SCIENTIFIC REPORTS

OPEN

A Validated Phenotyping Algorithm for Genetic Association Studies in Age-related Macular Degeneration

Received: 10 April 2015
Accepted: 13 July 2015
Published: 10 August 2015

Joseph M. Simonett[1], Mahsa A. Sohrab[1], Jennifer Pacheco[2], Loren L. Armstrong[3], Margarita Rzhetskaya[3], Maureen Smith[2], M. Geoffrey Hayes[2,3,4,5] & Amani A. Fawzi[1]

Age-related macular degeneration (AMD), a multifactorial, neurodegenerative disease, is a leading cause of vision loss. With the rapid advancement of DNA sequencing technologies, many AMD-associated genetic polymorphisms have been identified. Currently, the most time consuming steps of these studies are patient recruitment and phenotyping. In this study, we describe the development of an automated algorithm to identify neovascular (wet) AMD, non-neovascular (dry) AMD and control subjects using electronic medical record (EMR)-based criteria. Positive predictive value (91.7%) and negative predictive value (97.5%) were calculated using expert chart review as the gold standard to assess algorithm performance. We applied the algorithm to an EMR-linked DNA bio-repository to study previously identified AMD-associated single nucleotide polymorphisms (SNPs), using case/control status determined by the algorithm. Risk alleles of three SNPs, rs1061170 (*CFH*), rs1410996 (*CFH*), and rs10490924 (*ARMS2*) were found to be significantly associated with the AMD case/control status as defined by the algorithm. With the rapid growth of EMR-linked DNA biorepositories, patient selection algorithms can greatly increase the efficiency of genetic association study. We have found that stepwise validation of such an algorithm can result in reliable cohort selection and, when coupled within an EMR-linked DNA biorepository, replicates previously published AMD-associated SNPs.

Age-related macular degeneration (AMD) is a multifactorial neurodegenerative disease that is the leading cause of blindness in western individuals over the age of 65[1–4]. Clinical presentation of AMD is heterogeneous, with many genetic and environmental risk factors contributing to its pathogenesis[5]. The rate of identification of AMD-associated genetic risk factors, including but not limited to single nucleotide polymorphisms (SNPs) in *CFH*, *ARMS2* and *HTRA1* genes, has increased rapidly with the utilization of genome-wide association studies (GWAS)[6–9]. These studies have led to a better understanding of AMD pathophysiology, creation of genetic based prediction models and a plethora of AMD pharmacogenomics studies[8,10–15]. GWAS studies have also identified environmental exposures that interact with AMD genetic risk factors, highlighting the importance of developing accurate criteria for clinical phenotyping in order to discriminate disease and control populations[16,17]. One important barrier to genetic association

[1]Department of Ophthalmology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. [2]Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. [3]Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. [4]Department of Anthropology, Northwestern University, Evanston, IL. [5]Northwestern Comprehensive Center on Obesity, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. Correspondence and requests for materials should be addressed to A.A.F. (email: amani.fawzi@northwestern.edu)

studies is the time consuming process of patient recruitment, phenotyping, and DNA collection necessary to build sufficiently powered cohorts. This process can be accelerated by implementing electronic medical record (EMR)-linked DNA bio-repositories, which allow multiple unrelated fields of research to share a large, common pool of genetic data coupled to a searchable EMR, significantly facilitating phenotype-genotype comparisons[18–22]. The use of high-throughput clinical phenotyping (HTCP) algorithms that apply specific inclusion and exclusion criteria to clinical data, available through the EMR, could generate a large cohort of potentially eligible study subjects and, in the case of EMR-linked DNA bio-repositories, would capitalize on previously genotyped or imputed data[23,24]. As the number and size of EMR-linked DNA bio-repositories grow, the need for accurate, validated HTCP algorithms continues to increase[25].

HTCP algorithms have been used in ophthalmologic research; a recent systematic review identified seven North American studies that reported EMR based algorithms to identify patients with a diagnosis of uveitis[26]. All of the identified studies used inclusion criteria based on international classification of disease-9 (ICD-9) codes and three of these seven studies described validation of their algorithm. These studies investigated the epidemiology and treatment response of uveitis, but not genetic associations. Another study investigating ocular complications after anti-vascular endothelial growth factor therapy employed an algorithm to identify exudative AMD, which included any patient with ICD-9 diagnosis codes 364.00, 364.05, 364.10, 364.11, however this study did not describe a validation procedure for their algorithm[27].

Despite the rapid growth of SNP association studies in AMD populations, validated HTCP algorithms have yet to be implemented in the field. We hypothesized that a patient cohort identified with an HTCP algorithm from an EMR-linked DNA bio-repository can be used to perform a SNP association study using previously acquired DNA samples to replicate published AMD associations. To test our hypothesis, we developed an HTCP algorithm to identify AMD and control patients using EMR data, which we validated through expert chart review. We then applied our HTCP algorithm to an institutional EMR-linked bio-repository to test our hypothesis.

Methods

Ethics. The study was approved by the Institutional Review Board at Northwestern University and adhered to the tenets set forth by the Declaration of Helsinki; informed consent was obtained from patients prior to enrollment in this study.

Algorithm Development. We developed the algorithm to identify all AMD cases using the criterion of AMD ICD-9 codes entered by an ophthalmologist (362.50, 362.51, 362.52, 362.16, 362.57). To classify cases as "wet" AMD cases within this population, we additionally required a current procedural terminology (CPT) code (J2778: ranibizumab injection, J9035, J3490 or J3590: bevacizumab injection), or an order or prescription for ranibizumab , bevacizumab, or aflibercept. This initial algorithm was tested by unsupervised random selection of 20 suspected AMD patient charts (10 dry and 10 wet cases of AMD) from the Northwestern University Department of Ophthalmology. Based on this initlal pilot study, we revised the HTCP algorithm to require subjects to be $\geq$60 years of age at the time of the first AMD diagnosis and to have $\geq$2 visits that were associated with the AMD ICD-9 codes. Furthermore, an ICD-9 code starting with 362.5 on the same date as the procedural CPT code or medication order was required for "wet" AMD classification. All AMD cases not meeting the wet AMD criteria were labeled as "dry" AMD (Fig. 1). Patients were classified as controls if they had $\geq$1 ophthalmology visit within the last two years, were $\geq$60 years of age at the time of the visit, and did not receive an AMD or AMD-associated diagnosis (we excluded the following non-specific or unrelated ICD-9 codes 362 or 377.21).

The modified HTCP algorithm was re-tested on the Northwestern University clinical EMR, and was set up to extract a selection of 100 charts: 30 dry AMD, 30 wet AMD and 40 control charts. None of the 20 charts from the original pilot study were included in this validating study. The annonymized charts selected by the algorithm were verified by experienced graders/retina specialists (MAS, AAF) who examined the clinical notes and retinal imaging (retinal fundus photographs, angiography and optical coherence tomography) of these 100 charts. A case was considered to have a diagnosis of "wet" AMD if at least one eye was determined to have wet AMD pathology by the graders. The positive predictive value (PPV), negative predictive value (NPV), and false negative rate (FNR) for overall AMD diagnosis and "wet vs dry" sub-classification were calculated to ensure we have achieved our preset criterion for HTCP having >90% accuracy. Once validated, the HTCP algorithm was then applied to all 11,075 subjects enrolled in the NUGene Project, an EMR-linked DNA bio-repository at Northwestern University Center for Genetic Medicine, to identify AMD cases and controls for the subsequent genetic association study. We specifically ensured that none of the patients/charts used in the validation or pilot studies were participants in the NUGene Project.

Genotyping. We initially selected 11 SNPs based on previous studies showing significant association with AMD: rs1061170, rs1410996 (*CFH*), rs10490924 (*ARMS2*), rs11200638 (*HTRA1*), rs2230199 (*C3*), rs833069 (*VEGFA*), rs8017304 (*RAD51L1*), rs4151667, rs541862, rs641153 (*CFB),* rs9332739 (*C2)*[10,11,28–32]. Through prior investigations, as part of the Electronic Medical Records and Genomics (eMERGE) network, genome-wide genotype data, imputed to >36 million SNPs using the 1000 Genomes Project

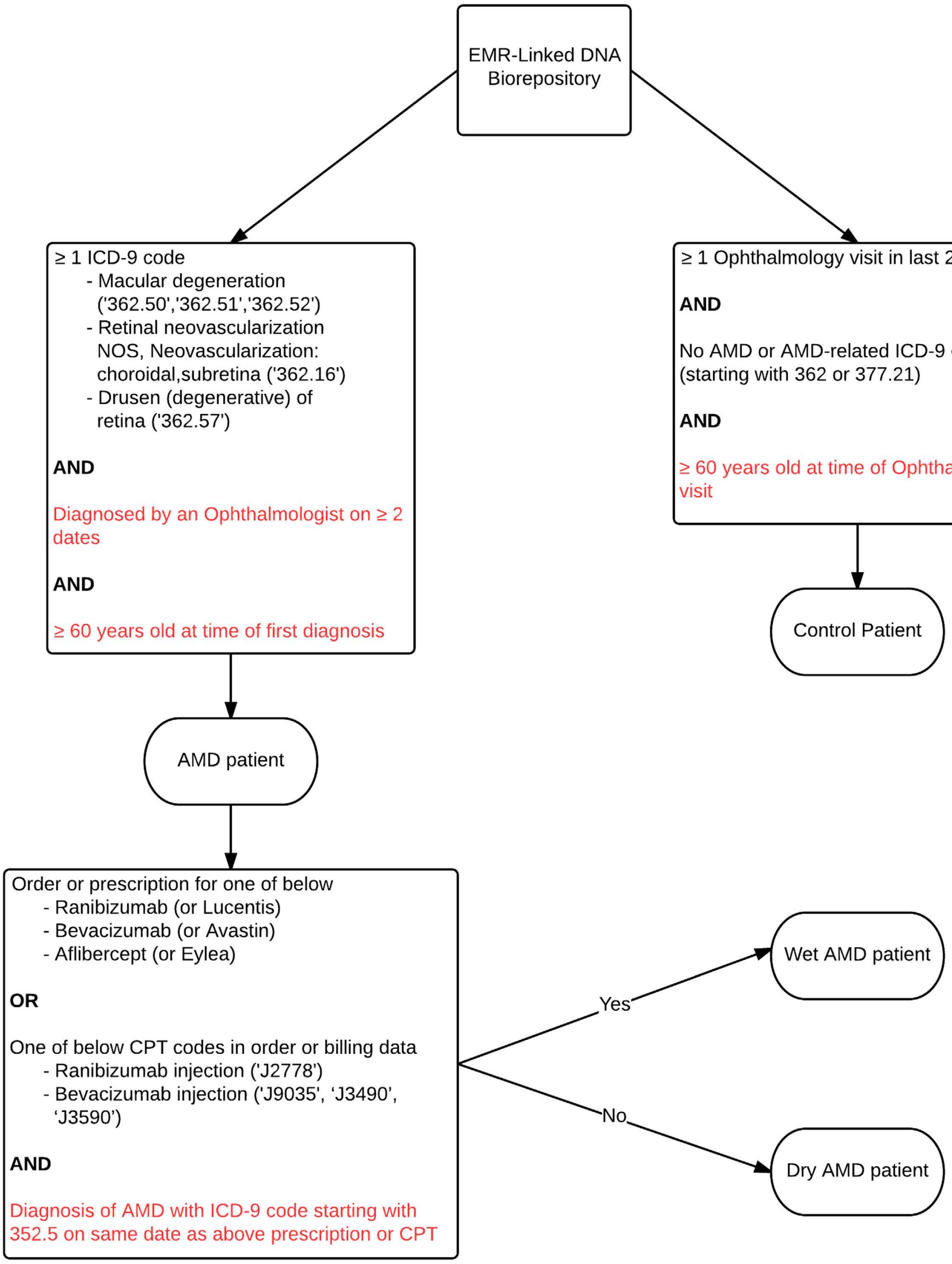
EMR-Linked DNA Biorepository
≥ 1 ICD-9 code
- Macular degeneration ('362.50','362.51','362.52')
- Retinal neovascularization NOS, Neovascularization: choroidal,subretina ('362.16')
- Drusen (degenerative) of retina ('362.57')
AND
Diagnosed by an Ophthalmologist on ≥ 2 dates
AND
≥ 60 years old at time of first diagnosis
≥ 1 Ophthalmology visit in last 2 years
AND
No AMD or AMD-related ICD-9 codes (starting with 362 or 377.21)
AND
≥ 60 years old at time of Ophthalmology visit
Control Patient
AMD patient
Order or prescription for one of below
- Ranibizumab (or Lucentis)
- Bevacizumab (or Avastin)
- Aflibercept (or Eylea)
OR
One of below CPT codes in order or billing data
- Ranibizumab injection ('J2778')
- Bevacizumab injection ('J9035', 'J3490', 'J3590')
AND
Diagnosis of AMD with ICD-9 code starting with 352.5 on same date as above prescription or CPT
Yes
No
Wet AMD patient
Dry AMD patient

Figure 1. High-throughput clinical phenotyping algorithm outline. Final HTCP algorithm applied to EMR-linked DNA biorepository. Red criteria were added after first round of case selection/expert chart review. ICD-9: International Classification of Disease-9, CPT: current procedural terminology.

cosmopolitan reference panel with IMPUTE2 were available for 38 of the 61 AMD cases and 167 of the 332 controls subjects identified with the HTCP algorithm[21,33,34]. For AMD case subjects that did not have available genotype data, we performed direct genotyping of 7 of the 11 SNPs (7 SNPs at *CFH, ARMS2, HTRA1, C3, VEGFA* and *RAD51L1*). Selection of these SNPs was based on a *P* value threshold of <0.19 from allele case/control associations using the imputed genotype data from the initial 38 cases and 167 controls. SNPs were genotyped directly from bio-banked genomic DNA samples that were previously isolated from peripheral blood leukocytes with the Gentra Autopure system at the Northwestern University Center for Genetic Medicine Genomics Core Facility. SNP genotyping was performed by direct sequencing of PCR products using either the forward or reverse primer used as sequencing primer (NEB Q5® High-Fidelity DNA Polymerase following manufacturer's recommended protocols) at the Northwestern University Center for Genetic Medicine Genomics Core Facility. Primers and T_m are listed in Supplementary Table 1. We excluded genetic data from non-European-ancestry patients (4 out of 61 AMD cases and 25 out of 167 controls). None of the final 57 AMD cases had been included in the prior pilot or validation studies.

Classification	PPV	NPV	FNR
Overall AMD	91.7%	97.5%	1.8%
Dry AMD	73.3%	95.7%	12.0%
Wet AMD	86.7%	92.9%	16.1%

Table 1. Final Algorithm Performance Metrics. Final algorithm performance metrics for classifying cases as AMD, regardless of dry vs wet status, and for determining dry vs wet status of identified AMD cases. PPV: Positive predictive value, NPV: negative predictive value, FNR: false negative rate.

Statistical Analysis. Demographic and past medical history data was compared between algorithm identified cases and controls and analyzed using Chi-squared and student-t tests. For genotyped SNPs, a logistic regression between AMD case-control status and allelic dosage was performed using PLINK to estimate the odds ratios (OR) and 95% confidence intervals to assess the significant association for each SNP with the diagnosis of AMD[35]. Regression models assumed an additive genetic model and adjusted for age, sex, body mass index (BMI), and smoking history, which are known confounders of AMD[36]. Similarly, we used the frequentist approach in SNPTEST v.2.2.0 for imputed SNPs to perform the same regression analyses[37]. SNP allele counts from both imputed and genotyped cases were used together to calculate final allelic ORs and *P* values. We considered SNP associations to be significant if they met a Bonferroni multiple comparison corrected *P* value threshold of $P < 0.05/11$ SNPs ~ 4.5×10^{-3}.

Results

Algorithm development. Of the 20 suspected AMD patient charts selected by the initial algorithm, 45% (9/20) of patients were correctly classified as either dry or wet AMD. In the dry AMD category, misclassified cases included pattern dystrophy, diabetic retinopathy and central serous chorioretinopathy. In the wet AMD category, misclassified cases included proliferative diabetic retinopathy, high myopia with lacquer crack, idiopathic macular scar, end-stage retinopathy of prematurity, and birdshot chorioretinopathy.

After additional modifications (≥2 diagnosis dates, ≥60 years of age at diagnosis and clinic visit, and requiring AMD diagnosis at the time of CPT codes and medication orders) to improve the algorithm, the readers performed revalidation of the algorithm using the clinical EMR database. Of the 60 patients classified by the algorithm as AMD and 40 classified as controls, we found overall 94% (94/100) of patients were correctly classified as either AMD or control. Of the 60 patients identified as having a diagnosis of AMD, 5 were misclassified, including a case of proliferative diabetic retinopathy, a case of posterior vitreous detachment with recurrent vitreous hemorrhage, a case of atypical angioid streaks, a case of pattern dystrophy, and a case of ruptured macroaneurysm. Of the 40 patients identified as controls, 1 case was misclassified and found to have a large macular scar with history of polypoidal choroidal vasculopathy. For the overall AMD classification, the PPV was 91.67% (55/60), NPV was 97.50% (39/40), and FNR was 1.79% (1/56) (Table 1).

We further evaluated the algorithm performance in terms of AMD sub-classification (wet vs dry AMD). Of the 55 correctly classified AMD cases, 26 were identified as dry and 29 as wet AMD by the algorithm. Four of the 26 cases identified as dry AMD were found to be wet AMD following expert review. Three of the 29 cases identified as wet AMD were found to be dry AMD following expert review. PPV, NPV and FNR of wet and dry determination were calculated (Table 1).

Algorithm Application to EMR-linked DNA bio-repository to confirm established AMD SNP associations. The refined HTCP algorithm identified 61 AMD cases and 332 controls within the NUGene project bio-repository. Of these, 38 AMD cases and 167 controls had genome-wide imputed data available. All SNPs tested imputed well (IMPUTE2 info score ≥0.75). The remaining 23 AMD cases were genotyped for the 7 SNPs of interest. Low quality sequencing occurred at rs1410996 in one AMD case, which was excluded from the rs1410996 allele association analysis. No additional controls were genotyped. Following exclusion of non European-ancestry subjects, a total of 57 AMD cases (imputed + directly sequenced) and 142 controls remained.

Demographics and past medical history, including diagnosis of type 2 diabetes mellitus, glaucoma and cataracts, of the AMD cases and controls used in SNP analysis are shown in Table 2. Age at last eye exam, gender, smoking history and BMI data was available for all AMD cases and controls. AMD cases were significantly older (average age 78.3 vs 69.3), had a significantly higher rate of cataracts (86.9% vs 68.9%), and trended towards having a lower BMI than controls.

The risk alleles of 3 previously identified AMD associated SNPs, rs1061170 (*CFH*), rs1410996 (*CFH*), and rs10490924 (*ARMS2*), were significantly associated (P value $< 4.5 \times 10^{-3}$) with AMD case/control status as identified by the HTCP (Table 3). rs11200638 (*HTRA1*) and rs2230199 (*C3*) approached significance but did not meet the Bonferroni corrected *P* value threshold. Significant allele associations were in the same direction and had similar ORs to previous reports. rs833069 (*VEGFA*) and rs8017304

Phenotype	AMD identified cases	Control identified cases	*P*
Age at last eye exam	78.3	69.3	**5.9×10^{-16}**
Female sex	70.5%	78.4%	0.15
History of smoking	52.5%	52.1%	0.95
BMI	26.3	27.9	0.07
Type 2 DM	29.5%	34.1%	0.62
Glaucoma	42.6%	35.5%	0.36
Cataracts	86.9%	68.9%	**5.3×10^{-3}**

Table 2. Demographic Characteristics of HTCP Defined AMD cases and controls. BMI: Body mass index, DM: Diabetes mellitus. Bolded *P* values are statistically significant ($P < 0.05$).

SNP	Near by gene	Previously reported risk allele	RAF in AMD Cases	RAF in controls	OR (CI)	*P*	Previously reported OR (CI)	Previously reported RAF	Published source
rs1061170	*CFH*	C	0.580	0.363	2.43 (1.55–3.79)	**2.3E-04**	1.86 (1.77–1.97)	0.49	Sofat *et al.*, 2012
rs10490924	*ARMS2*	T	0.321	0.201	1.88 (1.15–3.07)	**2.0E-03**	2.76 (2.72–2.80)	0.30	Fritsche *et al.*, 2013
rs1410996	*CFH*	C	0.732	0.584	1.94 (1.20–3.14)	**2.6E-03**	1.98 (1.44–2.72)	0.60	Mori *et al.*, 2007
rs11200638	*HTRA1*	A	0.304	0.205	1.69 (1.03–2.78)	9.6E-03	1.80 (1.34–2.39)	0.31	Hadley *et al.*, 2010
rs2230199	*C3*	C	0.607	0.448	1.90 (1.22–2.97)	0.033	1.42 (1.37–1.47)	0.20	Fritsche *et al.*, 2013
rs8017304	*RAD51L1*	A	0.670	0.581	1.46 (0.92–2.31)	0.13	1.11 (1.08–1.14)	0.61	Fritsche *et al.*, 2013
rs833069	*VEGFA*	G	0.286	0.377	0.66 (0.41–1.06)	0.04	1.69 (1.26–2.26)	0.26	Galan *et al.*, 2010

Table 3. Pooled Imputed and Directly Genotyped Association Results Between Previously Identified AMD Risk Alleles and HTCP Defined AMD case/control status. SNP: Single nucleotide polymorphism, RAF: Risk allele frequency, OR (CI): Odds ratio (95% Confidence Interval). Bolded *P* values are statistically significant after Bonferroni correction ($P < 4.5 \times 10^{-3}$).

(*RAD51L1*), as well as the SNPs for which direct genotyping was not obtained in the additional cases, were not significantly associated with algorithm-identified AMD cases/control status in our study.

Discussion

Recent advances in genotyping and sequencing technology have significantly outpaced the development of HTCP phenotyping capabilities, causing labor-intensive patient identification and DNA collection to be the rate-limiting step in genetic association studies[38]. As the use of EMR-linked DNA bio-repositories expands, improved HTCP algorithms for cohort selection offers an appealing alternative to automate these processes and share clinically linked genotype data across research fields. These methods are of particular importance for chronic, complex diseases like AMD that are associated with a large number of genetic and environmental risk factors. HTCP algorithms will require a multi-step validation method that achieves sufficiently high phenotyping accuracy, in particular high PPV, necessary for identifying genetic variants that are associated with multifactorial diseases.

We found that relying on ICD-9 codes alone for AMD patient selection was not satisfactory. The addition of ICD-9-linked CPT codes or medication records and age restrictions improved algorithm accuracy at identifying both AMD (wet and dry) and control patients. 61 of 11,075 subjects enrolled in the EMR-linked DNA bio-repository were assigned AMD case status by the HTCP algorithm (0.55%), expectedly lower than the prevalence reported in the age 50 and older population as this database contained subjects of all ages. Demonstrating associations with SNPs previously shown to be associated with AMD, with odds ratios and confidence intervals that substantially overlap with those from the literature, argues that the PPV and NPV achieved by this algorithm are high enough to properly identify AMD cases and controls. The PPV for both dry and wet AMD subtype determination by the algorithm was <90% (73.3% and 86.7%, respectively) and the FNRs were significantly higher than the FNR for overall AMD case determination (12.0% and 16.1% vs 1.8%). The relative weakness of the algorithm in discriminating AMD subtypes is likely multifactorial and complicated by the spectrum of AMD retinal pathology. Further improvements could include the addition of searchable keywords within the EMR free text in addition to ICD9/CPT codes. Application to a larger AMD cohort will be needed to determine if the algorithm can be reliably used to differentiate wet and dry AMD.

Three of the seven previously reported AMD-associated SNPs examined in this study were found to be significantly associated (at Bonferroni correction threshold) with case/control status identified by the HTCP algorithm, and one of the SNPs (rs833069) trended in the opposite direction from that previously published. Potential explanations for why more SNPs were not significantly associated with case/control status include inadequate power due to sample size or insufficient phenotyping accuracy. Additionally, genetic heterogeneity between the previous populations used to identify the SNP associations and the NUGene project population may partially account for these findings. Increased compatibility between EMR systems will allow for more wide spread application of HTCP algorithms and will greatly increase the potential sample size.

One barrier to widespread use of HTCP algorithms is the lack of standardization across EMRs. Although the HTCP algorithm was sufficiently accurate in this study to identify some genetic associations, accuracy will become more difficult when algorithms are applied across multiple EMR systems, especially when they attempt to include complex additional criteria such as treatment response. Although ICD-9 and CPT codes are shared across US EMR systems, validation of this HTCP algorithm on external EMRs will be important in assessing its performance in other EMR systems. Shared billing codes and increased utilization of EMR organization tools continue to improve, including efforts to develop external informatics infrastructure on which to normalize EMR data[39]. Particularly, for the purpose of addressing Meaningful Use standards, phenotyping algorithms have been successfully applied to clinical EMR platforms and accurately identified specific patient cohorts[40,41]. Additionally, introduction of active learning to HTCP algorithms has been shown to decrease the number of clinical data annotations necessary to achieve a precise classification model[42]. Similar approaches in modifying EMR-linked DNA bio-repositories and HTCP algorithms are possible and would likely improve the classification function.

Multiple AMD associated SNPs, including those confirmed in this study, are believed to play an important role in the progression through specific AMD stages[43]. Utilizing combinations of genetic, clinical and demographic data, several AMD progression risk prediction models have been validated in independent patient cohorts[44–47]. Accuracy of modeling the risk for progression from early stage AMD to advanced stages of either geographic atrophy (GA) or choroidal neovascularization (CNV) improves with the inclusion of relevant genetic markers (GA: C-statistic = 0.94, CNV: C-statistic = 0.96) compared to phenotype-only models (C-statistic = 0.63 to 0.89)[12]. Additionally, AMD risk prediction models that incorporate a higher number of AMD associated SNPs within the *CFH* gene, compared to models that use only one or two, have been shown to be more accurate[12]. Therefore, great insight can be gained by exploring the conferred risk of haplotype combinations, even when the association between individual haplotypes and AMD are already known. Utilization of rapid association studies and validated HTCP algorithms may be an ideal method for identifying and confirming additional SNPs for this purpose.

Results from pharmaco-genomic studies on the effect of AMD-associated SNPs on clinical response to anti-VEGF treatment have been conflicting[14,15,48]. Expansion of an AMD HTCP algorithm to include treatment response has the potential to strengthen these pharmaco-genomic studies by accessing larger treatment cohorts. Identification of eye specific AMD status rather than individual AMD status will be important when exploring pharmacogenomics and treatment response in future studies; ICD-9 and CPT codes are currently not linked to a specific eye. Future use of ICD-10, which includes left and right eye specific diagnostic codes, may facilitate eye level HTCP classification.

The American Academy of Ophthalmology (AAO) 2012 recommendations for genetic testing discussed the value of genetic testing for multifactorial disorders such as AMD[49]. They recommended avoiding routine genetic screening until a specific treatment or surveillance is proven to be beneficial to patients with a given genotype. Significant advances need to be made in AMD research, including those discussed above, before personal sequencing data can directly influence patient care[50]. We believe that validated HTCP algorithms, such as the one developed in this study, when combined with EMR-linked DNA bio-repositories will become valuable tools to increase the efficiency of association studies and should be used to bring us closer to the ultimate goal of personalized medicine in AMD treatment.

References

1. Swaroop, A., Chew, E. Y., Rickman, C. B. & Abecasis, G. R. Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annual review of genomics and human genetics* **10,** 19–43, doi: 10.1146/annurev.genom.9.081307.164350 (2009).
2. Klein, R., Klein, B. E. & Linton, K. L. Prevalence of age-related maculopathy. The Beaver Dam Eye Study. *Ophthalmology* **99,** 933–943 (1992).
3. Bonastre, J. *et al.* The epidemiology, economics and quality of life burden of age-related macular degeneration in France, Germany, Italy and the United Kingdom. *The European journal of health economics : HEPAC : health economics in prevention and care* **3,** 94–102, doi: 10.1007/s10198-002-0104-y (2002).
4. Friedman, D. S. *et al.* Prevalence of age-related macular degeneration in the United States. *Archives of ophthalmology* **122,** 564–572, doi: 10.1001/archopht.122.4.564 (2004).
5. Seddon, J. M. & Chen, C. A. The epidemiology of age-related macular degeneration. *International ophthalmology clinics* **44,** 17–39 (2004).

6. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science (New York, N.Y.)* **308,** 421–424, doi: 10.1126/science.1110189 (2005).
7. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)* **308,** 385–389, doi: 10.1126/science.1109557 (2005).
8. Tong, Y. *et al.* LOC387715/HTRA1 gene polymorphisms and susceptibility to age-related macular degeneration: A HuGE review and meta-analysis. *Molecular vision* **16,** 1958–1981 (2010).
9. Yates, J. R. *et al.* Complement C3 variant and the risk of age-related macular degeneration. *The New England journal of medicine* **357,** 553–561, doi: 10.1056/NEJMoa072618 (2007).
10. Fritsche, L. G. *et al.* Seven new loci associated with age-related macular degeneration. *Nature genetics* **45,** 433–439, 439e431-432, doi: 10.1038/ng.2578 (2013).
11. Sofat, R. *et al.* Complement factor H genetic variant and age-related macular degeneration: effect size, modifiers and relationship to disease subtype. *International journal of epidemiology* **41,** 250–262, doi: 10.1093/ije/dyr204 (2012).
12. Perlee, L. T. *et al.* Inclusion of genotype with fundus phenotype improves accuracy of predicting choroidal neovascularization and geographic atrophy. *Ophthalmology* **120,** 1880–1892, doi: 10.1016/j.ophtha.2013.02.007 (2013).
13. Seddon, J. M., Reynolds, R., Yu, Y., Daly, M. J. & Rosner, B. Risk models for progression to advanced age-related macular degeneration using demographic, environmental, genetic, and ocular factors. *Ophthalmology* **118,** 2203–2211, doi: 10.1016/j.ophtha.2011.04.029 (2011).
14. Matsumiya, W. *et al.* Evaluation of clinical and genetic indicators for the early response to intravitreal ranibizumab in exudative age-related macular degeneration. *Pharmacogenomics* **15,** 833–843, doi: 10.2217/pgs.14.51 (2014).
15. Hagstrom, S. A. *et al.* Pharmacogenetics for genes associated with age-related macular degeneration in the Comparison of AMD Treatments Trials (CATT). *Ophthalmology* **120,** 593–599, doi: 10.1016/j.ophtha.2012.11.037 (2013).
16. Naj, A. C. *et al.* Genetic factors in nonsmokers with age-related macular degeneration revealed through genome-wide gene-environment interaction analysis. *Annals of human genetics* **77,** 215–231, doi: 10.1111/ahg.12011 (2013).
17. Courtenay, M. D. *et al.* Set-based joint test of interaction between SNPs in the VEGF pathway and exogenous estrogen finds association with age-related macular degeneration. *Investigative ophthalmology* & *visual science*, doi: 10.1167/iovs.14-14494 (2014).
18. Kho, A. N. *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Science translational medicine* **3,** 79re71, doi: 10.1126/scitranslmed.3001807 (2011).
19. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *American journal of human genetics* **86,** 560–572, doi: 10.1016/j.ajhg.2010.03.003 (2010).
20. Oetjens, M. *et al.* Utilization of an EMR-biorepository to identify the genetic predictors of calcineurin-inhibitor toxicity in heart transplant recipients. Paper presented at Pacific Symposium on Biocomputing: 2013, Hawaii, US. Place of publication: *Biocomputing* 253-264, doi: 10.1142/9789814583220_0025 (2014).
21. McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics* **4,** 13, doi: 10.1186/1755-8794-4-13 (2011).
22. Wilke, R. A. *et al.* The emerging role of electronic medical records in pharmacogenomics. *Clinical pharmacology and therapeutics* **89,** 379–386, doi: 10.1038/clpt.2010.260 (2011).
23. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in medicine: official journal of the American College of Medical Genetics* **15,** 761–771, doi: 10.1038/gim.2013.72 (2013).
24. Wilke, R. A. *et al.* Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clinical medicine* & *research* **5,** 1–7, doi: 10.3121/cmr.2007.726 (2007).
25. Newton, K. M. *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association: JAMIA* **20,** e147–154, doi: 10.1136/amiajnl-2012-000896 (2013).
26. Williams, S. E., Carnahan, R. & McPheeters, M. L. A systematic review of validated methods for identifying uveitis using administrative or claims data. *Vaccine* **31 Suppl 10,** K88–97, doi: 10.1016/j.vaccine.2013.03.077 (2013).
27. Day, S. *et al.* Ocular complications after anti-vascular endothelial growth factor therapy in Medicare patients with age-related macular degeneration. *American journal of ophthalmology* **152,** 266–272, doi: 10.1016/j.ajo.2011.01.053 (2011).
28. Hadley, D. *et al.* Analysis of six genetic risk factors highly associated with AMD in the region surrounding ARMS2 and HTRA1 on chromosome 10, region q26. *Investigative ophthalmology* & *visual science* **51,** 2191–2196, doi: 10.1167/iovs.09-3798 (2010).
29. Galan, A. *et al.* Association of age-related macular degeneration with polymorphisms in vascular endothelial growth factor and its receptor. *Ophthalmology* **117,** 1769–1774, doi: 10.1016/j.ophtha.2010.01.030 (2010).
30. Huang, L. *et al.* Different hereditary contribution of the CFH gene between polypoidal choroidal vasculopathy and age-related macular degeneration in Chinese Han people. *Investigative ophthalmology* & *visual science* **55,** 2534–2538, doi: 10.1167/iovs.13-13437 (2014).
31. Thakkinstian, A. *et al.* The association between complement component 2/complement factor B polymorphisms and age-related macular degeneration: a HuGE review and meta-analysis. *American journal of epidemiology* **176,** 361–372, doi: 10.1093/aje/kws031 (2012).
32. Nakata, I. *et al.* Significance of C2/CFB variants in age-related macular degeneration and polypoidal choroidal vasculopathy in a Japanese population. *Investigative ophthalmology* & *visual science* **53,** 794–798, doi: 10.1167/iovs.11-8468 (2012).
33. Verma, S. S., Andrade de, M., Tromp, G. C., Kuivaniemi, H. S., Pugh, E. & Namjou-Khales, B. Imputation and QC for combining multiple Genome-Wide Datasets. *Front. Genet.* **5,** 370, doi: 10.3389/fgene.2014.00370 (2014)
34. Crawford, D. C. *et al.* eMERGEing progress in genomics-the first seven years. *Frontiers in genetics* **5,** 184, doi: 10.3389/fgene.2014.00184 (2014).
35. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81,** 559–575, doi: 10.1086/519795 (2007).
36. Jonasson, F. *et al.* Five-year incidence, progression, and risk factors for age-related macular degeneration: the age, gene/environment susceptibility study. *Ophthalmology* **121,** 1766–1772, doi: 10.1016/j.ophtha.2014.03.013 (2014).
37. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39,** 906–913, doi: 10.1038/ng2088 (2007).
38. Hennekam, R. C. & Biesecker, L. G. Next-generation sequencing demands next-generation phenotyping. *Human mutation* **33,** 884–886, doi: 10.1002/humu.22048 (2012).
39. Rea, S. *et al.* Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *Journal of biomedical informatics* **45,** 763–771, doi: 10.1016/j.jbi.2012.01.009 (2012).
40. Pathak, J. *et al.* Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association : JAMIA* **20,** e341–348, doi: 10.1136/amiajnl-2013-001939 (2013).

41. Li, D. *et al.* Modeling and executing electronic health records driven phenotyping algorithms using the NQF Quality Data Model and JBoss(R) Drools Engine. *AMIA … Annual Symposium proceedings/AMIA Symposium. AMIA Symposium* **2012,** 532–541 (2012).
42. Chen, Y. *et al.* Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association: JAMIA* **20,** e253–259, doi: 10.1136/amiajnl-2013-001945 (2013).
43. Yu, Y., Reynolds, R., Rosner, B., Daly, M. J. & Seddon, J. M. Prospective assessment of genetic effects on progression to different stages of age-related macular degeneration using multistate Markov models. *Investigative ophthalmology & visual science* **53,** 1548–1556, doi: 10.1167/iovs.11-8657 (2012).
44. Chiu, C. J. *et al.* A risk score for the prediction of advanced age-related macular degeneration: development and validation in 2 prospective cohorts. *Ophthalmology* **121,** 1421–1427, doi: 10.1016/j.ophtha.2014.01.016 (2014).
45. Seddon, J. M., Reynolds, R., Yu, Y. & Rosner, B. Validation of a prediction algorithm for progression to advanced macular degeneration subtypes. *JAMA ophthalmology* **131,** 448–455, doi: 10.1001/jamaophthalmol.2013.2578 (2013).
46. Hageman, G. S. *et al.* Clinical validation of a genetic model to estimate the risk of developing choroidal neovascular age-related macular degeneration. *Human genomics* **5,** 420–440 (2011).
47. Buitendijk, G. H. *et al.* Prediction of age-related macular degeneration in the general population: the Three Continent AMD Consortium. *Ophthalmology* **120,** 2644–2655, doi: 10.1016/j.ophtha.2013.07.053 (2013).
48. Imai, D. *et al.* CFH, VEGF, and PEDF genotypes and the response to intravitreous injection of bevacizumab for the treatment of age-related macular degeneration. *Journal of ocular biology, diseases, and informatics* **3,** 53–59, doi: 10.1007/s12177-010-9055-1 (2010).
49. Stone, E. M. *et al.* Recommendations for genetic testing of inherited eye diseases: report of the American Academy of Ophthalmology task force on genetic testing. *Ophthalmology* **119,** 2408–2410, doi: 10.1016/j.ophtha.2012.05.047 (2012).
50. Stone, E. M. Genetic testing for age-related macular degeneration: Not indicated now. *JAMA ophthalmology* doi: 10.1001/jamaophthalmol.2015.0369 (2015).

Acknowledgement

We acknowledge NIH-EY021470 (AAF), Illinois society for the prevention of blindness (MAS), Research to Prevent Blindness, NY (Department of Ophthalmology, Northwestern University), The NUgene Biobank (Center for Genetic Medicine, Northwestern University), and *U01HG004609* and *U01HG006388* (NHGRI funded eMERGE I and II study, Northwestern University site).

Author Contributions

J.S., M.A.S. and A.F. wrote the main manuscript text, J.P., L.A., M.R., M.S. and M.G.H. prepared tables 1–3. All authors critically reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Simonett, J. M. *et al.* A Validated Phenotyping Algorithm for Genetic Association Studies in Age-related Macular Degeneration. *Sci. Rep.* **5**, 12875; doi: 10.1038/srep12875 (2015).