 20 instead of 2 genetic  
337 principal components (data not shown). While the yield of only few known AMD signals in this  
338 UK Biobank GWAS may be disappointing, this is not fully unexpected given an effective sample  
339 size<sup>25</sup> of 13,130 and a power estimate of  $\sim 80\%$  (assuming no misclassification and reported effect  
340 sizes) to detect associations with genome-wide significance for only 4 of the 34 established  
341 variants (*CFH*, *ARMS2/HTRA1*, *C3*, *C2/CFB/SKIV2L*, **Supplemental Table 5**).

342 In summary, our GWAS on automated AMD in UK Biobank detected the established  
343 *ARMS2/HTRA1* locus, an unknown locus around *HERC2* with genome-wide significance, and  
344 the established *CFH* locus to some extent.

345

### 346 **Applying the developed MLA to account for misclassification for selected variants**

347 Due to our simulation results and theory<sup>5,7</sup>, we expected our GWAS on automated (error-prone)  
348 AMD to yield biased estimates and, when the misclassification was differential towards the  
349 genetic variant, even potentially false signals. We applied our developed MLAs for 26 selected  
350 variants: (i) the 3 lead variants detected here with (near) genome-wide significance (*CFH*:  
351 rs6695321, *ARMS2/HTRA1*: rs370974631, *HERC2*: rs12913832), (ii) the 3 reported independent  
352 variants in the *CFH* locus with MAF $\geq$ 5% (rs61818925, rs570618, rs10922109; 2 of these  
353 correlated to the here identified *CFH* lead variant), and (iii) the other 20 of the 34 reported lead  
354 variants<sup>10</sup>, for which we had reasonable power in this analysis (including 1 reported  
355 *ARMS2/HTRA1* variant correlated to here identified variant). This yielded a total of ~23  
356 independent variants.

357 Our MLAs estimated simultaneously (1) sensitivity and specificity of the eye-specific  
358 misclassification process and (2) genetic association accounting for the misclassification. With  
359 regard to sensitivity and specificity, we found (i) an overall sensitivity of 64.5% (95%-CI: 60.1%,  
360 68.7%) and a specificity of 98.6% (98.4%, 98.8%), i.e. a false-negative “any AMD” proportion of  
361 35.5% and a false-positive of 1.4%, (ii) no dependency of the sensitivity on any selected variant  
362 ( $P > 0.05 / (23 \times 2) = 1.09 \times 10^{-3}$ ) and no dependency of the specificity, except for two variants: *HERC2*  
363 lead variant, rs12913832, and the reported *CFH* lead variant rs10922109 ( $OR_{spec} = 0.64$ ,  
364  $P_{spec} = 7.38 \times 10^{-9}$  and  $OR_{spec} = 1.36$ ,  $P_{spec} = 2.29 \times 10^{-4}$ , respectively; **Supplemental Table 6**,  
365 **Appendix E**). Therefore, we found a misclassification that was associated with some genetic  
366 variants (differential), which could induce bias into all directions and severe lack of type-I error  
367 control.

368 When comparing genetic association estimates from our MLA1 and MLA2 with the naïve  
369 analysis for our three detected lead variants, we found interesting patterns (**Figure 2**,  
370 **Supplemental Table 7a**). (i) For *CFH* and *ARMS2/HTRA1*, we found consistent effect estimates  
371 across the three analyses, with larger confidence intervals when using the more complex models  
372 MLA1 or MLA2. (ii) For *HERC2*, MLA1 yielded comparable results to the naïve analysis, but when

373 accounting for differential misclassification (MLA2), the effect vanished (MLA2: OR=1.03,  
374 P=0.76; MLA1: OR=1.34, P=1.11x10<sup>-12</sup>; naïve: OR=1.26, P=4.16x10<sup>-16</sup>). When applying MLA1  
375 and MLA2 to the three reported *CFH* locus variants and the further 20 of the 34 reported lead  
376 variants, we found the following (**Supplemental Table 7b/c**): (i) effect estimates for all three *CFH*  
377 variants increased when applying MLA2 compared to the naïve analysis. This was particularly  
378 interesting for the reported *CFH* lead variant rs10922109, where we now observed a nominally  
379 significant association into the reported direction (MLA2: OR=1.15, P=0.047; naïve: OR=1.00,  
380 P=0.98; **Supplemental Table 7c**). This is in line with the observed dependency of the specificity  
381 on this *CFH* variant. (ii) For the other 20 reported lead variants, many variants showed increased  
382 effect estimates by MLA2 compared to the naïve analysis (effect estimates mostly more  
383 comparable to reported effect sizes<sup>10</sup>; **Figure 3c**). Altogether, MLA results confirmed the *CFH*  
384 and *ARMS2/HTRA1* loci and unmasked the *HERC2* finding as false positive.

385

### 386 **Misclassification depended to eye and fundus image color**

387 Interestingly, our *HERC2* lead variant, rs129138329, is precisely the variant for which the G allele  
388 was considered causal for blue eyes<sup>26</sup>. We were able to support this in our AugUR<sup>27,28</sup> study  
389 (n=1026; reported “light eye color” for 14%, 36%, or 97% of participants with A/A, G/A, or G/G,  
390 respectively). Eye color is discussed as AMD risk factor, but the debate is on blue eyes to  
391 increase risk due to increased susceptibility to UV-radiation<sup>29</sup>, which is in contrast to our  
392 observation of brown eyes to increase AMD risk and a challenge for interpreting this finding. It  
393 was interesting to see the *HERC2* rs129138329 association vanish when accounting for  
394 rs129138329-associated misclassification. This was in line with the observed strong association  
395 of the specificity with this variant (OR<sub>spec</sub>=0.64 per A allele, **Supplemental Table 6a**) resulting in  
396 3.0%, 1.9%, or 1.2% of false-positive AMD classifications among persons with A/A, A/G, or G/G,  
397 respectively. This notion of a larger misclassification among A/A versus G/G individuals was  
398 further supported by the larger fraction of manually ungradable images that were deemed  
399 gradable by the automatic classification among A/A versus G/G (54.5% versus 38.8%,  
400 respectively; **Figure 4**). When visually inspecting fundus images per genotype group, the images

401 for A/A had a darker appearance than those for A/G or G/G (**Figure 4**), which we were able to  
402 quantify by means of average gray level per image of 46.4, 49.0, or 53.6, respectively. Therefore,  
403 the *HERC2* signal appeared to be an artefact due to a larger misclassification for brown eyes  
404 linked to darker fundus images. One may hypothesize that the darker eye color had reduced light  
405 exposure during fundus photography, which gave rise to darker images and more misclassified  
406 AMD-free eyes. The notion of a differential misclassification due to eye color was further  
407 supported by the fact that the full *HERC2* signal disappeared by modelling a misclassification  
408 dependency on the causal variant for eye color (rs129138329, **Supplemental Figure 1a/b**), while  
409 some signal remained when modelling a misclassification dependency on the respective *HERC2*  
410 variant in the model (**Supplemental Figure 1c**). In summary, we found the MLA2 not only to  
411 effectively remove the artefact signal of the naïve GWAS, but also to help understand the  
412 dependencies of the misclassification.

413

## 414 **DISCUSSION**

415 GWAS on machine learning derived classification of imaging-based diseases, like AMD, can be  
416 expected to accelerate knowledge gain and drug target development<sup>30</sup>, since it will enable  
417 substantially increased sample sizes and refined, homogeneous phenotyping. To this date, there  
418 was no GWAS reported using a machine learning derived classification for AMD or any other  
419 imaging-based disease – to our knowledge. We here present a GWAS on machine learning  
420 derived AMD in UK Biobank highlighting chances and challenges. By this GWAS on AMD  
421 combined with an evaluation of emerging genetic signals via our newly developed MLA, we were  
422 able to detect known AMD loci and to distinguish true loci from artefacts.

423         Such artefacts, i.e. false positives, can derive from a misclassification that is associated  
424 with a genetic variant. Our data and analyses provide a compelling example for such an artefact:  
425 our MLA revealed the *HERC2* signal as false positive signal and suggested darker eye color and  
426 darker fundus images as a relevant source of misclassification for this machine learning  
427 algorithm. It is perceivable that the misclassification process of other algorithms for AMD and for  
428 other image-based diseases will depend on one or the other characteristic as well, and that such

429 a characteristic is picked up by some genetic variants due to the abundant range of genetically  
430 pinpointed characteristics (see e.g. NHGRI-EBI GWAS Catalog<sup>31</sup>), which can yield artefact  
431 signals when left unaccounted.

432 Our MLA, developed for bilateral diseases, does not only quantify the misclassification  
433 and the dependencies, but also guards against bias and artefacts in association analyses. Similar  
434 approaches are available for classic diseases<sup>7,8</sup>. Thus, this concept can be generalized to other  
435 algorithms and other image-based diseases. Our work here links the theory of misclassification  
436 to machine learning derived disease classification, which can be generalized also to  
437 measurement error and quantitative phenotypes.

438 We recommend a GWAS combined with a post-GWAS evaluation of emerging genetic  
439 effects for non-differential and differential misclassification not only to search for GWAS signals  
440 on image-based, machine-learning derived disease phenotypes. We also recommend such a  
441 GWAS as a quality control for diseases like AMD, where strong genetic signals are known: a  
442 GWAS on AMD ascertained by any classification approach, manual or automatic, should be able  
443 to detect at least the two strong known signals around *ARMS2/HTRA1* and *CFH*. When a GWAS  
444 does not detect these signals, this indicates issues that can be anything from mis-matched bio-  
445 samples, analytical errors, or imperfect disease ascertainment – like from machine learning  
446 algorithms as highlighted here. A GWAS can be a quick guide towards phenotype classification  
447 quality when genomic data is available.

448 Overall, we illustrate chances and challenges of machine learning derived disease  
449 classification in GWAS, and the applicability of our MLA to guard against bias and artefacts.



## 450 **Appendices**

### 451 **Appendix A. MLA to adjust for response misclassification in bilateral diseases.**

452 We developed an MLA to adjust for response misclassification from an error-prone, entity-specific  
453 disease classification in bilateral diseases. Here we illustrate it based on the example of age-  
454 related macular degeneration, where AMD can occur in each eye (eye-specific AMD) and the  
455 person-specific binary outcome is defined as worse-eye outcome, i.e. “AMD in at least one eye”,  
456 and modeled using logistic regression. We assume that we have an error-prone, eye-specific  
457 AMD classification (e.g. from a machine-learning based automated classification) available for  
458 nearly all eyes and true, gold-standard classifications (e.g. manual classification) for a subset of  
459 individuals from validation data.

460 Let  $(Z_{1i}, Z_{2i}) \in \{0,1\}$  be the true, binary disease stages in the two eyes of study participant  $i$ , i.e.  
461  $(Z_{1i} = 1, Z_{2i} = 0)$  means that participant  $i$  suffers from AMD in the left eye and is unaffected from  
462 AMD in the right. When estimating the association of person-specific risk factors with AMD, one  
463 often defines a binary person-specific disease status as worse-entity AMD,  $Y_i := \max(Z_{1i}, Z_{2i})$ ,  
464  $Z_{1i}, Z_{2i} \in \{0,1\}$ , and uses logistic regression to estimate the association of some covariates  $X$   
465 with AMD: the person-specific disease status  $Y_i$  equals 1, if at least one eye of individual  $i$  is  
466 classified as AMD, and  $Y_i$  equals 0, if both eyes are unaffected. As described previously<sup>16</sup>, such  
467 a worse-eye disease status can be misclassified because of two reasons: either, because of  
468 missing disease information in one of two eyes (in this case disease can be overlooked), or  
469 because of error-prone disease status for any of the two eyes. Here we assume that we observed  
470 an error-prone, eye-specific disease status  $(Z_{1i}^*, Z_{2i}^*)$  for each of the two eyes of a “main study”  
471 participant  $i$  and additionally the true disease status in each of the two eyes  $(Z_{1j}, Z_{2j})$  for a subset  
472 of study participants  $j$  from the “validation study”. For all participants from the main study (error-  
473 prone classifications only) or the validation subset (error-prone and true classification), there is  
474 the additional issue that the disease information can be missing in one of two eyes, because of  
475 missing or ungradable fundus images. Since the automated (error-prone) and manual (gold  
476 standard, “true”) classification may judge differently on whether an image is gradable or  
477 ungradable, any possible subset of  $(Z_{1i}, Z_{2i}, Z_{1i}^*, Z_{2i}^*)$  might be the available information for a

478 specific study participant. To obtain valid estimates for the association of covariates with the true  
479 AMD status, we set up a likelihood based on the conditional probabilities of the observed error-  
480 prone and/or true eye-specific disease classifications given covariates. The product of these  
481 conditional probabilities over all individuals forms the likelihood, which has to be numerically  
482 optimized with respect to the regression parameters to obtain estimates. The different likelihood  
483 contributions for the individuals depend on the available AMD classifications (true and/or error-  
484 prone for one or both eyes).

485 The general problem of response misclassification when AMD information is missing in one of  
486 two eyes and/or the eye-specific classification suffers from misclassification with known  
487 classification probabilities has already been evaluated in a previous publication<sup>16</sup>. There, we also  
488 derived the corresponding likelihood contributions for the different scenarios of available outcome  
489 data. Here, we add the aspect that validation data is available for some study participants or,  
490 more specifically, a collection of error-free (gold-standard) classified single eyes, and that we  
491 model the eye-specific misclassification process based on information from this validation data.  
492 In the following, we describe the general idea and provide formulas for the respective likelihood  
493 contributions:

494 The assumed logistic regression model for the true worse-eye disease corresponds to the  
495 assumption that  $\max(Z_{1i}, Z_{2i}) = Y_i \sim \text{Bernoulli}(\pi_i)$ , where we model the success probability based  
496 on a linear predictor via  $\pi_i = 1/(1 + \exp(-x_i'\beta)) = \text{Logist}(x_i'\beta)$ ;  $x_i$  is a vector of observed person-  
497 specific covariates and  $\beta$  the vector of corresponding regression coefficients. It follows that  
498  $P(Y_i = 1|x_i) = \pi_i$ . If we focus on single-eye disease classifications, there exist four different  
499 pattern of true disease classifications  $(Z_{1i}, Z_{2i})$ :  $(1,1)$ ,  $(1,0)$ ,  $(0,1)$ ,  $(0,0)$ . From the assumed logistic  
500 regression model for  $Y_i$ , it follows that  $P(Z_{1i} = 0, Z_{2i} = 0|x_i) = 1 - \pi_i$ . Based on the law of total  
501 probability, we can derive  $P(Z_{1i} = 1, Z_{2i} = 1|x_i) = P(Z_{1i} = 1, Z_{2i} = 1|x_i, Y_i = 1) \times P(Y_i = 1|x_i)$  and  
502 we define the person-specific conditional probability of being affected by AMD in both eyes given  
503 AMD in at least one eye as  $\delta_i := P(Z_{1i} = 1, Z_{2i} = 1|x_i, Y_i = 1)$ . When assuming symmetric  
504 probabilities for disease in one but not the other eye for left and right eyes (i.e. same probabilities

505 to be affected in the left but not the right eye and vice versa), the conditional probability mass  
 506 function of the two-entity disease status distribution can be written concisely as

$$\begin{array}{c|cc}
 P(\cdot, \cdot | x_i) & Z_{2i} = 1 & Z_{2i} = 0 \\
 \hline
 Z_{1i} = 1 & \delta_i \pi_i & \frac{1 - \delta_i}{2} \pi_i \\
 Z_{1i} = 0 & \frac{1 - \delta_i}{2} \pi_i & 1 - \pi_i
 \end{array} \quad (1)$$

507 which specifies the *true data model*. If we look at a single eye selected randomly from both eyes,  
 508 we can derive (without loss of generality for  $Z_{1i}$ ):

$$509 \quad P(Z_{1i} = 1 | x_i) = P(Z_{1i} = 1, Z_{2i} = 1 | x_i) + P(Z_{1i} = 1, Z_{2i} = 0 | x_i) = \left( \frac{1}{2} + \frac{1}{2} \delta_i \right) \pi_i \quad (2)$$

510 We now assume that we observed potentially misclassified single eye disease stages ( $Z_{1i}^*, Z_{2i}^*$ )  
 511 for each participant and describe the *misclassification process* based on the sensitivity and  
 512 specificity of the classification,

$$513 \quad P(Z_{1i}^* = 1 | Z_{1i} = 1, x_i) = \pi_{1i} \quad (3)$$

$$514 \quad P(Z_{1i}^* = 0 | Z_{1i} = 0, x_i) = \pi_{0i},$$

515 with  $l = 1, 2$ ;  $\pi_{1i}$  and  $\pi_{0i}$  are the person-specific sensitivity and specificity from the eye-specific  
 516 classification process. We assume that the eye-specific classification process within an individual  
 517 is independent in the two eyes, i.e.:

$$518 \quad P(Z_{1i}^* = z_{1i}^*, Z_{2i}^* = z_{2i}^* | Z_{1i} = z_{1i}, Z_{2i} = z_{2i}, x_i) = P(Z_{1i}^* = z_{1i}^* | Z_{1i} = z_{1i}, x_i) \times P(Z_{2i}^* = z_{2i}^* | Z_{2i} = z_{2i}, x_i).$$

519 Based on the *true data model* and the description of the *misclassification process* via sensitivity  
 520 and specificity, we can now express the conditional probabilities of all combinations of observed  
 521 outcomes, by using Bayes' rule and the law of total probability. If all four AMD classifications  
 522 were observed for an individual (individual with full validation data, true and error-prone disease  
 523 status for each of the two eyes), we can derive the following (omitting a random variable notation  
 524 and only using the small z's for the observed data):

$$\begin{aligned}
 525 \quad P(z_{1i}^*, z_{2i}^*, z_{1i}, z_{2i} | x_i) &= P(z_{1i}^*, z_{2i}^* | z_{1i}, z_{2i}, x_i) \times P(z_{1i}, z_{2i}, | x_i) \\
 526 \quad &= P(z_{1i}^* | z_{1i}, x_i) \times P(z_{2i}^* | z_{2i}, x_i) \times P(z_{1i}, z_{2i}, | x_i).
 \end{aligned}$$

527 Here, we fraction the conditional probability of the observed data into terms of the eye-specific  
 528 classification process (depending on sensitivity or specificity when the observed true outcome  
 529  $z_{ii}$  is 1 or 0, respectively, (3)) and the true data model (1). If only the two eye-specific error-prone  
 530 classifications are observed (individual in the main study, not part of the validation subset), the  
 531 law of total probability can be used and the conditional probability can be expressed as

$$\begin{aligned}
 532 \quad P(z_{1i}^*, z_{2i}^* | x_i) &= \sum_{z_{1i}, z_{2i} \in \{0,1\}} P(z_{1i}^*, z_{2i}^* | z_{1i}, z_{2i}, x_i) \times P(z_{1i}, z_{2i}, | x_i) \\
 533 \quad &= \sum_{z_{1i}, z_{2i} \in \{0,1\}} P(z_{1i}^* | z_{1i}, x_i) \times P(z_{2i}^* | z_{2i}, x_i) \times P(z_{1i}, z_{2i}, | x_i),
 \end{aligned}$$

534 This again yields an expression that depends on the eye-specific classification probabilities (3)  
 535 and the *true data model* (1).

536 If only a classification for one error-prone outcome was observed (e.g.  $Z_{1i}^* = z_{1i}^*$ ), the conditional  
 537 probability is given by

$$538 \quad P(z_{1i}^* | x_i) = P(z_{1i}^* | Z_{1i} = 0, x_i) \times P(Z_{1i} = 0 | x_i) + P(z_{1i}^* | Z_{1i} = 1, x_i) \times P(Z_{1i} = 1 | x_i),$$

539 where the first terms in each summand depends on the specificity and the sensitivity of the eye-  
 540 specific observation process; an expression for the second was already given above (equation  
 541 (2)).

542 When three classifications were observed, e.g.  $(Z_{1i}, Z_{1i}^*, Z_{2i}^*)$  or  $(Z_{1i}, Z_{2i}, Z_{1i}^*)$ , we can derive

$$\begin{aligned}
 544 \quad P(z_{1i}, z_{1i}^*, z_{2i}^* | x_i) &= P(z_{1i}^*, z_{2i}^* | z_{1i}, Z_{2i} = 0, x_i) \times P(z_{1i}, Z_{2i} = 0 | x_i) + P(z_{1i}^*, z_{2i}^* | z_{1i}, Z_{2i} = 1, x_i) \times P(z_{1i}, Z_{2i} = 1 | x_i) \\
 545 \quad &= P(z_{1i}^* | z_{1i}, x_i) \times P(z_{2i}^* | Z_{2i} = 0, x_i) \times P(z_{1i}, Z_{2i} = 0 | x_i) \\
 546 \quad &\quad + P(z_{1i}^* | z_{1i}, x_i) \times P(z_{2i}^* | Z_{2i} = 1, x_i) \times P(z_{1i}, Z_{2i} = 1 | x_i),
 \end{aligned}$$

543 and

$$\begin{aligned}
 547 \quad & \\
 548 \quad P(z_{1i}, z_{2i}, z_{1i}^* | x_i) &= P(z_{1i}^* | z_{1i}, z_{2i}, x_i) \times P(z_{1i}, z_{2i} | x_i) = P(z_{1i}^* | z_{1i}, x_i) \times P(z_{1i}, z_{2i} | x_i).
 \end{aligned}$$

549 All conditional probabilities characterizing the *true data model* and the *misclassification process*,  
 550 i.e. (i) the probability of true worse-eye AMD  $P(Y_i = 1 | x_i) = \pi_i$ , (ii) the probability of AMD in both  
 551 eyes given AMD in at least one eye  $P(Z_{1i} = 1, Z_{2i} = 1 | Y_i = 1, x_i) = \delta_i$ , (iii) the eye-specific  
 552 sensitivity  $P(Z_{1i}^* = 1 | Z_{1i} = 1, x_i) = \pi_{1i}$  and (iv) the eye-specific specificity of the error-prone

553 classification  $P(Z_{ii}^* = 0 | Z_{ii} = 0, x_i) = \pi_{0i}$ , can potentially vary with person-specific characteristics.  
554 We therefore decided to model them based on the logistic function of a linear predictor, where  
555 relevant covariates (characteristics) can be specified for each probability. Combining all these  
556 expressions, we can set up the whole likelihood based on the derived conditional probabilities  
557 and numerically optimize with respect to the regression coefficients of the linear predictors for  $\pi_i$ ,  
558  $\delta_i$ ,  $\pi_{1i}$ , and  $\pi_{0i}$ . Standard errors of the maximum likelihood estimates are derived based on  
559 standard likelihood theory from the square root of the diagonal elements of the inverse of the  
560 observed Fisher information (Hessian) and used for inference. An implementation of the MLA in  
561 the statistical programming language R<sup>32</sup> is available (**Web Resources**)

562

563 **Appendix B. Simulation study to evaluate consequences of ignoring misclassification and**  
564 **the performance of the MLA in correcting it.**

565 We performed a simulation study to evaluate the consequences of ignoring response  
566 misclassification and to evaluate the performance of the derived MLA in data scenarios similar  
567 to the situations in AMD studies. For each simulation scenario (data generating process), we  
568 simulated 1000 datasets, applied different models to the sampled data and evaluated the  
569 distribution of effect estimates, frequencies of significant statistical tests and coverage  
570 frequencies of confidence intervals for a central covariate of interest.

571 To sample data mimicking studies on AMD with internal validation data, we performed the  
572 following steps:

- 573 1) We sampled the true binary “worse-eye” AMD data  $Y$  for 5000 individuals by sampling from  
574 a Bernoulli distribution, where we modelled the success probability based on the logistic  
575 function of a linear predictor (corresponding to the assumed data generating process in  
576 logistic regression). For the linear predictor, we used an intercept of -0.25 (corresponding to  
577 an average probability of person-specific AMD of ~0.44) and a continuous standard normal  
578 covariate  $X$ . We varied the log OR of  $X$  on  $Y$  between zero (simulation under  $H_0$  of no effect)  
579 and one.

- 580 2) To create the true eye-specific disease data (two binary observations per individual,  $(Z_1, Z_2)$ )  
581 we specified the conditional probability of being affected in both eyes given disease in at least  
582 one eye (i.e.  $Y = 1$  based on “worse-eye definition),  $\delta$ , to be (on average)  $\delta = 1/(1 +$   
583  $\exp(-1)) = 0.73$ . We assumed this probability to be either constant or varying with the  
584 continuous covariate  $X$  based on formula  $\delta = 1/(1 + \exp(-(1 + 1 \times X))) = \text{Logist}(1 + 1 \times X)$ .  
585 For all individuals with sampled  $Y=1$ , we sampled a Bernoulli variable based on probability  $\delta$ ,  
586 to decide whether they were affected in both eyes or not. If they were affected on only one  
587 eye, we sampled randomly from the left or right.
- 588 3) To mimic the situation of missing information in one of two eyes, we sampled a Bernoulli  
589 random variable for each individual based on a fixed success probability (e.g. 0.75), to  
590 indicate whether information on both eyes was available. If not, we removed the disease  
591 information from a randomly selected eye.
- 592 4) To obtain eye-specific error-prone outcome data  $(Z_1^*, Z_2^*)$ , we conditioned on the true, sampled  
593 observations  $(Z_1, Z_2)$ , and sampled the error-prone outcomes based on specified  
594 classification probabilities, the sensitivity  $P(Z^* = 1|Z = 1)$  and specificity  $P(Z^* = 0|Z = 0)$ .  
595 Sensitivity and specificity were either fixed (non-differential misclassification, e.g.  
596  $\text{sens}=\text{spec}=0.9$ ) or varying between individuals based on the formula  $\text{sens} = \text{Logist}(2.20 +$   
597  $\beta_{\text{sens}} \times X)$  for different values of  $\beta_{\text{sens}}$  (analogously for the specificity, corresponding to an  
598 average  $\text{sens}=\text{spec}=0.9$ ).
- 599 5) Afterwards, we split the data into two parts, the “main study” and the “validation” subset based  
600 on defined fractions (e.g.  $n^{\text{val}} = 1000$ ,  $n^{\text{main}} = 4000$ ). For the validation subset we kept both,  
601 the true and the error-prone eye-specific AMD observations  $(Z_1, Z_2, Z_1^*, Z_2^*)$ ; for the main study,  
602 we kept only the error-prone outcomes  $(Z_1^*, Z_2^*)$  (or only the respective information for one of  
603 the two eyes, when information in one eye was missing for an individual).
- 604 6) For the naïve analysis ignoring response misclassification, we defined an observed, binary  
605 naïve person-specific outcome  $Y_{\text{obs}}^*$  the following way: for individuals from the validation data,  
606 we used the true eye-specific disease information; for individuals from the main study data,  
607 we used the error-prone eye-specific information. When disease information was available

608 for both eyes, we defined  $Y_{\text{obs}}^* = \max(Z_1, Z_2)$  or  $Y_{\text{obs}}^* = \max(Z_1^*, Z_2^*)$ , respectively; for  
609 observations with information only on one eye  $Z_1$ , we used  $Y_{\text{obs}}^* = Z_1$  or  $Y_{\text{obs}}^* = Z_1^*$ . For  
610 individuals from the validation data with information on both eyes,  $Y_{\text{obs}}^* = \max(Z_1, Z_2)$   
611 corresponds to the true  $Y$ ; for all others,  $Y_{\text{obs}}^*$  might be misclassified.

612 For each sampled dataset we estimated three models: 1) standard logistic regression based on  
613 the error-prone naïve worse-entity outcome  $Y_{\text{obs}}^*$ , 2) the derived MLA (see above) modelling the  
614 probability of person-specific AMD and the probability of AMD in both eyes given AMD in at least  
615 one eye,  $\delta$ , based on covariate  $X$ , while assuming a constant eye-specific sensitivity and  
616 specificity and accounting for missing information in one of two eyes (MLA1), and 3) the derived  
617 MLA allowing for a dependency of sensitivity and specificity on  $X$  (MLA2).

618

### 619 **Appendix C. Power analysis for reported lead variants based on UK Biobank sample size.**

620 We wanted to evaluate the impact of using the MLA on selected variants including the 34 reported  
621 lead variants known for their association with advanced AMD. Given reported effect sizes and  
622 effect allele frequencies (EAF), we expected the power to detect some of these 34 associations  
623 to be limited in a sample size of approximately 3,500 cases (and more controls). Therefore, we  
624 aimed to assess the power to detect reported genetic associations for AMD in the available data  
625 of UK Biobank, to focus our analyses with the MLA only on adequately powered reported  
626 associations and to avoid overinterpreting results from underpowered analyses. It is, however,  
627 not fully straight forward how to compute power for the scenario of “any AMD” from machine  
628 learning based disease classification, due to the power-diminishing effect of misclassification and  
629 some uncertainty of what effect size to use. We chose to use the reported<sup>10</sup> EAFs in advanced  
630 AMD cases and AMD-free controls for the established 34 lead variants and computed the power  
631 for a t-Test on EAFs for differently sized groups, given the 3,544 cases and 44,521 controls  
632 derived from the automated “any AMD” classification in the UK Biobank GWAS data  
633 (**Supplemental Table 2**). The standard error of the difference in EAFs between cases and  
634 controls was derived based on the formula

635 
$$se_{diff} = \sqrt{\frac{n_{case} \times eaf_{case} \times (1 - eaf_{case}) + n_{contr} \times eaf_{contr} \times (1 - eaf_{contr})}{n_{case} + n_{contr}}}$$

636 Based on these power calculations, we selected all lead variants with at least 80% power to yield  
637 nominally significant associations in UK Biobank. By this, we made the assumptions that EAFs  
638 in advanced AMD cases are transferable to EAFs of “any AMD” cases and that no  
639 misclassification was present in the machine learning derived any AMD classification. Therefore,  
640 this is probably an overestimate of available power. We performed the power analysis, however,  
641 mainly to dismiss variants with an obvious lack of power, while trying to include as many variants  
642 as reasonable in our analyses using the MLA.

643

#### 644 **Appendix D. MLA avoids bias and excess of type-I error in simulation studies.**

645 In our simulation study, we investigated bias and type-I error of logistic-regression based  
646 association estimates for a binary worse-entity outcome  $Y := \max(Z_1, Z_2) \in \{0,1\}$  and a  
647 continuous covariate  $X$ , when error-prone single-entity observations  $(Z_1^*, Z_2^*) \in \{0,1\}$  are  
648 observed instead of the true entity-specific disease classifications  $(Z_1, Z_2) \in \{0,1\}$ . When utilizing  
649 the error-prone observations for deriving the worse-entity outcomes  $Y^* := \max(Z_1^*, Z_2^*)$ , the entity-  
650 specific misclassification is passed on to the worse-entity disease stage. We compare the  
651 performance of the naïve analysis (logistic regression ignoring misclassification) and the two  
652 versions of our MLA for different simulation scenarios.

653 In the naïve analysis, we found a similar pattern for bilateral disease misclassification as reported  
654 for classic diseases<sup>5,7</sup>: (i) under the null hypothesis (**Table 1, Supplemental Table 1**,  $\beta_Y = 0$ ),  
655 we found biased estimates and a lack of type-I error control (potential for false-positive  
656 association findings) for differential misclassification. With non-differential misclassification,  
657 estimates were unbiased and type-I error frequencies were at the desired levels. (ii) When  $X$  was  
658 associated with true AMD (**Table 1, Supplemental Table 1**,  $\beta_Y = 1$ ), effect estimates were  
659 biased towards the null for non-differential misclassification and into any direction for differential  
660 misclassification. Specific for the bilateral disease situation was (iii) increasing bias with



661 increasingly missing AMD in one of the two eyes, and (iv) a larger bias by decreased specificity  
662 than by decreased sensitivity. (**Table 1, Supplemental Table 1**).

663 In logistic regression, the larger the misclassification probabilities, the larger the bias of  
664 estimates<sup>5</sup>, with similar influence of increased probabilities for false-positive and false-negative  
665 classifications for balanced data. In the following, we provide an explanation of the findings (iii)  
666 and (iv) for bilateral diseases from above. Finding (iii) is explained by the fact that an increased  
667 fraction of missing eyes implies a reduced sensitivity for person-specific AMD: AMD in the  
668 missing eye can be overlooked, which can lead to a false-negative person-specific AMD  
669 classification if only the missing eye of an individual is affected. Finding (iv) was that decreased  
670 specificity had larger impact on bias than decreased sensitivity, e.g. for (sens, spec)=(0.9, 0.9)  
671 and a fraction of 25% of individuals with “missing eyes” and a true log OR of X on Y of 1 the  
672 observed bias was -0.27. When the sensitivity was reduced to 0.8 (specificity=0.9), the bias  
673 increased (in absolute value) to -0.32; when the specificity was reduced to 0.8 (sensitivity=0.9),  
674 the bias increased to -0.39. This can be explained by rewriting the probability of misclassification  
675 in the worse-entity outcome,  $P(Y^* \neq Y)$  as

$$\begin{aligned} 676 \quad P(Y^* \neq Y) &= P(Y^* = 1|Y = 0)P(Y = 0) + P(Y^* = 0|Y = 1)P(Y = 1) \\ 677 \quad &= P(\max(Z_1^*, Z_2^*) = 1|Z_1 = 0, Z_2 = 0)P(Y = 0) + P(Z_1^* = 0, Z_2^* = 0|\max(Z_1, Z_2) = 1)P(Y = 1) \\ 678 \quad &= (1 - \text{spec}^2)P(Y = 0) + ((1 - \text{sens})^2\delta + \text{spec}(1 - \text{sens})(1 - \delta))P(Y = 1), \end{aligned}$$

679 This illustrates the dependency of  $P(Y^* \neq Y)$  on entity-specific sensitivity, specificity, probability  
680 of disease in both entities given disease in one eye  $\delta$ , and the fraction of truly affected individuals  
681  $P(Y = 1)$ . This probability can be evaluated for different combinations of parameters: for example,  
682 in the simulation study, we assumed  $P(Y = 1) = 0.44$ ,  $\delta = 0.75$  (**Appendix B**), which leads to a  
683 misclassification probability of 12%, 14%, or 22% for (sens, spec)=(0.9, 0.9), (sens, spec)=(0.8,  
684 0.9), or (sens, spec)=(0.9, 0.8), respectively, illustrating the larger impact of reducing specificity.  
685 This is even more true in scenarios with a lower fraction of affected individuals: if we assume a  
686 probability of person-specific disease of 0.10 instead of 0.44, we obtain misclassification  
687 probabilities of 17%, 18%, or 33%, for the same combinations of sensitivity and specificity. A  
688 reduced entity-specific specificity increases the probability of falsely classifying healthy entities

689 towards disease, and falsely classifying only one of two healthy entities towards disease is  
690 sufficient to misclassify the person-specific disease status.

691 When applying the MLA1, we found it to effectively correct for bias and to yield the expected  
692 confidence interval coverage rates (~95%) when the misclassification was non-differential, but  
693 we found it to still result in biased estimates and excess type-I error when the misclassification  
694 was differential (**Table 1, Supplemental Table 1**). When applying the MLA2, we found it effective  
695 in bias correction and type-I error control under all misclassification scenarios, but with larger  
696 standard errors due to the larger number of parameters in the model (**Table 1, Supplemental**  
697 **Table 1**). Overall, our simulation results documented substantial bias and lack of type-I error  
698 control when the naïve analysis was applied to misclassified data and our MLA to effectively  
699 remove bias and keep type-I error when specified correctly.

700

#### 701 **Appendix E. Detailed results of MLA for the selected 26 variants.**

702 For estimating sensitivity and specificity, we found the following: (i) for the 3 lead variants from  
703 this GWAS (*CFH*, *ARMS2/HTRA1*, or *HERC2*, respectively), the MLA1-derived sensitivity and  
704 specificity (at mean age and two copies of the non-effect allele) showed only small differences  
705 between the 3 variants (sensitivity = 65%, 67%, 63%; specificity=98%, 98%, 99%, respectively,  
706 **Supplemental Table 6a**). From a model without including a genetic covariate, we obtained an  
707 overall sensitivity of 64.5% (95%-CI: 60.1%, 68.7%) and a specificity of 98.6% (98.4%, 98.8%).  
708 (ii) We did not find strong evidence for associations with age using MLA1 or MLA2 based on any  
709 of the 26 selected variants, except for an association of the specificity with age based on MLA1  
710 for the *HERC2* variant that disappeared when applying MLA2 (age- $P=6.71 \times 10^{-9}$  or 0.70,  
711 respectively, **Supplemental Table 6a**). (iii) Applying MLA2, we found no association of the  
712 sensitivity with any selected variant ( $P>0.05/(23 \times 2)$ ), but a strong association of the specificity  
713 with the *HERC2* lead variant rs12913832 and with the reported *CFH* lead variant rs10922109  
714 ( $OR_{spec}=0.64$ ,  $P_{spec}=7.38 \times 10^{-9}$  and  $OR_{spec}=1.36$ ,  $P_{spec}=2.29 \times 10^{-4}$ , respectively; **Supplemental**  
715 **Table 6**).

716 Second, we obtained genetic association estimates from MLA1 and MLA2 accounting for  
717 misclassification and compared these with naïve analysis estimates. We found interesting  
718 patterns: (i) when applying MLA1, we found comparable, slightly increased effect estimates for  
719 the *CFH*, *ARMS2/HTRA1*, and *HERC2* lead variant when compared to the naïve analysis (MLA1:  
720 OR=1.23, 1.48, 1.34; P=1.69x10<sup>-6</sup>, 8.9x10<sup>-18</sup>, 1.11x10<sup>-12</sup>; naïve: OR=1.14, 1.30, 1.26, P=6.18x10<sup>-7</sup>,  
721 2.44x10<sup>-20</sup>, 4.16x10<sup>-16</sup>; **Figure 2, Supplemental Table 7a**). (ii) When applying MLA2, we found  
722 similar effect estimates for *CFH* and *ARMS2/HTRA1* compared to MLA1 and naïve analysis  
723 (OR=1.19 or 1.28, respectively), which is in line with limited bias due to differential  
724 misclassification. We also found larger P-values (P=0.02 or 2.47x10<sup>-4</sup>, respectively, which is in  
725 line with larger uncertainty when estimating more model parameters. In contrast, we found a  
726 completely vanished effect estimate for the *HERC2* variant (MLA2: OR=1.03, P=0.76; **Figure 2,**  
727 **Supplemental Table 7a**), indicating a bias in the naïve analysis and MLA1 when ignoring a  
728 differential misclassification. (iii) Effect estimates for the 3 reported *CFH* variants increased when  
729 applying MLA2 compared to the naïve analysis. This was particularly interesting for the reported  
730 *CFH* lead variant rs10922109, where we now observed a nominally significant association into  
731 the reported direction (MLA2: OR=1.15, P=0.047; naïve: OR=1.00, P=0.98; **Supplemental Table**  
732 **7c**). This is in line with the observed association of the specificity with this *CFH* variant. (iv) For  
733 the other 20 reported lead variants, we found many variants with increased effect estimates by  
734 MLA1 or MLA2 compared to the naïve analysis; effect estimates were mostly more comparable  
735 to reported effect sizes for advanced AMD<sup>10</sup> (**Figure 3c**). For one variant, this MLA2 analysis  
736 yielded an effect into the opposite direction compared to the reported effect direction, which is  
737 the *C9* lead variant (OR=0.83, P=0.59). With an effect allele frequency of 1%, it is the rarest  
738 analyzed variant of the 26 selected variants and estimates from the reported association as well  
739 as for the MLA2 analysis have low precision (i.e. large standard errors).

740

#### 741 **Supplemental Data**

742 Supplemental Data include one figure and seven tables.

743

744 **Declaration of Interest**

745 I.M.H. received funding from Roche Diagnostics for a project unrelated to this work.

746

747 **Acknowledgements**

748 This work was funded by DFG HE 3690/5-1 (to I.M.H.) and NIH R01 RES511967 (to I.M.H.), the  
749 University of Regensburg and the Ludwig-Maximilians-University München. The UK Biobank  
750 (accessed via application number 33999) was established by the Wellcome Trust medical charity,  
751 Medical Research Council, Department of Health, Scottish Government and the Northwest  
752 Regional Development Agency. It has also had funding from the Welsh Assembly Government,  
753 British Heart Foundation and Diabetes UK.

754

755 **Web Resources**

756 An open source R implementation of the MLA to account for misclassification in bilateral disease  
757 in genetic association analyses is available at:

758 <https://www.genepi-regensburg.de/MLA-bilateral/> (upon publication)

759 Convolutional Neural Net Ensemble used for automated AMD classification and  
760 recommendations by the authors:

761 <https://github.com/RegensburgMedicalImageComputing/ARIANNA>;

762 IrfanView: <https://www.irfanview.com/>;

763 GWAS catalogue: <https://www.ebi.ac.uk/gwas/>

764

765 **References**

766 1. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak,  
767 J.A.W.M., van Ginneken, B., and Sánchez, C.I. (2017). A survey on deep learning in medical  
768 image analysis. *Med. Image Anal.* 42, 60–88.

769 2. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012).  
770 A unifying view on dataset shift in classification. *Pattern Recognit.* 45, 521–530.

771 3. Csurka, G. (2017). A comprehensive survey on domain adaptation for visual applications. In

- 772 Domain Adaptation in Computer Vision Applications, (Springer), pp. 1–35.
- 773 4. Heinze-Deml, C., and Meinshausen, N. (2017). Conditional variance penalties and domain  
774 shift robustness. *ArXiv Prepr. ArXiv1710.11469*.
- 775 5. Neuhaus, J. (1999). Bias and efficiency loss due to misclassified responses in binary  
776 regression. *Biometrika* 86, 843–855.
- 777 6. Hausman, J.A., Abrevaya, J., and Scott-Morton, F.M. (1998). Misclassification of the  
778 dependent variable in a discrete-response setting. *J. Econom.* 87, 239–269.
- 779 7. Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error  
780 in Nonlinear Models* (Chapman and Hall/CRC).
- 781 8. Lyles, R.H., Tang, L., Superak, H.M., King, C.C., Celentano, D.D., Lo, Y., and Sobel, J.D.  
782 (2011). Validation data-based adjustments for outcome misclassification in logistic regression:  
783 an illustration. *Epidemiology* 22, 589–597.
- 784 9. Klein, R., Meuer, S.M., Myers, C.E., Buitendijk, G.H.S., Rochtchina, E., Choudhury, F., de  
785 Jong, P.T.V.M., McKean-Cowdin, R., Iyengar, S.K., Gao, X., et al. (2014). Harmonizing the  
786 Classification of Age-related Macular Degeneration in the Three-Continent AMD Consortium.  
787 *Ophthalmic Epidemiol.* 21, 14–23.
- 788 10. Fritsche, L.G., Igl, W., Bailey, J.N.C., Grassmann, F., Sengupta, S., Bragg-Gresham, J.L.,  
789 Burdon, K.P., Hebbiring, S.J., Wen, C., Gorski, M., et al. (2016). A large genome-wide association  
790 study of age-related macular degeneration highlights contributions of rare and common variants.  
791 *Nat. Genet.* 48, 134–143.
- 792 11. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic,  
793 D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping  
794 and genomic data. *Nature* 562, 203–209.
- 795 12. Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E., and Bressler, N.M. (2017).  
796 Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using  
797 Deep Convolutional Neural Networks. *JAMA Ophthalmol.* 135, 1170.
- 798 13. Ting, D.S.W., Cheung, C.Y.-L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H.,  
799 Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., et al. (2017). Development and Validation of a Deep

- 800 Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From  
801 Multiethnic Populations With Diabetes. *JAMA* 318, 2211.
- 802 14. Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M.E., Linkohr, B.,  
803 Peters, A., Heid, I.M., Palm, C., and Weber, B.H.F. (2018). A Deep Learning Algorithm for  
804 Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular  
805 Degeneration from Color Fundus Photography. *Ophthalmology* 125, 1410–1420.
- 806 15. Peng, Y., Dharssi, S., Chen, Q., Keenan, T.D., Agrón, E., Wong, W.T., Chew, E.Y., and Lu,  
807 Z. (2019). DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based  
808 Age-related Macular Degeneration Severity from Color Fundus Photographs. *Ophthalmology*  
809 126, 565–575.
- 810 16. Günther, F., Brandl, C., Heid, I.M., and Küchenhoff, H. (2019). Response misclassification in  
811 studies on bilateral diseases. *Biom. J.* 61, 1033–1048.
- 812 17. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M.,  
813 Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes  
814 for genotype imputation. *Nat. Genet.* 48, 1279–1283.
- 815 18. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C.,  
816 Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and  
817 disease. *Nature* 526, 82–89.
- 818 19. Keane, P.A., Grossi, C.M., Foster, P.J., Yang, Q., Reisman, C.A., Chan, K., Peto, T., Thomas,  
819 D., Patel, P.J., and UK Biobank Eye Vision Consortium (2016). Optical Coherence Tomography  
820 in the UK Biobank Study - Rapid Automated Analysis of Retinal Thickness for Large Population-  
821 Based Studies. *PLoS One* 11, e0164095.
- 822 20. Davis, M.D., Gangnon, R.E., Lee, L.-Y., Hubbard, L.D., Klein, B.E.K., Klein, R., Ferris, F.L.,  
823 Bressler, S.B., Milton, R.C., and Age-Related Eye Disease Study Group (2005). The Age-Related  
824 Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17.  
825 *Arch. Ophthalmol. (Chicago, Ill. 1960)* 123, 1484–1498.
- 826 21. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model  
827 association for biobank-scale datasets. *Nat. Genet.* 50, 906–908.

- 828 22. Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., Waterworth,  
829 D., Beckmann, J.S., and Bergmann, S. (2011). Methods for testing association between  
830 uncertain genotypes and quantitative traits. *Biostatistics* 12, 1–17.
- 831 23. Devlin, A.B., Roeder, K., and Devlin, B. (2013). Genomic Control for Association. 55, 997–  
832 1004.
- 833 24. Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring  
834 population-specific haplotype structure and linking correlated alleles of possible functional  
835 variants. *Bioinformatics* 31, 3555–3557.
- 836 25. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J., and GoT2D investigators (2013).  
837 Recommended joint and meta-analysis strategies for case-control association testing of single  
838 low-count variants. *Genet. Epidemiol.* 37, 539–550.
- 839 26. Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P.N., Stark, M.S., Hayward, N.K., Martin, N.G.,  
840 and Montgomery, G.W. (2008). A single SNP in an evolutionary conserved region within intron  
841 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet.* 82, 424–  
842 431.
- 843 27. Stark, K., Olden, M., Brandl, C., Dietl, A., Zimmermann, M.E., Schelter, S.C., Loss, J.,  
844 Leitzmann, M.F., Böger, C.A., Luchner, A., et al. (2015). The German AugUR study: study  
845 protocol of a prospective study to investigate chronic diseases in the elderly. *BMC Geriatr.* 15,  
846 130.
- 847 28. Brandl, C., Zimmermann, M.E., Günther, F., Barth, T., Olden, M., Schelter, S.C., Kronenberg,  
848 F., Loss, J., Küchenhoff, H., Helbig, H., et al. (2018). On the impact of different approaches to  
849 classify age-related macular degeneration: Results from the German AugUR study. *Sci. Rep.* 8,  
850 8675.
- 851 29. Chakravarthy, U., Wong, T.Y., Fletcher, A., Piau, E., Evans, C., Zlateva, G., Buggage, R.,  
852 Pleil, A., and Mitchell, P. (2010). Clinical risk factors for age-related macular degeneration: a  
853 systematic review and meta-analysis. *BMC Ophthalmol.* 10, 31.
- 854 30. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham,  
855 P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug

856 indications. *Nat. Genet.* *47*, 856–860.

857 31. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C.,  
858 McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog  
859 of published genome-wide association studies, targeted arrays and summary statistics 2019.  
860 *Nucleic Acids Res.* *47*, D1005–D1012.

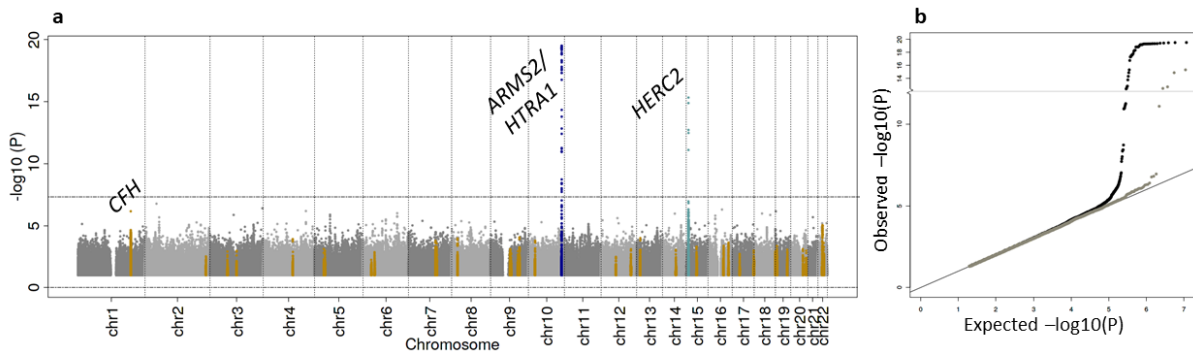
861 32. R Core Team (2019). R: A Language and Environment for Statistical Computing.

862

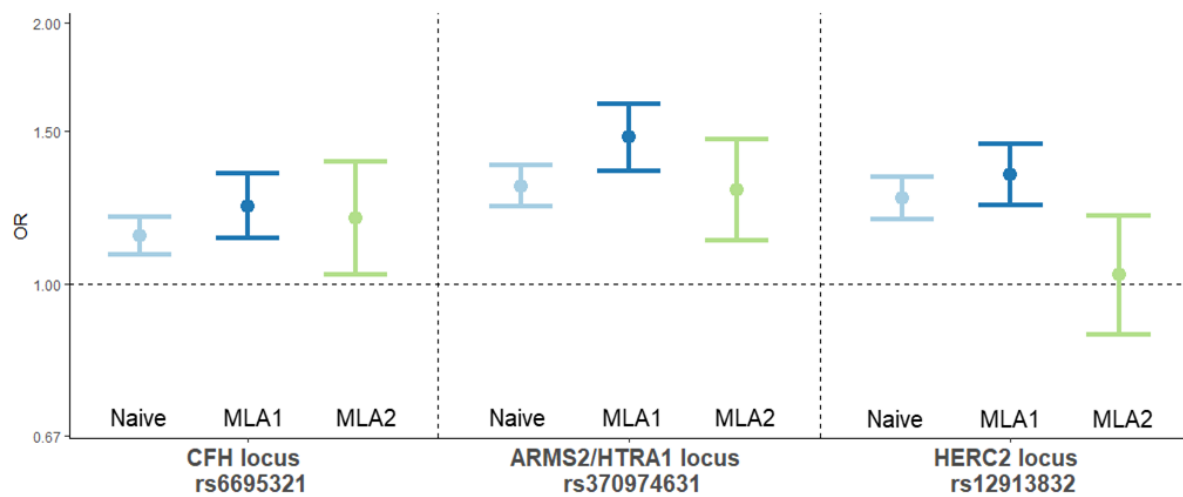


## FIGURES

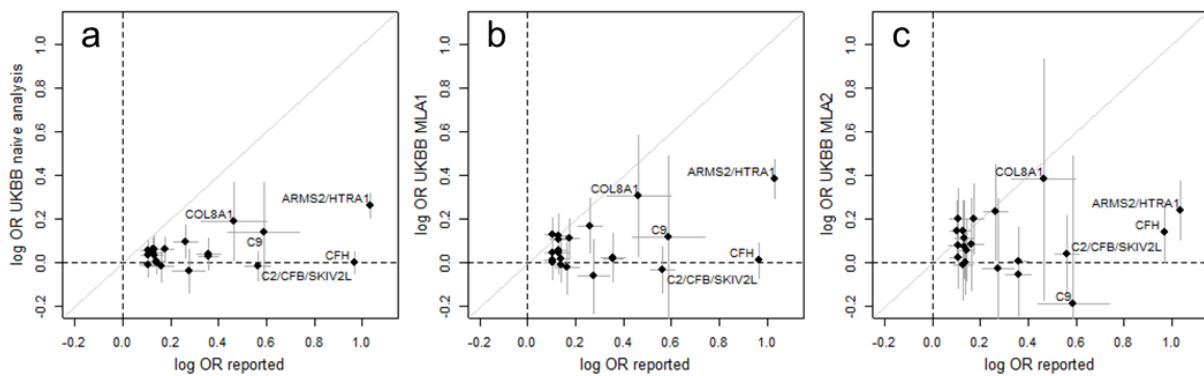
**Figure 1. GWAS results in UK Biobank based on automatically derived “any AMD” from naïve analysis.** Association analyses were conducted using the error-prone, machine learning derived AMD classification in UK Biobank participants with 3,544 “any AMD” cases and 44,521 controls via logistic regression adjusted for age and two genetic principal components, the *naïve analysis* ignoring misclassification. Shown are **a)** Manhattan Plot of 11,567,158 analyzed variants; dark blue: genome-wide significant and previously established<sup>10</sup> locus, light blue: unknown genome-wide significant locus, orange: other 33 previously established loci for advanced AMD), and **b)** expected versus observed  $-\log_{10} P$ -values; black: all variants, grey: all variants outside the 34 previously reported loci.






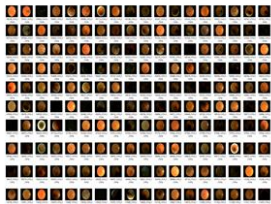
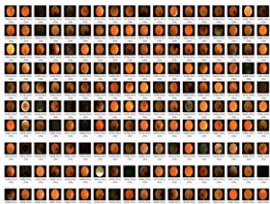
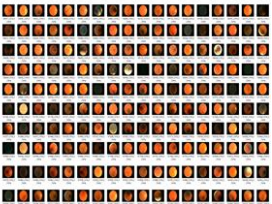
**Figure 2. Genetic effect estimates for the 3 lead variants in UK Biobank without and with accounting for misclassification.** Shown are genetic effect estimates and 95% confidence intervals for 3 lead variants from the GWAS on automated AMD classification with 3,544 “any AMD” cases and 44,521 controls from 3 models: without accounting for the misclassification; naïve analysis, light blue. With accounting for non-differential misclassification, i.e. no dependency on the genetic variant; MLA1, dark blue. And accounting for a differential misclassification, i.e. dependency on the genetic variant; MLA2, light green. Both MLAs accounted for missing AMD information in one of two eyes and a misclassification that depended on age.



**Figure 3. Comparison of 21 reported genetic effect estimates for advanced AMD with estimates for automatically derived “any AMD” from UK Biobank without and with accounting for misclassification.** We selected the 21 reported AMD lead variants, for which we had  $\geq 80\%$  power to detect them in this UK Biobank sample size with nominal significance. Shown are log OR effect estimates and 95% confidence intervals reported for advanced AMD on x-axis versus UK Biobank estimates for automatically derived “any AMD” on y-axis from **a)** the naïve analysis (logistic regression ignoring misclassification, **b)** MLA1, and **c)** MLA2.



**Figure 4. Evidence for differential misclassification in automatically derived AMD with respect to the *HERC2* variant rs12913832.** Shown are (i) estimated odds ratios from the naïve analysis ignoring misclassification and various characteristics per genotype group: (ii) the fraction of persons with self-reported “light eye color” in the AugUR study, (iii) randomly selected fundus images in UKBB, (iv) image-lightness quantified by mean average grayscale, (v) proportion of false-positive AMD in the automated classification (1-specificity) and 95% confidence interval estimated via MLA2, and (vi) observed proportion of manually ungradable images that were deemed gradable by the algorithm and classified as “any AMD” or “AMD-free”.

| <i>HERC2</i> (rs12913832)                                      |  |  |  |
|--|--|--|--|
| Genotype   | AA  | AG  | GG  |
| <b>Odds Ratio</b><br>(GWAS, naïve analysis)                    | 1.58   | 1.26   | Ref.   |
| <b>% light eye color</b><br>(AugUR study)                      | 14%  | 36%  | 97%  |
| <b>UKBB fundus images</b>                                      |   |    |   |
| <b>Average grayscale</b>                                       | 46.4   | 49.0   | 53.6   |
| <b>% false positive AMD</b>                                    | 3.0% [2.3%, 3.7%]  | 1.9% [1.7%, 2.2%]  | 1.2% [1.1%, 1.4%]  |
| <b>% manually ungradable with automated AMD classification</b> | 54.5% [32.2%, 75.6%]   | 54.8% [46.3%, 63.1%]   | 38.8% [32.1%, 45.9%]   |

## TABLES

**Table 1. Simulation results on effect estimates and empirical type-I error in naïve and MLA-analysis.** We evaluated the performance of naïve and MLA analysis of a quantitative covariate X and a binary bilateral disease Y, e.g. person-specific AMD, simulating various scenarios. For each scenario, we sampled 1000 data sets à 5000 individuals, 4000 with only error-prone eye-specific AMD classification, and 1000 with additional true AMD classification. Shown are performance measures from three models, naïve analysis, MLA1, or MLA2 assuming non-differential/differential misclassification regarding X, respectively, in various simulation scenarios. For the eight scenarios shown here, we assumed no association of X with  $\delta$ , the probability of AMD in both eyes given  $\geq 1$  affected eye; results were similar when modelling an association of X with  $\delta$ , see **Supplemental Table 1**. For each model and scenario, we report mean effect estimates  $\widehat{\beta}_Y$ , log OR per unit increase in standard-normal X, over all simulation runs, and the associated root mean squared error (RMSE), fraction of nominally significant effect estimates (% with  $P < 0.05$ ), and coverage frequencies of 95%-confidence intervals.

| Simulation Scenario                       |      |        |           |                |                | $\widehat{\beta}_Y$ |      |       |      |      |      | % with $P < 0.05$ |      |      | Cov. Freq. |       |       |
|---|------|--------|-----------|----------------|----------------|---------------------|------|-------|------|------|------|-------------------|------|------|------------|-------|-------|
| Sens                                      | Spec | %miss. | $\beta_Y$ | $\beta_{sens}$ | $\beta_{spec}$ | Naïve               |      | MLA1  |      | MLA2 |      | Naive             | MLA1 | MLA2 | Naive      | MLA1  | MLA2  |
|   |      |        |           |                |                | Mean                | RMSE | Mean  | RMSE | Mean | RMSE |                   |      |      |            |       |       |
| <b>Non-differential misclassification</b> |      |        |           |                |                |                     |      |       |      |      |      |                   |      |      |            |       |       |
| 0.9                                       | 0.9  | 0.25   | 0         | 0              | 0              | 0.00                | 0.03 | 0.00  | 0.04 | 0.00 | 0.04 | 5.3%              | 4.6% | 4.6% | 94.7%      | 95.4% | 95.4% |
| 0.9                                       | 0.9  | 0.25   | 1         | 0              | 0              | 0.73                | 0.27 | 1.00  | 0.05 | 1.00 | 0.05 | 100%              | 100% | 100% | 0.0%       | 96.5% | 96.3% |
| 0.9                                       | 0.9  | 0.75   | 1         | 0              | 0              | 0.69                | 0.31 | 1.00  | 0.06 | 1.00 | 0.07 | 100%              | 100% | 100% | 0.0%       | 94.4% | 93.5% |
| 0.8                                       | 0.8  | 0.25   | 1         | 0              | 0              | 0.56                | 0.44 | 1.00  | 0.06 | 1.00 | 0.07 | 100%              | 100% | 100% | 0.0%       | 95.0% | 95.0% |
| 0.8                                       | 0.9  | 0.25   | 1         | 0              | 0              | 0.68                | 0.32 | 1.00  | 0.05 | 1.00 | 0.06 | 100%              | 100% | 100% | 0.0%       | 97.0% | 95.9% |
| 0.9                                       | 0.8  | 0.25   | 1         | 0              | 0              | 0.61                | 0.39 | 1.00  | 0.06 | 1.00 | 0.06 | 100%              | 100% | 100% | 0.0%       | 95.3% | 94.8% |
| <b>Differential misclassification</b>     |      |        |           |                |                |                     |      |       |      |      |      |                   |      |      |            |       |       |
| 0.9                                       | 0.9  | 0.25   | 0         | -1             | 1              | -0.38               | 0.38 | -0.46 | 0.46 | 0.00 | 0.05 | 100%              | 100% | 4.7% | 0.0%       | 0.0%  | 95.3% |
| 0.9                                       | 0.9  | 0.25   | 1         | 1              | -1             | 1.14                | 0.14 | 1.39  | 0.40 | 1.00 | 0.06 | 100%              | 100% | 100% | 4.8%       | 0.0%  | 95.1% |

Sens/Spec = average sensitivity and specificity of error-prone, eye-specific AMD classification; %miss. = fraction of randomly selected individuals with missing AMD classification in one of two eyes;  $\beta_Y$  = log OR of X on true AMD,  $\beta_{sens}$  = log OR of X on the sensitivity or  $\beta_{spec}$  = log OR of X on the specificity of the eye-specific misclassification process, respectively.

**Table 2. Confusion matrices comparing manual and automated AMD classification per eye and per person.** Shown are absolute numbers and conditional classification probabilities, i.e. in row *i* and column *j*,  $P(\text{automated} = j \mid \text{manual} = i)$  as %, with *i, j* = "Ungradable", "No AMD", "Any AMD": **a)** for all eyes in the validation data; 4001 eyes of 2,013 individuals. **b)** For all persons in the overlap between validation data and GWAS; 1,327 persons.

---

**a) per eye (4,001 eyes, 2,013 individuals)**

| Manual     | Automated classification |            |           | Sum         |
|------------|--------------------------|------------|-----------|-------------|
|            | Ungradable               | No AMD     | Any AMD   |             |
| Ungradable | 813 (74%)                | 185 (17%)  | 103 (9%)  | 1101 (100%) |
| No AMD     | 107 (4%)                 | 2207 (90%) | 138 (6%)  | 2452 (100%) |
| Any AMD    | 20 (4%)                  | 103 (23%)  | 325 (73%) | 448 (100%)  |

---

**b) per person (1,327 individuals)**

| Manual        | Automated classification |           | Sum        |
|---------------|--------------------------|-----------|------------|
|               | No AMD                   | Any AMD   |            |
| Ungradable/NA | 202 (79%)                | 53 (21%)  | 255 (100%) |
| No AMD        | 750 (91%)                | 72 (9%)   | 822 (100%) |
| Any AMD       | 58 (23%)                 | 192 (77%) | 250 (100%) |

---

NA = true AMD status based on worse eye not available, since one eye was manually ungradable and the second AMD-free