Received: 11 December 2019 | Revised: 18 June 2020 | Accepted: 21 June 2020

DOI: 10.1002/gepi.22336

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

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

Felix Guenther[1,2] | **Caroline Brandl**[1,3] | **Thomas W. Winkler**[1] | **Veronika Wanner**[1] | **Klaus Stark**[1] | **Helmut Kuechenhoff**[2] | **Iris M. Heid**[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany

[2]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of Munich, Munich, Germany

[3]Department of Ophthalmology, University Hospital Regensburg, Regensburg, Germany

Correspondence
Helmut Kuechenhoff, Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of Munich, 80539 Munich, Germany.
Email: kuechenhoff@stat.uni-muenchen.de

Iris M. Heid, Department of Genetic Epidemiology, University of Regensburg, 93053 Regensburg, Germany.
Email: iris.heid@klinik.uni-regensburg.de

Felix Guenther, Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of Munich, 80539 Munich, Germany.
Email: felix.guenther@stat.uni-muenchen.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG HE 3690/5-1; National Institutes of Health, Grant/Award Number: NIH R01 EY RES 511967

Abstract

Imaging technology and machine learning algorithms for disease classification set the stage for high-throughput phenotyping and promising new avenues for genome-wide association studies (GWAS). Despite emerging algorithms, there has been no successful application in GWAS so far. We establish machine learning-based phenotyping in genetic association analysis as misclassification problem. To evaluate chances and challenges, we performed a GWAS based on automatically classified age-related macular degeneration (AMD) in UK Biobank (images from 135,500 eyes; 68,400 persons). We quantified misclassification of automatically derived AMD in internal validation data (4,001 eyes; 2,013 persons) and developed a maximum likelihood approach (MLA) to account for it when estimating genetic association. We demonstrate that our MLA guards against bias and artifacts in simulation studies. By combining a GWAS on automatically derived AMD and our MLA in UK Biobank data, we were able to dissect true association (*ARMS2*/*HTRA1*, *CFH*) from artifacts (near *HERC2*) and identified eye color as associated with the misclassification. On this example, we provide a proof-of-concept that a GWAS using machine learning-derived disease classification yields relevant results and that misclassification needs to be considered in analysis. These findings generalize to other phenotypes and emphasize the utility of genetic data for understanding misclassification structure of machine learning algorithms.

KEYWORDS

age-related macular degeneration (AMD), genome-wide association study, machine learning-based disease classification, response misclassification, UK Biobank

Helmut Kuechenhoff and Iris M. Heid jointly supervised this work.

[Correction added on 01 January, 2021: In compliance with contract with Projekt DEAL, co-correspondence authorship has been assigned to Felix Guenther]

1 | INTRODUCTION

Imaging technology allows for noninvasive access to detailed disease features in large studies and genome-wide association studies (GWAS) on such disease phenotypes can be expected to accelerate knowledge gain. However, image-based disease classification can be challenging for large sample sizes due to time-intensive, tiresome manual inspection. This limitation can be overcome by automated disease classification via machine learning and particularly deep learning algorithms. Such emerging approaches (Litjens et al., 2017) can classify diseases effortlessly also for huge sample sizes as needed for GWAS or other Omics approaches.

Deep learning algorithms require enormous input data with available gold standard classification, to "learn" classification reliably. Once trained and tested, the algorithms can be applied to external image data, but they cannot critically reflect unusual findings or incorporate unforeseen aspects, for which the human eye and brain have unmet capability. At the current time, input data to train algorithms are limited and often specific to a certain setting (e.g., patients from a clinic). Some characteristics that appear useful for disease classification in one setting might be misinterpreted in another, which can hamper transferability of trained models; a topic discussed as dataset shift or domain shift (Csurka, 2017; Heinze-Deml & Meinshausen, 2017; Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, & Herrera, 2012). Most predictions of deep learning algorithms for image-based disease classification will be error-prone and the structure of misclassification will generally be unknown. When using automated disease classification as outcome for association analyses and GWAS, the underlying response misclassification is usually unaccounted for, giving rise to biased effect estimates and potentially false-positive associations (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Hausman, Abrevaya, & Scott-Morton, 1998; Neuhaus, 1999). Extent and structure of the misclassification process can be assessed by *internal validation data*, that is, a subset of participants with both automated and gold standard classification, which can also be utilized to account for response misclassification in statistical models (Carroll et al., 2006; Lyles et al., 2011).

At present, it is unclear whether machine learning-based disease classification is of any utility for association analyses, particularly for detecting disease signals in GWAS. We thus set out to evaluate machine learning-derived disease classification in GWAS on the example of age-related macular degeneration (AMD) and we developed a statistical approach accounting for the implied response misclassification. AMD is an ideal role model, as a common disease ascertained via imaging of the central retina (Klein et al., 2014) and with particularly strong known genetic effects (Fritsche et al., 2016). The manual grading of images for AMD requires a substantial effort by trained staff and is currently an obstacle for homogeneous disease classification within and across large studies. For example, in UK Biobank (Bycroft et al., 2018), >135,000 color fundus images are available for >68,000 study participants, but there is no manually classified AMD available so far. Several machine learning algorithms have been emerging to classify AMD: they show promising performance, but still yield misclassified predictions, have acknowledged issues due to domain shift or insufficient sample size for training, or lack validation in external studies (Burlina et al., 2017; Grassmann et al., 2018; Peng et al., 2019; Ting et al., 2017). So far, there is no GWAS on fundus image ascertained AMD available in UK Biobank, manually classified or machine learning based.

2 | MATERIALS AND METHODS

2.1 | Machine learning-based disease classification in GWAS as misclassification problem

We consider a binary disease Y, for which each individual has a true status of disease (disease yes/no). A *gold standard* classification often involves manual grading of medical images via trained medical staff, which is considered here to correspond to the true disease classification. When applying a trained machine learning algorithm on medical images, we yield an automated disease classification Y^* for each individual. For an individual i with true disease status $Y_i = y_i$, the classification can either be correct or wrong ($y^*_i = y_i$, or $y^*_i \neq y_i$). If a gold standard classification is available (for at least a subset of study participants, internal validation data), the performance of the algorithm can be quantified by cross-tabulation of the observed error-prone y^* and the gold-standard classification y across all participants in the validation substudy (confusion matrix); the (mis-) classification process can be characterized by classification probabilities $P(Y^*{=}k|Y = l)$, for $l,k \in \{0, 1\}$. For $l = k = 1$ and $l = k = 0$, these probabilities correspond to the sensitivity and specificity of the algorithm, respectively.

In the following, we focus on *bilateral diseases* due to our motivating example of an eye disease (AMD): for each individual i, two entity-specific binary disease variables $Z_{1i}, Z_{2i} \in \{0, 1\}$ (here: AMD per eye) are used to define the binary person-specific disease status as the "worse entity disease status" $Y_i := \max(Z_{1i}, Z_{2i})$, corresponding to "AMD in at least one eye" versus "AMD in none of the two eyes" in our example. The error-prone

machine learning-based classification of entity-specific disease Z^*_{1i}, Z^*_{2i}, will propagate to error-prone person-specific disease status, $Y_i^* = \max(Z_{1i}^*, Z_{2i}^*)$, when compared to the manually graded "true" Y_i.

We were interested in evaluating the potential and consequences of such automatically classified disease in GWAS. The standard approach in GWAS is logistic regression for modeling the association of a genetic variant (observed as genotypes $\in\{0,1,2\}$ or imputed allelic dosages $\in[0,2]$) with a binary disease status, usually adjusted for other covariates like age, sex, and genetic principal components; Wald tests are used to test for genetic association, accounting for multiple testing by judging at a Bonferroni-corrected significance level of $p < 5 \times 10^{-8}$. When the association of the genetic variant with the true disease status Y (here: manually classified person-specific AMD) follows a logistic regression model, a naïve usage of the error-prone disease status Y* (here: automatically derived person-specific AMD) in standard logistic regression corresponds to the utilization of a misspecified model for the observed data (*naïve association analysis*). This has known consequences of decreased power, biased (genetic) association estimates, and potentially false-positive associations (Carroll et al., 2006; Hausman et al., 1998; Neuhaus, 1999). With additional information on the misclassification process, it is possible to correct for the bias and inflated type-I error. However, it is in general not possible to recover power lost due to misclassification.

2.2 | MLA to adjust for response misclassification in bilateral disease

In contrast to classical diseases and logistic regression (Carroll et al., 2006; Hausman et al., 1998; Neuhaus, 1999), no method is currently available to adjust for response misclassification in bilateral diseases. As described previously (Günther, Brandl, Heid, & Küchenhoff, 2019), the conceptual challenge is to account for two types of misclassification: (a) entity-specific misclassification that propagates to an error-prone person-specific disease status; and (b) person-specific misclassification from a missing disease status in one of the two entities. We thus developed an MLA to account for the fact that we are using an error-prone response $Y^*_i := \max(Z^*_{1i}, Z^*_{2i})$, $Z^*_{1i}, Z^*_{2i} \in \{0,1\}$, in the association analysis, while the true disease $Y_i := \max(Z_{1i}, Z_{2i})$, $Z_{1i}, Z_{2i} \in \{0,1\}$ is assumed to follow a logistic regression model.

Details are provided in Appendix A. The general idea of the MLA is to factorize the likelihood of the observed, error-prone response data into two parts, the model for the association between risk factor and true (but in general unobserved) response (*true association model*) and a model for the misclassification process (*misclassification model*). We adapted this well-established methodology for analyzing misclassified binary response data (Carroll et al., 2006; Lyles et al., 2011) to the scenario of bilateral disease with a "worse-entity" disease definition (i.e., the person-specific disease status is defined as the status of the worse entity). We assume conditional independence of the classification in the two entities z^*_{li}, z^*_{2i} of an individual i, given the true disease status. This assumption can be checked by validation data. Then, we have

$$P(z_{1i}^*, z_{2i}^* | x_i) = \sum_{z_{1i}, z_{2i} \in \{0,1\}} \underbrace{P(z_{1i}^* \mid z_{1i}, x_i) \times P(z_{2i}^* \mid z_{2i}, x_i)}_{\text{misclassification model}} \times \underbrace{P(z_{1i}, z_{2i}, \mid x_i)}_{\text{true association model}}.$$

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

Our developed approach allows us to adjust for both the entity-specific misclassification from an automated classification and the misclassification of the person-specific status when one entity is ungradable. Altogether, we model four parameters in the MLA: (a) the conditional probability of worse entity disease given the covariate of interest; (b) the probability of disease in both entities conditional on the disease in at least one entity (to adjust for missing information of one of two entities); as well as (c) the sensitivity and (d) the specificity of the entity-specific misclassification process. For each parameter, the conditional probabilities are modeled using the logistic function (as in standard logistic regression) allowing for a dependency on a parameter-specific set of person-specific covariates. An open source R (R Core Team, 2019) implementation is available.

2.3 | Simulation study to investigate the performance of the MLA

We repeatedly simulated association data for a standard normal covariate X and a (true and error-prone) binary outcome of a *bilateral* disease. To do this, we (a) sampled

the true, person-specific worse entity status associated with X for 5,000 individuals, (b) derived the true entity-specific disease status (e.g., manual eye-specific AMD classification) given assumptions, (c) sampled the entity-specific error-prone disease status (e.g., automated AMD classification), and (d) derived an error-prone, person-specific disease status. Afterward, we removed the true disease status for 4,000 individuals, yielding a subset of 1,000 with both true and error-prone disease status available (validation data). In different simulation scenarios, we varied sensitivity and specificity of the entity-specific classification. Classification probabilities were either constant for all individuals (nondifferential misclassification) or varying with X (differential misclassification). We also varied the fraction of individuals with missing classification in one of two entities (25–75%). Data were sampled with or without an effect of X on the true person-specific response Y ($\beta_Y \in \{0, 1\}$, log odds ratio [OR]) and on the probability δ of having disease in both entities given disease in at least one entity ($\beta_\delta \in \{0, 1\}$, log OR). We estimated the covariate effect using the naive analysis (logistic regression, which ignores misclassification) and the developed MLA1 and MLA2 accounting for response misclassification without (MLA1) and with allowing (MLA2) for differential misclassification, respectively. To compare the performance of the naïve analysis and the derived MLA, we investigated the distribution of effect estimates $\hat{\beta}_Y$ across 1,000 simulation runs in each scenario, computed the mean squared error of estimates relative to true effects, frequencies of rejected tests for no association, and coverage frequencies of 95% confidence intervals (CI). A detailed description of the simulation study, data sampling, and estimated models is given in Appendix B.

2.4 | UK Biobank study information and data

UK Biobank recruited ~500,000 individuals aged 40–69 years from across the United Kingdom. Genetic data are available from the Affymetrix UK Biobank Axiom Array imputed to the Haplotype Reference Consortium (McCarthy et al., 2016) and the UK10K haplotype resource (Walter et al., 2015); details described elsewhere (Bycroft et al., 2018). The UK Biobank baseline data contains 135,500 fundus images of 68,400 individuals. The images are taken with the Topcon 3D OCT-1000 Mark II system with a field angle of 45° without application of mydriasis (Keane et al., 2016). The images can be utilized for automated or manual AMD classification; however, there is no image-based AMD classification publicly available so far.

2.5 | AMD classification in UK Biobank derived from a machine learning algorithm and manually

We performed an automated AMD classification for 68,400 individuals with available fundus images in UK Biobank with additional manual classification in a subset of 2,013 participants, as described in Figure 1.

In epidemiological studies, AMD is usually classified per eye via manual grading of color fundus images by trained graders using established classification systems. One such system is the nine-step Age-Related Eye Disease Study (AREDS) severity scale (Davis et al., 2005), which defines early AMD combining a six-step drusen area scale with a five-step pigmentary abnormality scale and is therefore particularly detailed and time-consuming when applied manually. Another more recent system is the Three Continent AMD Consortium severity scale (3CC; Klein et al., 2014), which defines early AMD based on drusen size, drusen area, and the presence of pigmentary abnormalities and is thus more practical to apply manually. While the definition of "advanced AMD" is fairly robust across systems, each system defines "early" or "intermediate" AMD differently, but provides a clear assignment strategy to "no," "early/intermediate," or "advanced AMD" (or "no" and "any AMD").

To obtain an eye-specific AMD status for the 135,500 images of the UK Biobank (≤1 image per eye; 67,100 individuals with images for both eyes, 1,300 with image for only one eye), we applied a published convolutional neural network ensemble (Grassmann et al., 2018) to the fundus images following recommendations of the authors. The ensemble was trained to classify each image into the AREDS nine-step severity scale or three additional categories for advanced AMD (GA, NV, mixed GA + NV, "AREDS9 + 3 steps") or "ungradable." From this, we derived the person-specific automated AMD status as the AMD status of the worse eye (i.e., the higher score of the AREDS9 + 3) or as the status of the only eye, if applicable. We collapsed AREDS AMD severity steps 2–9 or any of the three advanced AMD categories to "any AMD."

To generate internal validation data, we selected a subset of UK Biobank individuals for additional manual grading. When randomly sampling participants, one would expect to catch only few AMD individuals; we thus selected (a) persons with high genetic risk score for AMD based on the known 52 variants for advanced AMD (Fritsche et al., 2016; >99th percentile, $n = 829$); (b) persons with low genetic risk score (<1st percentile, $n = 828$); and (c) persons with self-reported AMD not

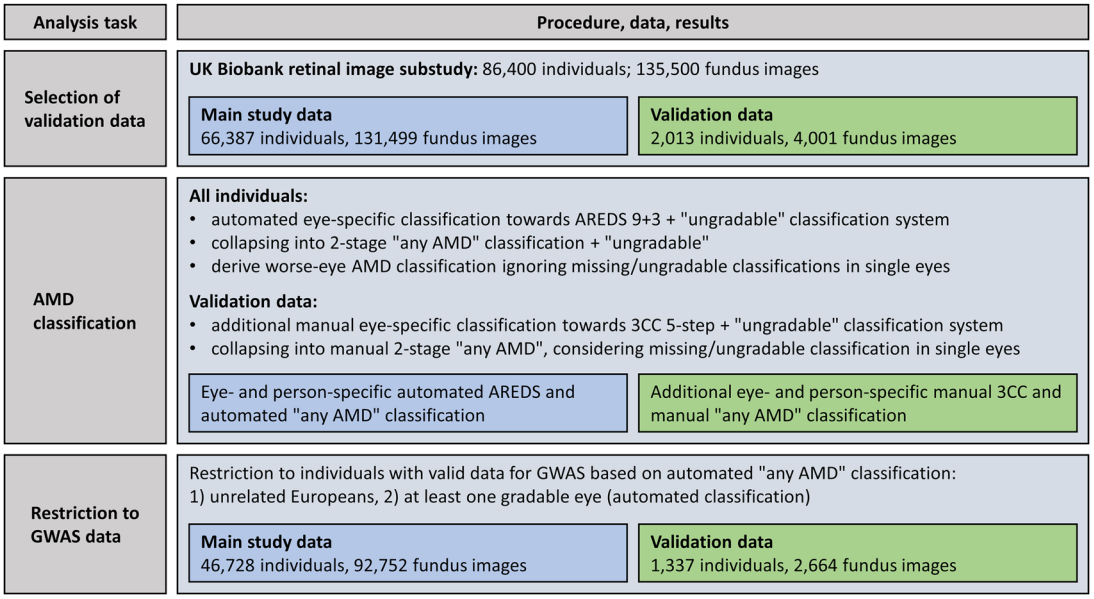

FIGURE 1 Schematic diagram of AMD classification and analyzed data. 3CC, Three Continent AMD Consortium severity scale; AMD, age-related macular degeneration; AREDS, Age-Related Eye Disease Study severity scale; GWAS, genome-wide association studies

already selected ($n = 356$). Results of the machine learning-based AMD classification were not used to select individuals into the validation subset and we can therefore validly estimate the algorithm's classification performance (sensitivity/specificity given the manual classification).

Each of the two eyes of the selected 2,013 individuals was manually classified for AMD according to the 3CC system (Klein et al., 2014) by a trained ophthalmologist (five AMD categories, 1 for no AMD, 3 for early, 1 for advanced AMD, and 1 "ungradable"). We collapsed the five AMD categories to "any AMD," "no AMD," or "ungradable" and derived a person-specific AMD status as the AMD status of the worse eye. Assuming neglectable misclassification in the eye-specific manually classified AMD status, this corresponds to the true person-specific AMD status if both eyes are manually gradable or one eye is manually ungradable and the second eye is manually graded as having AMD. If one eye is ungradable and the second, gradable eye is manually classified as "no AMD," the true person-specific disease status is unknown.

We derived eye-specific as well as person-specific confusion matrices based on the detailed (AREDS9 + 3 and five-category 3CC) and collapsed classifications. To conduct the GWAS with automatically derived "any AMD," we restricted the data with available automated AMD classification to unrelated individuals of European ancestry with valid GWAS data (see below), and derived the confusion matrices also for the restricted validation data.

2.6 | Genetic association analyses for AMD without and with accounting for misclassification

We performed a GWAS on the automatically derived "any AMD" versus "no AMD" in unrelated UK Biobank participants (relatedness status >3rd degree) of European ancestry (self-report "White," "British," "Irish," or "Any other White background") as recommended (Loh, Kichaev, Gazal, Schoech, & Price, 2018). For each variant, we applied standard logistic regression (i.e., the naïve analysis ignoring misclassification in the automatically derived AMD status) under the additive genotype model and applied a Wald-test as implemented in QUICKTEST (Kutalik et al., 2011). We included age and the first two genetic principal components as covariates. We excluded variants with low minor allele count (MAC < 400, calculated as MAC = $2 \times N \times$ MAF, sample size N, minor allele frequency MAF) or with low imputation quality (rsq < 0.4) yielding 11,567,158 analyzed variants. To correct for potential population stratification, we applied a Genomic Control correction ($\lambda = 1.01$ based on the analyzed variants excluding the 34 known AMD loci; Devlin, Roeder, & Devlin, 2013).

We selected genome-wide significant variants ($p_{GC} < 5.0 \times 10^{-8}$), clumped them into independent regions (≥500 kB between independent regions) and selected the variant with lowest p value in each region ("lead variant"). We also selected 21 of the 34 reported lead variants from the established advanced AMD loci, for which

we had ≥80% power to detect them in a UK Biobank sample size of 3,544 cases and 44,521 controls with Bonferroni-adjusted significance—under the assumption that the reported effect sizes for advanced AMD were the true effect sizes and ignoring any misclassification in the AMD classification (Appendix C). Information on linkage disequilibrium in Europeans was obtained from LDLink (Machiela & Chanock, 2015). Enrichment of directionally consistent or enrichment of nominally significant association for the 21 reported lead variants (when compared to the reported direction in literature) was tested based on the Exact Binomial test for H_0:Prob = .5 or H_0:Prob = .05, respectively.

To evaluate the robustness of the genetic association upon accounting for the misclassification, we applied the derived MLAs for the selected variants. For this, we modeled the conditional probability of AMD depending on age, genetic variant, and two genetic principal components (as in the naïve analysis). The MLAs accounted for the misclassification of the eye-specific automated classification and for the person-specific misclassification from missing AMD status in one of two eyes. For the misclassification process of the eye-specific automated classification (quantified by sensitivity and specificity), we allowed for a linear association with age and modeled two scenarios for the association with the genetic variant: (a) no association (nondifferential, MLA1) or (b) linear association (differential misclassification, MLA2). We compared association estimates of the naive analysis with MLA1- and MLA2-analysis and judged significance at Bonferroni-corrected significance levels for a family-wise error rate of 0.05. To allow for comparisons across different models, we did not apply genomic control correction for these comparative analyses. In addition, we evaluated the robustness of findings from the naïve analysis for the selected lead variants upon adjusting for 20 instead of 2 genetic principal components.

To follow-up on the *HERC2* lead variant finding (see Section 3), we quantified lightness of fundus images by calculating gray levels for the "RGB" fundus images (weighted sum of R, G, and B values, $0.30 \times R + 0.59 \times G + 0.11 \times B$, as implemented in IrfanView).

3 | RESULTS

3.1 | Linking misclassification theory to machine learning disease classification

We here establish the usage of machine learning-derived disease classification in genetic association analyses as a response misclassification problem in logistic regression (see Section 2). We present a newly developed maximum likelihood approach (MLA) for *bilateral diseases* like AMD (see Section 2). This includes two versions: (a) assuming *nondifferential misclassification* (MLA1, i.e., no dependency of misclassification probabilities on the covariate of interest, here the genetic variant) and (b) allowing for *differential misclassification* (MLA2, i.e., dependency on the covariate of interest). There are existing MLAs for considering response misclassification in logistic regression using internal validation data (Carroll et al., 2006; Lyles et al., 2011): these MLAs refer to *classic diseases* where the misclassification is on the person-specific disease status. Our developed approach provides a general framework for bilateral diseases with entity-specific misclassification that propagates to person-specific disease misclassification. Our approach also allows for missing classification in one of two entities, which is a second source of bias in association analyses for bilateral diseases as reported previously (Günther et al., 2019). We exemplify our approach on machine learning-derived AMD compared to manually graded AMD. Since machine learning algorithms for AMD are trained on images with human manual AMD grading as benchmark, we assume the manual classification to be gold standard.

We evaluated the performance of the naïve analysis and our developed MLA1 and MLA2 in a simulation study with different misclassification scenarios. By this, we documented substantial bias when the naïve analysis was applied to misclassified data, which was comparable to the theory for classic (nonbilateral) diseases (Carroll et al., 2006; Neuhaus, 1999). Naïve association estimates were biased toward zero in case of nondifferential misclassification and in any direction in case of differential misclassification. In the latter scenario, we observed a lack of type I error control for the naïve analysis. Furthermore, we showed our MLA1 and MLA2 to effectively remove bias and keep type I error when specified correctly (Tables 1 and S1 and Appendix D). In case of differential misclassification, MLA1 (assuming nondifferential misclassification) yields biased estimates and a lack of type I error control as well, comparable to the naïve analysis.

3.2 | AMD in UK Biobank based on automated classification and validation data

We applied a published convolutional neural network ensemble (Grassmann et al., 2018) to automatically derive eye- and person-specific AMD classifications for 68,400 UK Biobank participants with fundus images at

TABLE 1 Simulation results on effect estimates and empirical type I error in naïve and MLA-analysis

Simulation scenario						$\hat{\beta}_Y$						Percent with $p < .05$			Cov. Freq.		
Sens.	Spec.	Percent miss.	β_Y	β_{sens}	β_{spec}	Naïve		MLA1		MLA2		Naive	MLA1	MLA2	Naive	MLA1	MLA2
						Mean	RMSE	Mean	RMSE	Mean	RMSE						
Nondifferential misclassification																	
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%
Differential misclassification																	
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%

Note: We evaluated the performance of naïve and MLA analysis of a quantitative covariate X and a binary bilateral disease Y, for example, person-specific AMD, simulating various scenarios. For each scenario, we sampled 1,000 datasets à 5,000 individuals, 4,000 with only error-prone eye-specific AMD classification, and 1,000 with additional true AMD classification. Shown are performance measures from three models, naïve analysis, MLA1, or MLA2 assuming nondifferential/differential misclassification regarding X, respectively, in various simulation scenarios. For the eight scenarios shown here, we assumed no association of X with δ, the probability of AMD in both eyes given ≥1 affected eye; results were similar when modeling an association of *X* with δ, see Table S1. For each model and scenario, we report mean effect estimates $\hat{\beta}_Y$, 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 $p < .05$), and coverage frequencies of 95% CI. %miss., fraction of randomly selected individuals with missing AMD classification in one of two eyes; sens/spec, average sensitivity and specificity of error-prone, eye-specific AMD classification; β_Y, log OR of *X* on true AMD; β_{sens}, log OR of *X* on the sensitivity; β_{spec}, log OR of *X* on the specificity of the eye-specific misclassification process.

Abbreviations: AMD, age-related macular degeneration; MLA, maximum likelihood approach; OR, odds ratio; RMSE, root mean square error.

TABLE 2 Confusion matrices comparing manual and automated AMD classification per eye and per person

(a) Per eye (4,001 eyes, 2,013 individuals)				
	Automated classification			
Manual	Ungradable	No AMD	Any AMD	Sum
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 (1,337 individuals)				
	Automated classification			
Manual classification		No AMD	Any AMD	Sum
Ungradable/NA[a] (NA)		210 (80%)	53 (20%)	263 (100%)
No AMD		750 (91%)	72 (9%)	822 (100%)
Any AMD		58 (23%)	194 (77%)	252 (100%)

Note: Shown are absolute numbers and conditional classification probabilities, that is, in row i and column j,P(automated = j | manual = i) as %, with i, j = "Ungradable," "No AMD," "Any AMD": (a) for all eyes in the validation data; 4,001 eyes of 2,013 individuals. (b) For all individuals in the overlap between validation data and GWAS; 1,337 individuals, all gradable with automated classification.
Abbreviations: AMD, age-related macular degeneration; GWAS, genome-wide association studies.
[a]NA, true AMD status based on worse eye not available, since one eye was manually ungradable and the second AMD-free.

baseline (135,000 eyes; Table S2a). From this, we derived eye-specific "any AMD" status (i.e., any early AMD stage or advanced AMD versus AMD-free) and person-specific "any AMD" status based on the worse eye (see Section 2). Among the 68,400 participants, 10,128 were ungradable for AMD in both eyes by the automated classification (i.e., missing person-specific AMD status by the automated classification, 14.8%), 4,870 were classified as "any AMD" and 53,402 as AMD-free (Table S2b). Among the 58,272 automatically gradable participants (of these: 20.2% automatically gradable only in one eye), 8.4% had AMD and 91.6% were AMD-free. This included 48,065 unrelated individuals of European ancestry with GWAS data (3,544 "any AMD" cases, 44,521 AMD-free controls; 19.8% with only one eye gradable; Table S2b).

To quantify the performance of automated AMD classification, we manually classified AMD in a subset as internal validation data (4,001 images, ≤1 image per eye, 2,013 individuals). When comparing automated to manual (true) "any AMD" status, we found an eye-specific sensitivity of 73% and specificity of 90% in the full validation data and a person-specific sensitivity of 77% and specificity of 91% among the participants in the GWAS (Table 2). We found no structural differences between the full validation data and when restricting to the GWAS data (1,337 individuals, Table S3a,b). Both, the manual and automated classification included the category "ungradable." Among the 4,001 eyes, 1,101 were manually ungradable, of which the automatic classification yielded 74% as ungradable as well, but classified 9% as AMD and 17% as AMD-free, which raises concerns about these classifications. In summary, we found the automated classification to yield reasonable, but error-prone results.

3.3 | GWAS on automated AMD classification in naïve analysis identifies two loci

While we have some idea about the extent of the misclassification from validation data and about its impact on genetic association estimates from simulations, it is unclear whether the automated any AMD classification is "good enough" for GWAS. We conducted a GWAS for person-specific automatically derived "any AMD" in UK Biobank (3,544 "any AMD" cases; 44,521 controls) applying logistic regression as usual, which is without accounting for misclassification (naïve analysis). We found 53 variants with genome-wide significance ($p_{GC} < 5.0 \times 10^{-8}$) spread across two distinct loci (defined as lead variant and proxies +/− 500 kB, Figure 2a,b; Table S4a): the known *ARMS2/HTRA1* locus (lead variant here rs370974631, $p_{GC} = 3.1 \times 10^{-20}$, effect allele frequency [EAF] = 0.23) and an unknown locus for AMD near *HERC2* (lead variant rs12913832, $p_{GC} = 4.7 \times 10^{-16}$, EAF = 0.23). This *ARMS2/HTRA1* lead variant was highly correlated to the reported lead variant for advanced AMD, rs3750846, and effect estimates were

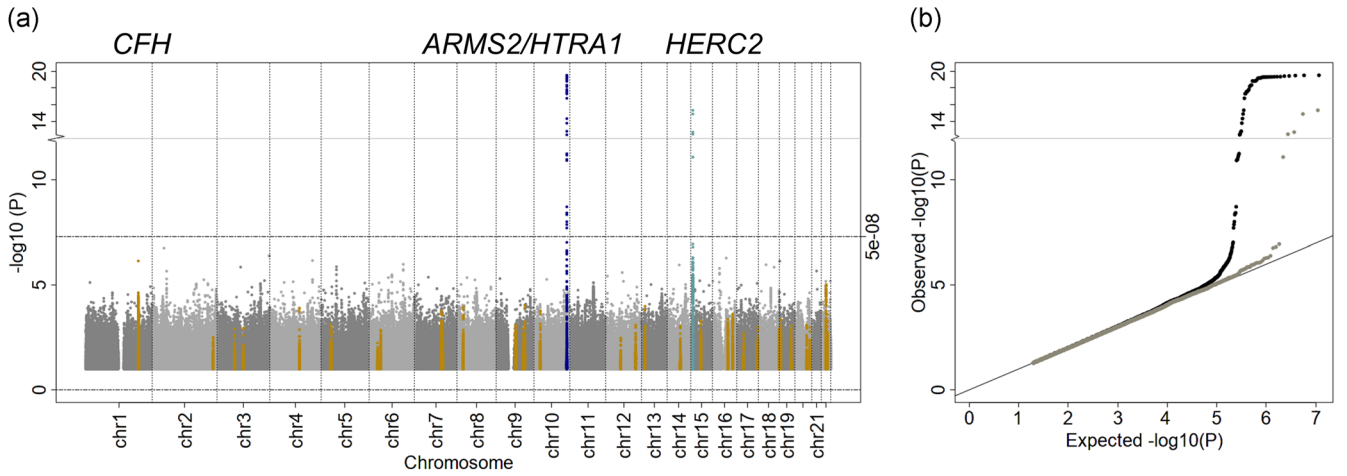

FIGURE 2 GWAS results in UK Biobank based on automatically derived "any AMD" from naïve analysis. Association analyses were conducted using the error-prone, machine learning-derived AMD classification in UK Biobank participants with 3,544 "any AMD" cases and 44,521 controls via logistic regression adjusted 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 (Fritsche et al., 2016) locus, light blue: unknown genome-wide significant locus, orange: other 33 previously established loci for advanced AMD), and (b) expected versus observed −log10 *p* values; black: all variants, gray: all variants outside the 34 previously reported loci. 3CC, Three Continent AMD Consortium; AMD, age-related macular degeneration; GWAS, genome-wide association studie

directionally consistent ($r^2 = .93$; Table S4b). The next best known locus is the *CFH* locus, which showed close to genome-wide significance here (smallest *p* value $p_{GC} = 7.0 \times 10^{-7}$, rs6695321, EAF = 0.62): rs6695321 is in linkage disequilibrium with two reported *CFH* variants (rs61818925, rs570618: $r^2 = .63$ or $r^2 = .40$, D′ = 0.81 or D′ = 1.00, EAF = 0.58 or 0.36, respectively; Table S4b) suggesting that rs6695321 captures the signals of these two reported variants.

Among the reported lead variants of the 34 advanced AMD loci (Fritsche et al., 2016), we had ≥80% power to detect 21 of these with Bonferroni-adjusted significance (Table S5). When comparing effect sizes of these 21 variants from this analysis on "any AMD" in UK Biobank with reported effect sizes for advanced AMD, we found 15 with directional consistency ($p_{Bin} = 0.078$) and 7 with directionally consistent nominal significance ($p_{Bin} = 4.9 \times 10^{-5}$; Figure 4a and Table S4c). The overall smaller effect sizes for automated "any AMD" compared to reported effect sizes for advanced AMD can be explained by a bias from misclassified automated AMD and by smaller effect sizes for early AMD merged into the definition of "any AMD." For the other 13 of the 34 variants, we refrained from interpreting results due to lack of power in this analysis (Table S4c). Results were similar when adjusting for 20 instead of 2 genetic principal components (data not shown). While the yield of only few known AMD signals in this UK Biobank GWAS may be disappointing, this is not fully unexpected given an effective sample size (Ma, Blackwell, Boehnke, Scott, & GoT2D investigators, 2013) of 13,130 and a power estimate of ~80% (assuming no misclassification and reported effect sizes) to detect associations with genome-wide significance for only 6 of the 34 established variants (*CFH*, *C2/CFB/SKIV2L, ARMS2/HTRA1*, *C3, APOE, SYN3/TIMP3*; Table S5).

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

3.4 | Applying the developed MLA to account for misclassification for selected variants

Due to our simulation results and theory (Carroll et al., 2006; Neuhaus, 1999), we expected our GWAS on automated (error-prone) AMD to yield biased estimates and, when the misclassification was differential toward the genetic variant, even potentially false signals. We applied our developed MLAs for 26 selected variants: (a) the three lead variants detected here with (near) genome-wide significance (*CFH*: rs6695321, *ARMS2/HTRA1:* rs370974631, *HERC2:* rs12913832), (b) the three reported independent variants in the *CFH* locus with MAF ≥ 5% (rs61818925, rs570618, rs10922109; two of these

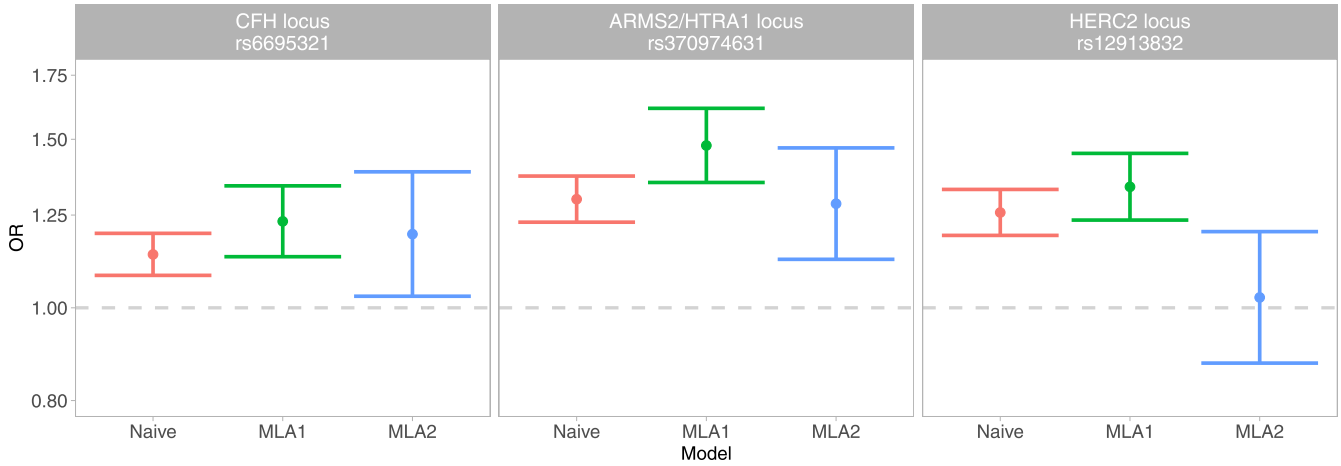

FIGURE 3 Genetic effect estimates for the three lead variants in UK Biobank without and with accounting for misclassification. Shown are genetic effect estimates (odds ratios [OR]) and 95% confidence intervals for three lead variants from the GWAS on automated AMD classification with 3,544 "any AMD" cases and 44,521 controls from three models: without accounting for the misclassification; *naïve analysis*, red. With accounting for nondifferential misclassification, that is, no dependency on the genetic variant; MLA1, green. And accounting for a differential misclassification, that is, dependency on the genetic variant; MLA2, blue. Both MLAs accounted for missing AMD information in one of two eyes and a misclassification associated with age. *Y*-axis is on log-scale. AMD, age-related macular degeneration; GWAS, genome-wide association studies; MLA, maximum likelihood approach

correlated to the here identified *CFH* lead variant), and (c) the other 20 of the 34 reported lead variants (Fritsche et al., 2016), for which we had reasonable power in this analysis (including 1 reported *ARMS2/HTRA1* variant correlated to here identified variant). This yielded a total of ~23 independent variants.

Our MLAs estimated simultaneously (a) sensitivity and specificity of the eye-specific misclassification process and (b) genetic association accounting for the misclassification. With regard to sensitivity and specificity, we found (a) an overall sensitivity of 64.5% (95% CI: 60.1%, 68.7%) and a specificity of 98.6% (98.4%, 98.8%), that is, a false-negative "any AMD" proportion of 35.5% and a false-positive of 1.4%; (b) few evidence for an association of the sensitivity with any selected variant ($p > .05/(23 \times 2) = 1.09 \times 10^{-3}$) and no association with the specificity, except for two variants: *HERC2* lead variant, rs12913832, and the reported *CFH* lead variant rs10922109 ($OR_{spec} = 0.64$, $p_{spec} = 7.38 \times 10^{-9}$ and $OR_{spec} = 1.36$, $p_{spec} = 2.29 \times 10^{-4}$, respectively; Table S6 and

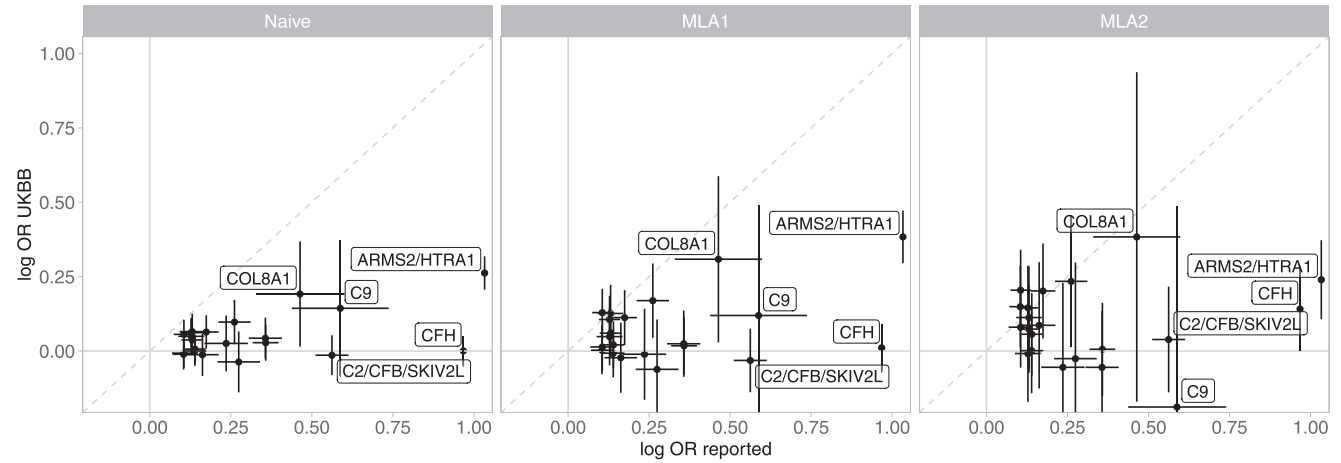

FIGURE 4 Comparison of 21 reported genetic effect estimates for advanced AMD with estimates for automatically derived "any AMD" from UK Biobank without and with accounting for misclassification. We selected the 21 reported AMD lead variants, for which we had ≥80% power to detect them in this UK Biobank sample size with Bonferroni-adjusted 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 the naïve analysis (logistic regression ignoring misclassification), MLA1, and MLA2. AMD, age-related macular degeneration; MLA, maximum likelihood approach; OR, odds ratio

Appendix E). Therefore, we found a misclassification that was associated with some genetic variants (differential), which could induce bias into either direction as well as a severe lack of type I error control.

When comparing genetic association estimates from our MLA1 and MLA2 with the naïve analysis for our three detected lead variants, we found interesting patterns (Figure 3 and Table S7a). (a) For *CFH* and *ARMS2/HTRA1*, we found consistent effect estimates across the three analyses, with larger confidence intervals when using the more complex models MLA1 or MLA2. (b) For *HERC2*, MLA1 yielded comparable results to the naïve analysis, but when accounting for differential misclassification (MLA2), the effect vanished (MLA2: OR = 1.03, $p = .76$; MLA1: OR = 1.34, $p = 1.11 \times 10^{-12}$; naïve: OR = 1.26, $p = 4.16 \times 10^{-16}$). The results of MLA1 for this variant were as expected, since a model considering nondifferential misclassification leads in general, by assumption, to larger estimates and widened confidence intervals if any misclassification is present.

When applying MLA1 and MLA2 to the three reported *CFH* locus variants and the further 20 of the 34 reported lead variants, we found the following (Table S7b,c): (a) Effect estimates for all three *CFH* variants increased when applying MLA2 compared to the naïve analysis. This was particularly interesting for the reported *CFH* lead variant rs10922109, where we now observed a nominally significant association into the reported direction (MLA2: OR = 1.15, $p = .047$; naïve: OR = 1.00, $p = .98$; Table S7c). This is in line with the observed association of the specificity and this *CFH* variant. (b) For the other 20 reported lead variants, many variants showed increased effect estimates by MLA2 compared to the naïve analysis (effect estimates mostly more comparable to reported effect sizes; Fritsche et al., 2016; Figure 4c). Altogether, MLA results confirmed the *CFH* and *ARMS2/HTRA1* loci and unmasked the *HERC2* finding as false positive.

3.5 | Misclassification depended on eye and fundus image color

Interestingly, our *HERC2* lead variant, rs12913832, is precisely the variant for which the G allele was considered causal for blue eyes (Sturm et al., 2008). We were able to support this in our AugUR (Brandl et al., 2018; Stark et al., 2015) study ($n = 1026$; reported "light eye color" for 14%, 36%, or 97% of participants with A/A, G/A, or G/G, respectively). Eye color is discussed as AMD risk factor, but the debate is on blue eyes to increase risk due to increased susceptibility to UV-radiation (Chakravarthy et al., 2010), which is in contrast to our observation of brown eyes to increase AMD risk and a challenge for interpreting this finding. It was interesting to see the *HERC2* rs12913832 association vanish when accounting for rs12913832-associated misclassification. This was in line with the observed strong association of the specificity with this variant ($OR_{spec} = 0.64$ per A allele; Table S6a) resulting in 3.0%, 1.9%, or 1.2% of false-positive AMD classifications among persons with A/A, A/G, or G/G, respectively. This notion of a larger misclassification among A/A versus G/G individuals was further supported by the larger fraction of manually ungradable images that were deemed gradable by the automatic classification among A/A versus G/G (54.5% vs. 38.8%, respectively; Figure 5). When visually inspecting fundus images per genotype group, the images for A/A had a darker appearance than those for A/G or G/G (Figure 5), which we were able to quantify by means of average gray level per image of 46.4, 49.0, or 53.6, respectively. Therefore, the *HERC2* signal appeared to be an artifact due to a larger misclassification for brown eyes linked to darker fundus images. One may hypothesize that the darker eye color had reduced light exposure during fundus photography, which gave rise to darker images and more misclassified AMD-free eyes. The notion of a differential misclassification due to eye color was further supported by the fact that the full *HERC2* signal disappeared by modeling a misclassification dependency on the causal variant for eye color (rs12913832; Figure S1a,b), while some signal remained when modeling a misclassification dependency on the respective *HERC2* variant in the model (Figure S1c). In summary, we found the MLA2 not only to effectively remove the artifact signal of the naïve GWAS, but also to help understand the dependencies of the misclassification.

4 | DISCUSSION

GWAS on machine learning-derived classification of imaging-based diseases, like AMD, can be expected to accelerate knowledge gain and drug target development (Nelson et al., 2015), since it will enable substantially increased sample sizes and refined, homogeneous phenotyping. To this date, there was no GWAS reported using a machine learning-derived classification for AMD or any other imaging-based disease—to the best of our knowledge. We here present a GWAS on machine learning-derived AMD in UK Biobank highlighting chances and challenges. By this GWAS on AMD combined with an evaluation of emerging genetic signals via our newly developed MLA, we were able to detect known AMD loci and to distinguish true loci from artifacts.

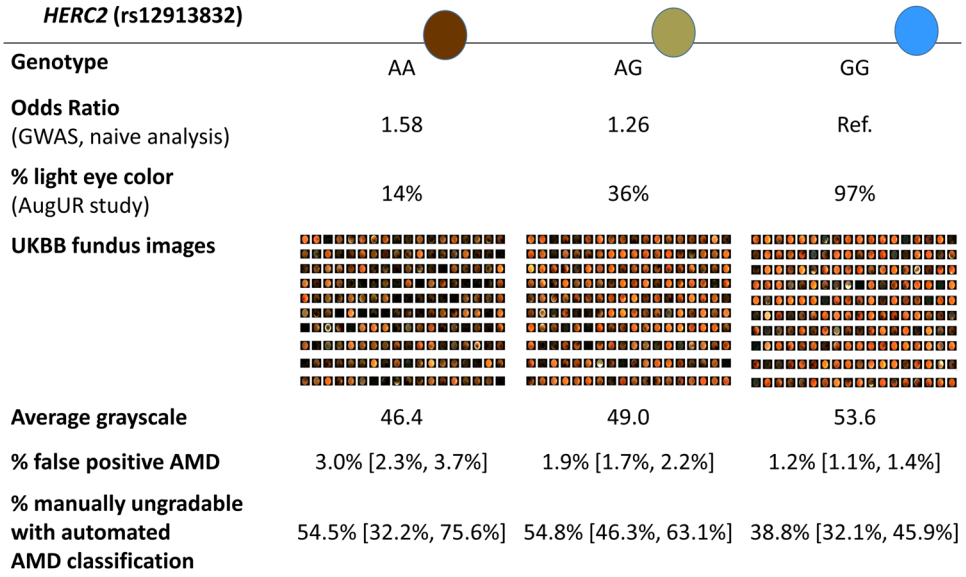

FIGURE 5 Evidence for differential misclassification in automatically derived AMD with respect to the *HERC2* variant rs12913832. Shown are (a) estimated odds ratios from the naïve analysis ignoring misclassification and various characteristics per genotype group; (b) the fraction of persons with self-reported "light eye color" in the AugUR study; (c) randomly selected fundus images in UK Biobank; (d) image-lightness quantified by mean average grayscale; (e) proportion of false-positive AMD in the automated classification (1-specificity) and 95% confidence intervals estimated via MLA2; and (f) observed proportion of manually ungradable images that were deemed gradable by the algorithm and classified as "any AMD" or "AMD-free." AMD, age-related macular degeneration; GWAS, genome-wide association study; MLA, maximum likelihood approach

Such artifacts, that is, false positives, can derive from misclassification that is associated with a genetic variant. Our data and analyses provide a compelling example for such an artifact: our MLA revealed the *HERC2* signal as false-positive signal and suggested darker eye color and darker fundus images as a relevant source of misclassification for this machine learning algorithm. It is perceivable that the misclassification process of other algorithms for AMD and for other image-based diseases will depend on one or the other characteristic as well, and that such a characteristic is picked up by some genetic variants due to the abundant range of genetically pinpointed characteristics (see, e.g., NHGRI-EBI GWAS Catalog; Buniello et al., 2019), which can yield artifact signals when left unaccounted.

Our MLA, developed for bilateral diseases, does not only quantify the misclassification and the dependencies, but also guards against bias and artifacts in association analyses. Our approach has certain limitations: since we use statistical modeling for the error-prone classification, the analysis is only valid if the corresponding assumptions hold. This concerns independence of entity-specific classification given the true disease status, the correct specification of the misclassification model based on the validation data, and a neglectable error in the gold standard classification. Similar approaches are available for classic diseases (Carroll et al., 2006; Lyles et al., 2011). Thus, this concept can be generalized to other algorithms and other image-based diseases. Our work here links the theory of misclassification to machine learning-derived disease classification, which can be generalized also to measurement error and quantitative phenotypes.

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

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

ACKNOWLEDGMENTS

This study was supported by the DFG HE 3690/5-1 (to I. M. H.) and NIH R01 EY RES 511967 (to I. M. H.), the University of Regensburg and the Ludwig Maximilians University Munich. The UK Biobank (accessed via application number 33999) was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government, and the Northwest Regional Development Agency. This study was also supported by the Welsh Assembly Government, British Heart Foundation and Diabetes UK. The authors would also like to thank the two anonymous reviewers whose comments helped improve and clarify this manuscript. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

Data that support the findings of this study are available as UK Biobank resource (accessed via application number 33999). The fundus-image derived AMD classifications will be returned to UK Biobank and can be accessed by other researchers via the Data Showcase. An open source R implementation of the developed maximum likelihood approach to account for misclassification in bilateral disease is available at: https://www.genepi-regensburg.de/MLA-bilateral/. The convolutional neural network ensemble used for automated AMD classification and recommendations by the authors can be found at: https://github.com/RegensburgMedicalImageComputing/ARIANNA. IrfanView: https://www.irfanview.com/; GWAS catalogue: https://www.ebi.ac.uk/gwas/.

ORCID

Felix Guenther http://orcid.org/0000-0001-6582-1174
Caroline Brandl https://orcid.org/0000-0001-8223-6137
Thomas W. Winkler https://orcid.org/0000-0003-0292-5421
Klaus Stark https://orcid.org/0000-0002-7832-1942
Helmut Kuechenhoff https://orcid.org/0000-0002-6372-2487

REFERENCES

Brandl, C., Zimmermann, M. E., Günther, F., Barth, T., Olden, M., Schelter, S. C., ... Heid, I. M. (2018). On the impact of different approaches to classify age-related macular degeneration: Results from the German AugUR study. *Scientific Reports*, *8*(1), 8675. https://doi.org/10.1038/s41598-018-26629-5

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burlina, P. M., Joshi, N., Pekala, M., Pacheco, K. D., Freund, D. E., & Bressler, N. M. (2017). Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmology*, *135*(11), 1170. https://doi.org/10.1001/jamaophthalmol.2017.3782

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Chakravarthy, U., Wong, T. Y., Fletcher, A., Piault, E., Evans, C., Zlateva, G., ... Mitchell, P. (2010). Clinical risk factors for age-related macular degeneration: A systematic review and meta-analysis. *BMC Ophthalmology*, *10*(1), 31. https://doi.org/10.1186/1471-2415-10-31

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Routledge. https://doi.org/10.4324/9780203771587

Csurka, G. (2017). A comprehensive survey on domain adaptation for visual applications, *Domain Adaptation in Computer Vision Applications* (pp. 1–35). Cham, Switzerland: Springer.

Davis, M. D., Gangnon, R. E., Lee, L.-Y., Hubbard, L. D., Klein, B. E. K., & Klein, R., ... Age-Related Eye Disease Study Group. (2005). The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17. *Archives of Ophthalmology*, *123*(11), 1484–1498. https://doi.org/10.1001/archopht.123.11.1484

Devlin, A. B., Roeder, K., & Devlin, B. (2013). Genomic control for association. *Biometrics*, *55*(4), 997–1004.

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., ... Heid, I. M. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, *48*(2), 134–143. https://doi.org/10.1038/ng.3448

Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M. E., Linkohr, B., ... Weber, B. H. F. (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, *125*(9), 1410–1420. https://doi.org/10.1016/j.ophtha.2018.02.037

Günther, F., Brandl, C., Heid, I. M., & Küchenhoff, H. (2019). Response misclassification in studies on bilateral diseases. *Biometrical Journal*, *61*(4), 1033–1048. https://doi.org/10.1002/bimj.201900039

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, *87*(2), 239–269. https://doi.org/10.1016/S0304-4076(98)00015-3

Heinze-Deml, C., & Meinshausen, N. (2017). Conditional variance penalties and domain shift robustness. arXiv:1710.11469 [stat.ML].

Keane, P. A., Grossi, C. M., Foster, P. J., Yang, Q., Reisman, C. A., & Chan, K., ... UK Biobank Eye Vision Consortium. (2016). Optical coherence tomography in the UK Biobank study–Rapid

automated analysis of retinal thickness for large population-based studies. *PLoS One*, *11*(10), e0164095. https://doi.org/10.1371/journal.pone.0164095

Klein, R., Meuer, S. M., Myers, C. E., Buitendijk, G. H. S., Rochtchina, E., Choudhury, F., ... Klein, B. E. K. (2014). Harmonizing the classification of age-related macular degeneration in the three-continent AMD Consortium. *Ophthalmic Epidemiology*, *21*(1), 14–23. https://doi.org/10.3109/09286586.2013.867512

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., ... Bergmann, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics*, *12*(1), 1–17. https://doi.org/10.1093/biostatistics/kxq039

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*(1995), 60–88. https://doi.org/10.1016/j.media.2017.07.005

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics*, *50*(7), 906–908. https://doi.org/10.1038/s41588-018-0144-6

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., & Sobel, J. D. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. *Epidemiology*, *22*(4), 589–597. https://doi.org/10.1097/EDE.0b013e3182117c85

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., & GoT2D Investigators (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology*, *37*(6), 539–550. https://doi.org/10.1002/gepi.21742

Machiela, M. J., & Chanock, S. J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, *31*(21), 3555–3557. https://doi.org/10.1093/bioinformatics/btv402

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., & Teumer, A., ... Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., ... Sanseau, P. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, *47*(8), 856–860. https://doi.org/10.1038/ng.3314

Neuhaus, J. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, *86*(4), 843–855. https://doi.org/10.1093/biomet/86.4.843

Peng, Y., Dharssi, S., Chen, Q., Keenan, T. D., Agrón, E., Wong, W. T., ... Lu, Z. (2019). DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, *126*(4), 565–575. https://doi.org/10.1016/j.ophtha.2018.11.015

R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from https://www.r-project.org/

Stark, K., Olden, M., Brandl, C., Dietl, A., Zimmermann, M. E., Schelter, S. C., ... Heid, I. M. (2015). The German AugUR study: Study protocol of a prospective study to investigate chronic diseases in the elderly. *BMC Geriatrics*, *15*(1), 130. https://doi.org/10.1186/s12877-015-0122-0

Stephane, C. (2018). pwr: Basic functions for power analysis. Retrieved from https://cran.r-project.org/package=pwr

Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., ... Montgomery, G. W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics*, *82*(2), 424–431. https://doi.org/10.1016/j.ajhg.2007.11.005

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., ... Wong, T. Y. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Journal of the American Medical Association*, *318*(22), 2211–2223. https://doi.org/10.1001/jama.2017.18152

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., ... Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–89. https://doi.org/10.1038/nature14962

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Guenther F, Brandl C, Winkler TW, et al. Chances and challenges of machine learning-based disease classification in genetic association studies illustrated on age-related macular degeneration. *Genetic Epidemiology*. 2020;44:759–777. https://doi.org/10.1002/gepi.22336

APPENDIX A: MLA TO ADJUST FOR RESPONSE MISCLASSIFICATION IN BILATERAL DISEASES

We developed a maximum likelihood approach (MLA) to adjust for response misclassification from an error-prone, entity-specific disease classification in bilateral diseases. Here, we illustrate it based on the example of age-related macular degeneration, where AMD can occur in each eye (eye-specific AMD) and the person-specific binary outcome is defined as worse eye outcome, that is, "AMD in at least one eye," and modeled using logistic regression. We assume that we have an error-prone, eye-specific AMD classification (e.g., from a machine learning-based

automated classification) available for nearly all eyes and true, gold-standard classifications (e.g., manual classification) for a subset of individuals from validation data.

Let $(Z_{1i}, Z_{2i}) \in \{0, 1\}$ be the true, binary disease stages in the two eyes of study participant i, that is, $(Z_{1i} = 1, Z_{2i} = 0)$ means that participant i suffers from AMD in the left eye and is unaffected from AMD in the right. When estimating the association of person-specific risk factors with AMD, one often defines a binary person-specific disease status as worse eye AMD,$Y_i := \max(Z_{1i}, Z_{2i})$, $Z_{1i}, Z_{2i} \in \{0, 1\}$, and uses logistic regression to estimate the association of some covariates X with AMD: the person-specific disease status Y_i equals 1, if at least one eye of individual i is classified as AMD, and Y_i equals 0, if both eyes are unaffected. As described previously (Günther et al., 2019), such a worse eye disease status can be misclassified because of two reasons: either, because of missing disease information in one of two eyes (in this case disease can be overlooked), or because of error-prone disease status for any of the two eyes. Here, we assume that we observed an error-prone, eye-specific disease status (Z_{1i}^*, Z_{2i}^*) for each of the two eyes of a "main study" participant i and additionally the true disease status in each of the two eyes (Z_{1j}, Z_{2j}) for a subset of study participants j from the "validation study." For all participants from the main study (error-prone classifications only) or the validation subset (error-prone and true classification), there is the additional issue that the disease information can be missing in one of two eyes, because of missing or ungradable fundus images. Since the automated (error-prone) and manual (gold standard, "true") classification may judge differently on whether an image is gradable or ungradable, any possible subset of $(Z_{1i}, Z_{2i}, Z_{1i}^*, Z_{2i}^*)$ might be the available information for a specific study participant. To obtain valid estimates for the association of covariates with the true AMD status, we set up a likelihood based on the conditional probabilities of the observed error-prone and/or true eye-specific disease classifications given covariates. The product of these conditional probabilities over all individuals forms the likelihood, which has to be numerically optimized with respect to the regression parameters to obtain estimates. The different likelihood contributions for the individuals depend on the available AMD classifications (true and/or error-prone for one or both eyes).

The general problem of response misclassification when AMD information is missing in one of two eyes and/or the eye-specific classification suffers from misclassification with known classification probabilities has already been evaluated in a previous publication (Günther et al., 2019). There, we also derived the corresponding likelihood contributions for the different scenarios of available outcome data. Here, we add the aspect that validation data are available for some study participants or, more specifically, a collection of error-free (gold-standard) classified single eyes, and that we model the eye-specific misclassification process based on information from this validation data.

In the following, we describe the general idea and provide formulas for the respective likelihood contributions.

The assumed logistic regression model for the true worse eye disease corresponds to the assumption that $\max(Z_{1i}, Z_{2i}) = Y_i \sim \text{Bernoulli}(\pi_i)$, where we model the success probability based 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-specific covariates and β the vector of corresponding regression coefficients. It follows that $P(Y_i = 1|x_i) = \pi_i$. If we focus on single-eye disease classifications, there exist four different pattern of true disease classifications (Z_{1i}, Z_{2i}): (1, 1), (1, 0),(0, 1), (0, 0). From the assumed logistic 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 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 we define the person-specific conditional probability of being affected by AMD in both eyes given AMD in at least one eye as $\delta_i := P(Z_{1i} = 1, Z_{2i} = 1|x_i, Y_i = 1)$. When assuming symmetric probabilities for disease in one but not the other eye for left and right eyes (i.e., same probabilities to be affected in the left but not the right eye and vice versa), the conditional probability mass function of the two-entity disease status distribution can be written concisely as

$P(\cdot,\cdot\|x_i)$	$Z_{2i} = 1$	$Z_{2i} = 0$
$Z_{1i} = 1$	$\delta_i\pi_i$	$\frac{1-\delta_i}{2}\pi_i$
$Z_{1i} = 0$	$\frac{1-\delta_i}{2}\pi_i$	$1 - \pi_i$

(1)

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

$$\begin{aligned} P(Z_{1i} = 1|x_i) &= P(Z_{1i} = 1, Z_{2i} = 1|x_i) \\ &+ P(Z_{1i} = 1, Z_{2i} = 0|x_i) = \left(\frac{1}{2} + \frac{1}{2}\delta_i\right)\pi_i. \end{aligned} \quad (2)$$

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

$$\begin{aligned} P(Z_{li}^* = 1|Z_{li} = 1, x_i) &= \pi_{1i} \\ P(Z_{li}^* = 0|Z_{li} = 0, x_i) &= \pi_{0i} \end{aligned} \quad (3)$$

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

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

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

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

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

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

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

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

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

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

When three classifications were observed, for example, $(Z_{1i} = z_{1i}, Z_{1i}^* = z_{1i}^*, Z_{2i}^* = z_{2i}^*)$ or $(Z_{1i} = z_{1i}, Z_{2i} = z_{2i}, Z_{1i}^* = z_{1i}^*)$, we can derive

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

and

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

All conditional probabilities characterizing the *true data model* and the *misclassification process*, that is, (a) the probability of true worse eye AMD $P(Y_i = 1|x_i) = \pi_i$, (b) the probability of AMD in both eyes given AMD in at least one eye $P(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, x_i) = \delta_i$, (c) the eye-specific sensitivity $P(Z_{1i}^* = 1|Z_{1i} = 1, x_i) = \pi_{1i}$, and (d) the eye-specific specificity of the error-prone classification $P(Z_{1i}^* = 0|Z_{1i} = 0, x_i) = \pi_{0i}$, can potentially vary with person-specific characteristics. We therefore decided to model them based on the logistic function of a linear predictor, where relevant covariates can be specified for each probability. Combining all these expressions, we can set up the whole likelihood based on the derived conditional probabilities and numerically optimize with respect to the regression coefficients of the linear predictors for π_i, δ_i, π_{1i}, and π_{0i}. Standard errors of the maximum likelihood estimates are derived based on standard likelihood theory from the square root of the diagonal elements of the inverse of the observed Fisher information (Hessian) and used for inference. An implementation of the MLA in the statistical programming language R (R Core Team, 2019) is available.

APPENDIX B: SIMULATION STUDY TO EVALUATE CONSEQUENCES OF IGNORING MISCLASSIFICATION AND THE PERFORMANCE OF THE MLA IN CORRECTING IT

We performed a simulation study to evaluate the consequences of ignoring response misclassification and to evaluate the performance of the derived MLA in data scenarios similar to the situations in AMD studies. For each simulation scenario (data generating process), we simulated 1,000 datasets, applied different models to the sampled data, and evaluated the distribution of effect estimates, frequencies of significant statistical tests, and coverage frequencies of confidence intervals for a central covariate of interest.

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

1. We sampled the true binary "worse-eye" AMD data Y for 5,000 individuals by sampling from a Bernoulli distribution, where we modeled the success probability based on the logistic function of a linear predictor (corresponding to the assumed data generating process in logistic regression). For the linear predictor, we used an intercept of −0.25 (corresponding to an average probability of person-specific AMD of ∼0.44) and a continuous standard normal covariate X. We varied the log OR of X on Y between zero (simulation under H_0 of no effect) and one.
2. To create the true eye-specific disease data (two binary observations per individual, (Z_1, Z_2)) we specified the conditional probability of being affected in both eyes given disease in at least one eye (i.e., Y = 1 based on "worse-eye definition), δ, to be (on average) $\delta = 1/(1 + \exp(-1)) = 0.73$. We assumed this probability to be either constant or varying with the continuous covariate X based on formula $\delta = 1/(1 + \exp(-(1 + 1 \times X))) = \text{Logist}(1 + 1 \times X)$. For all individuals with sampled Y = 1, we sampled a Bernoulli variable based on probability δ, to decide whether they were affected in both eyes or not. If they were affected on only one eye, we sampled randomly from the left or right.
3. To mimic the situation of missing information in one of two eyes, we sampled a Bernoulli random variable for each individual based on a fixed success probability (e.g., 0.75), to indicate whether information on both eyes was available. If not, we removed the disease information from a randomly selected eye.
4. To obtain eye-specific error-prone outcome data (Z_1^*, Z_2^*), we conditioned on the true, sampled observations (Z_1, Z_2), and sampled the error-prone outcomes based on specified classification probabilities, the sensitivity $P(Z^*=1|Z = 1)$ and specificity $P(Z^*=0|Z = 0)$. Sensitivity and specificity were either fixed (nondifferential misclassification, e.g., sens = spec= 0.9) or varying between individuals based on the formula $\text{sens} = \text{Logist}(2.20 + \beta_{\text{sens}} \times X)$ for different values of β_{sens} (analogously for the specificity, corresponding to an average sens = spec = 0.9).
5. Afterward, we split the data into two parts, the "main study" and the "validation" subset ($n^{\text{val}} = 1,000$, $n^{\text{main}} = 4,000$). For the validation subset we kept both, the true and the error-prone eye-specific AMD observations (Z_1, Z_2, Z_1^*, Z_2^*); for the main study, we kept only the error-prone outcomes (Z_1^*, Z_2^*) (or only the respective information for one of the two eyes, when information in one eye was missing for an individual).
6. For the naïve analysis ignoring response misclassification, we defined an observed, binary naïve person-specific outcome Y_{obs}^* the following way: for individuals from the validation data, we used the true eye-specific disease information; for individuals from the main study data, we used the error-prone eye-specific information. When disease information was available for both eyes, we defined $Y_{\text{obs}}^* = \max(Z_1, Z_2)$ or $Y_{\text{obs}}^* = \max(Z_1^*, Z_2^*)$, respectively; for observations with information only on one eye Z_1, we used $Y_{\text{obs}}^* = Z_1$ or $Y_{\text{obs}}^* = Z_1^*$. For individuals from the validation data with information on both eyes, $Y_{\text{obs}}^* = \max(Z_1, Z_2)$ corresponds to the true Y; for all others, Y_{obs}^* might be misclassified.

For each sampled dataset we estimated three models: (a) standard logistic regression based on the error-prone naïve worse entity outcome Y_{obs}^*, (b) the derived MLA (see above) modeling the probability of person-specific AMD and the probability of AMD in both eyes given AMD in at least one eye, δ, based on covariate X, while assuming a constant eye-specific sensitivity and specificity and accounting for missing information in one of two eyes (MLA1), and (c) the derived MLA allowing for a dependency of sensitivity and specificity on X (MLA2).

APPENDIX C: POWER ANALYSIS FOR REPORTED LEAD VARIANTS BASED ON UK BIOBANK SAMPLE SIZE

We wanted to evaluate the impact of using the MLA on selected variants including the 34 reported lead variants known for their association with advanced AMD. Given reported effect sizes and EAFs, we expected the power to detect some of these 34 associations to be limited in a sample size of approximately 3,500 cases and 44,500 controls. Therefore, we aimed to assess the power to detect reported genetic associations for AMD in the

available data of UK Biobank, to focus our analyses with the MLA only on adequately powered reported associations and to avoid over-interpreting noisy results from underpowered analyses. It is, however, not fully straight forward how to compute power for the scenario of "any AMD" from machine learning based disease classification, due to the power-diminishing effect of misclassification and some uncertainty of what effect size to use. We chose to use the reported (Fritsche et al., 2016) EAFs in advanced AMD cases and AMD-free controls for the established 34 lead variants and computed the power for a test on differences in (effect allele) fractions for differently sized groups (Cohen, 2013; Stephane, 2018). Group sizes correspond to the automated "any AMD" classification in UK Biobank GWAS data (Table S2). The number of observations in each group is two times the observed number of individuals, that is, $n_{\text{case}} = 2 \times 3{,}500$ and $n_{\text{contr}} = 2 \times 44{,}500$, since each individual contributes two (independent) alleles.

Based on these power calculations, we selected all lead variants with at least 80% power to yield Bonferroni-corrected ($\alpha = .05/34$) significant associations in UK Biobank. By this, we made the assumptions that EAFs in advanced AMD cases are transferable to EAFs of "any AMD" cases and that no misclassification was present in the machine learning-derived any AMD classification. Therefore, this is probably an overestimate of available power. We performed the power analysis, however, mainly to dismiss variants with an obvious lack of power.

APPENDIX D: MLA AVOIDS BIAS AND EXCESS OF TYPE I ERROR IN SIMULATION STUDIES

In our simulation study, we investigated bias and type I error of logistic regression-based association estimates for a binary worse entity outcome $Y := \max(Z_1, Z_2) \in \{0, 1\}$ and a continuous covariate X, when error-prone single-entity observations $(Z_1^*, Z_2^*) \in \{0,1\}$ are observed instead of the true entity-specific disease classifications $(Z_1, Z_2) \in \{0,1\}$. When utilizing the error-prone observations for deriving the worse entity outcomes $Y^* := \max(Z^*_1, Z^*_2)$, the entity-specific misclassification is passed on to the worse entity disease stage. We compare the performance of the naïve analysis (logistic regression ignoring misclassification) and the two versions of our MLA for different simulation scenarios.

In the naïve analysis, we found a similar pattern for bilateral disease misclassification as reported for classic diseases (Carroll et al., 2006; Neuhaus, 1999): (a) under the null hypothesis (Tables 1 and S1, $\beta_Y = 0$), we found biased estimates and a lack of type I error control (potential for false-positive association findings) for differential misclassification. With nondifferential misclassification, estimates were unbiased and type I error frequencies were at the desired levels. (b) When X was associated with true AMD (Tables 1 and S1, $\beta_Y = 1$), effect estimates were biased toward the null for nondifferential misclassification and into any direction for differential misclassification. Specific for the bilateral disease situation was (c) increasing bias with increasingly missing AMD in one of the two eyes, and (d) a larger bias by decreased specificity than by decreased sensitivity. (Tables 1 and S1).

In logistic regression, the larger the misclassification probabilities, the larger the bias of estimates (Neuhaus, 1999), with similar influence of increased probabilities for false-positive and false-negative classifications for balanced data. In the following, we provide an explanation of the findings (c) and (d) for bilateral diseases from above. Finding (c) is explained by the fact that an increased fraction of missing eyes implies a reduced sensitivity for person-specific AMD: AMD in the missing eye can be overlooked, which can lead to a false-negative person-specific AMD classification if only the missing eye of an individual is affected. Finding (d) was that decreased specificity had larger impact on bias than decreased sensitivity, for example, for (sens, spec) = (0.9, 0.9) and a fraction of 25% of individuals with "missing eyes" and a true log OR of X on Y of 1 the observed bias was −0.27. When the sensitivity was reduced to 0.8 (specificity = 0.9), the bias increased (in absolute value) to −0.32; when the specificity was reduced to 0.8 (sensitivity = 0.9), the bias increased to −0.39. This can be explained by rewriting the probability of misclassification in the worse entity outcome, $P(Y^* \neq Y)$ as

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

This illustrates the dependency of $P(Y^* \neq Y)$ on entity-specific sensitivity, specificity, probability of disease in both entities given disease in one eye δ, and the fraction of truly affected individuals $P(Y = 1)$. This probability can be evaluated for different combinations of parameters: for example, in the simulation study, we assumed $P(Y = 1) = 0.44, \delta = 0.75$ (Appendix B), which leads to a misclassification probability of 12%, 14%, or 22% for (sens, spec) = (0.9, 0.9), (sens, spec) = (0.8, 0.9), or (sens, spec) = (0.9, 0.8), respectively, illustrating the larger impact of reducing specificity. This is even more true in scenarios with a lower fraction of affected individuals: if

we assume a probability of person-specific disease of 0.10 instead of 0.44, we obtain misclassification probabilities of 17%, 18%, or 33%, for the same combinations of sensitivity and specificity. A reduced entity-specific specificity increases the probability of falsely classifying healthy entities toward disease, and falsely classifying only one of two healthy entities toward disease is sufficient to misclassify the person-specific disease status.

When applying the MLA1, we found it to effectively correct for bias and to yield the expected confidence interval coverage rates (~95%) when the misclassification was nondifferential, but we found it to still result in biased estimates and excess type I error when the misclassification was differential (Tables 1 and S1). When applying the MLA2, we found it effective in bias correction and type I error control under all misclassification scenarios, but with larger standard errors due to the larger number of parameters in the model (Tables 1 and S1). Overall, our simulation results documented substantial bias and lack of type I error control when the naïve analysis was applied to misclassified data and our MLA to effectively remove bias and keep type I error when specified correctly.

APPENDIX E: DETAILED RESULTS OF MLA FOR THE SELECTED 26 VARIANTS

For estimating sensitivity and specificity, we found the following: (a) for the three lead variants from this GWAS (*CFH*, *ARMS2/HTRA1*, or *HERC2*, respectively), the MLA1-derived sensitivity and specificity (at mean age and two copies of the noneffect allele) showed only small differences between the three variants (sensitivity = 65%, 67%, 63%; specificity = 98%, 98%, 99%, respectively, Table S6a). From a model without including a genetic covariate, we obtained an overall sensitivity of 64.5% (95% CI: 60.1%, 68.7%) and a specificity of 98.6% (98.4%, 98.8%). (b) We did not find strong evidence for associations with age using MLA1 or MLA2 based on any of the 26 selected variants, except for an association of the specificity with age based on MLA1 for the *HERC2* variant that disappeared when applying MLA2 (age: $p = 6.71 \times 10^{-9}$ or .70, respectively, Table S6a). (c) Applying MLA2, we found no association of the sensitivity with any selected variant ($p > .05/[23 \times 2]$), but a strong association of the specificity with the *HERC2* lead variant rs12913832 and with the reported *CFH* lead variant rs10922109 ($OR_{spec} = 0.64$, $P_{spec} = 7.38 \times 10^{-9}$ and $OR_{spec} = 1.36$, $P_{spec} = 2.29 \times 10^{-4}$, respectively; Table S6).

Second, we obtained genetic association estimates from MLA1 and MLA2 accounting for misclassification and compared these with naïve analysis estimates. We found interesting patterns: (a) when applying MLA1, we found comparable, slightly increased effect estimates for the *CFH*, *ARMS2/HTRA1,* and *HERC2* lead variant when compared to the naïve analysis (MLA1: OR = 1.23, 1.48, 1.34; $p = 1.69 \times 10^{-6}$, 8.9×10^{-18}, 1.11×10^{-12}; naïve: OR = 1.14, 1.30, 1.26, $p = 6.18 \times 10^{-7}$, 2.44×10^{-20}, 4.16×10^{-16}; Figure 2 and Table S7a). These results were as expected, since a model considering nondifferential misclassification leads in general, by assumption, to larger estimates and widened confidence intervals if any misclassification is present. (b) When applying MLA2, we found similar effect estimates for *CFH and ARMS2/HTRA1* compared to MLA1 and naïve analysis (OR = 1.19 or 1.28, respectively), which is in line with limited bias due to differential misclassification. We also found larger *p* values ($p = .02$ or 2.47×10^{-4}, respectively, which is in line with larger uncertainty when estimating more model parameters. In contrast, we found a completely vanished effect estimate for the *HERC2* variant (MLA2: OR = 1.03, $p = .76$; Figure 2 and Table S7a), indicating a bias in the naïve analysis and MLA1 when ignoring a differential misclassification. (c) Effect estimates for the three reported *CFH* variants increased when applying MLA2 compared to the naïve analysis. This was particularly interesting for the reported *CFH* lead variant rs10922109, where we now observed a nominally significant association into the reported direction (MLA2: OR = 1.15, $p = .047$; naïve: OR = 1.00, $p = .98$; Table S7c). This is in line with the observed association of the specificity with this *CFH* variant. (d) For the other 20 reported lead variants, we found many variants with increased effect estimates by MLA1 or MLA2 compared to the naïve analysis; effect estimates were mostly more comparable to reported effect sizes for advanced AMD (Fritsche et al., 2016; Figure 3c). For one variant, this MLA2 analysis yielded an effect into the opposite direction compared to the reported effect direction, which is the *C9* lead variant (OR = 0.83, $p = .59$). With an effect allele frequency of 1%, it is the rarest analyzed variant of the 26 selected variants and estimates from the reported association as well as for the MLA2 analysis have low precision (i.e., large standard errors).