IDENTIFYING GENETIC PLEIOTROPY THROUGH A LITERATURE-WIDE ASSOCIATION STUDY (LITWAS) AND A DEEP PHENOTYPE ASSOCIATION STUDY (DEEPAS) WITH DATA FROM THE AGE-RELATED EYE DISEASE STUDY 2 (AREDS2)

A thesis submitted to the University of Arizona College of Medicine – Phoenix
in partial fulfillment of the requirements for the Degree of Doctor of Medicine

Michael Simmons
Class of 2017

Mentor: Zhiyong Lu, PhD

Acknowledgements

We acknowledge the contributions of Ayush Singhal PhD, Freekje VanAsten PhD, Tiarnan Keenan, FRCOphth, PhD, and Emily Chew MD in their support and indespensible contributions to this work. Particularly, Dr. Singhal designed the text mining algorithm and conducted the LitWAS, and Dr. VanAsten performed the principle work for execution of the DeePAS.

This research was supported by the NIH Medical Research Scholars Program, a public-private partnership supported jointly by the NIH and generous contributions to the Foundation for the NIH from the Doris Duke Charitable Foundation, the Howard Hughes Medical Institute, the American Association for Dental Research, the Colgate-Palmolive Company, and other private donors. No funds from the Doris Duke Charitable Foundation were used to support research that used animals.

Abstract

Background/Significance: Genetic association studies simplify genotype-phenotype relationship investigation by considering only the presence of a given polymorphism and the presence or absence of a given downstream phenotype. Although such associations do not indicate causation, collections of phenotypes sharing association with a single genetic polymorphism may provide valuable mechanistic insights. In this thesis we explore such genetic pleiotropy with Deep Phenotype Association Studies (DeePAS) using data from the Age-Related Eye Study 2 (AREDS2). We also employ a novel text mining approach to extract pleiotropic associations from the published literature as a hypothesis generation mechanism.

Research Question: Is it possible to identify pleiotropic genetic associations across multiple published abstracts and validate these in data from AREDS2?

Methods: Data from the AREDS2 trial includes 123 phenotypes including AMD features, other ocular conditions, cognitive function and cardiovascular, neurological, gastrointestinal and endocrine disease. A previously validated relationship extraction algorithm was used to isolate descriptions of genetic associations with these phenotypes in MEDLINE abstracts. Results were filtered to exclude negated findings and normalize variant mentions. Genotype data was available for 1826 AREDS2 participants. A DeePAS was performed by evaluating the association between selected SNPs and all available phenotypes. Associations that remained significant after Bonferroni-correction were replicated in AREDS.

Results: LitWAS analysis identified 9372 SNPs with literature support for at least two distinct phenotypes, with an average of 3.1 phenotypes/SNP. PheWAS analyses revealed that two variants of the ARMS2-HTRA1 locus at 10q26, rs10490924 and rs3750846, were significantly associated with sub-retinal hemorrhage in AMD (rs3750846 OR 1.79 (1.41-2.27), p=1.17*10^{-7}). This associated remained significant even in populations of participants with neovascular AMD. Furthermore, odds ratios for the development of sub-retinal hemorrhage in the presence of the rs3750846 SNP were similar between incident and prevalent AREDS2 sub-populations (OR: 1.94

vs 1.75). This association was also replicated in data from the AREDS trial. No literature-defined pleiotropic associations tested remained significant after multiple-testing correction.

Conclusions: The rs3750846 variant of the ARMS2-HTRA1 locus is associated with sub-retinal hemorrhage. Automatic literature mining, when paired with clinical data, is a promising method for exploring genotype-phenotype relationships.

Table of Contents

List of Figures and Tables

TABLES

FIGURES

Introduction:

Age-related macular degeneration (AMD) is a neurodegenerative disease that affects the elderly. The macula is the portion of the retina responsible for high-acuity vision so people with macular degeneration lose vision in the center of their visual field while typically retaining their peripheral vision. AMD is the most common cause of blindness in the elderly [1]. Advanced stages of the disease take two distinct forms: a "wet" or exudative form characterized by retinal neovascularization and a "dry" or non-exudative form characterized by geographic atrophy of the retinal pigment epithelium (RPE). AMD is a complex disease with strong genetic underpinnings as recently illustrated in a landmark genome-wide association study conducted by the International AMD Genomics Research Consortium [2].

The genome-wide association study (GWAS) has been instrumental in identifying genotype-phenotype relationships and bridging the gap in medical knowledge between genetic variations and human disease [3]. In conducting a GWAS, researchers compare the incidence of typically millions of genetic variants (also referred to as single-nucleotide polymorphisms or SNPs) between two cohorts of people who either do or do not have a specific phenotype, such as a disease like AMD. Since Klein et al conducted the first GWAS study and published their discovery of the association between the complement factor H (CFH) polymorphism and AMD in 2005 [4], the popularity of GWAS as an instrument for genetic research has risen steadily each year enabling discovery of new gene-disease associations in many fields [3].

One of the key limitations of GWAS is their inability to detect genetic pleiotropy. Genetic pleiotropy is the relationship of multiple phenotypes with a single gene. For example, type 2 diabetes mellitus (DM2) and obesity are pleiotropic traits since both are associated with genetic variants in the FTO gene [5]. Identifying genetic pleiotropy is useful in biology and medicine because pleiotropic phenotypes provide insight into the function of genes and the pathophysiology of disease. Unfortunately, the very nature of the GWAS study precludes an ability to perceive pleiotropic relationships. This is because GWAS' are conducted through case-control comparisons that require assembling new cohorts of people with a specific condition to act as 'cases' for each phenotype studied. Identifying genetic pleiotropy through GWAS studies thus requires performing multiple studies and recruiting separate populations of cases and controls. This is how researchers uncovered the above association between DM2 and obesity.

A new research construct, the phenome-wide association study (PheWAS) was pioneered by the eMERGE Network in 2010 [6]. PheWAS can be conceptualized as a “reverse GWAS” — they involve defining cohorts of cases and controls not by the presence or absence of a specific phenotype, but rather by the presence or absence of a specific genetic polymorphism. PheWAS compare the association of a large number of phenotypes with the presence or absence of a specific allele. Multiple studies have shown the ability of PheWAS to replicate established genetic associations, find novel associations, and identify genetic pleiotropy [7,8]. PheWAS are typically conducted using billing codes and electronic health record (EHR) data from biobanks or DNA-linked electronic health records (EHRs). One drawback of this approach is that billing code data lacks biologically relevant granularity even though deep, granular phenotypes have potential to provide unique insights into gene function and disease pathophysiology.

GWAS studies are also inherently limited in their ability to explore genetic associations with deep, granular phenotypes. This is because of the difficulties associated with assembling cohorts of patients with a shared, granular phenotype. For example, genetic associations with age-related macular degeneration have been explored with GWAS since 2005 [4], yet although AMD is a heterogeneous disease, the NHGRI GWAS catalog contains association information for only a few specific aspects of this disease: advanced AMD, choroidal neovascularization, geographic atrophy, and smoking interaction [9].

Clinical trial data is an alternative source of phenotype information for PheWAS with a unique complement of advantages and disadvantages in comparison with EHR billing code data. For example, clinical trial data is typically more complete than EHR data, with fewer biases. While clinical trial data does not usually contain the breadth of phenotype information that can be found in EHRs, clinical trials often do contain much greater depth regarding specific phenotypes. To date, only a few groups have used clinical trial data for performing PheWAS analyses [10].

In this thesis, we describe conducting a modification of a PheWAS, which we refer to as a DeePAS (Deep Phenotype-Association Study), using data from the Age-Related Eye Disease Study 2 (AREDS2) and investigating genetic associations with deep, granular phenotypes related to AMD. The term DeePAS in this setting is more appropriate than PheWAS since we do not investigate a comprehensive list of phenotypes, and yet the selection of phenotypes we do investigate contains much greater granularity than traditional PheWAS. We also introduce the concept of a literature-wide association study (LitWAS) using a novel text mining algorithm as a means of hypothesis generation. Our algorithm identifies separate descriptions of genotype-phenotype relationships in PubMed abstracts for overlapping genetic variants. These descriptions of pleiotropy come from multiple, often disparate studies and can only be regarded as hypothetical pleiotropic associations. We thus explore these LitWAS-generated hypotheses using DeePAS performed with the AREDS2 data.

Table 1 contains a list of key abbreviations and their definitions for this thesis.

Table 1. Common Abbreviations

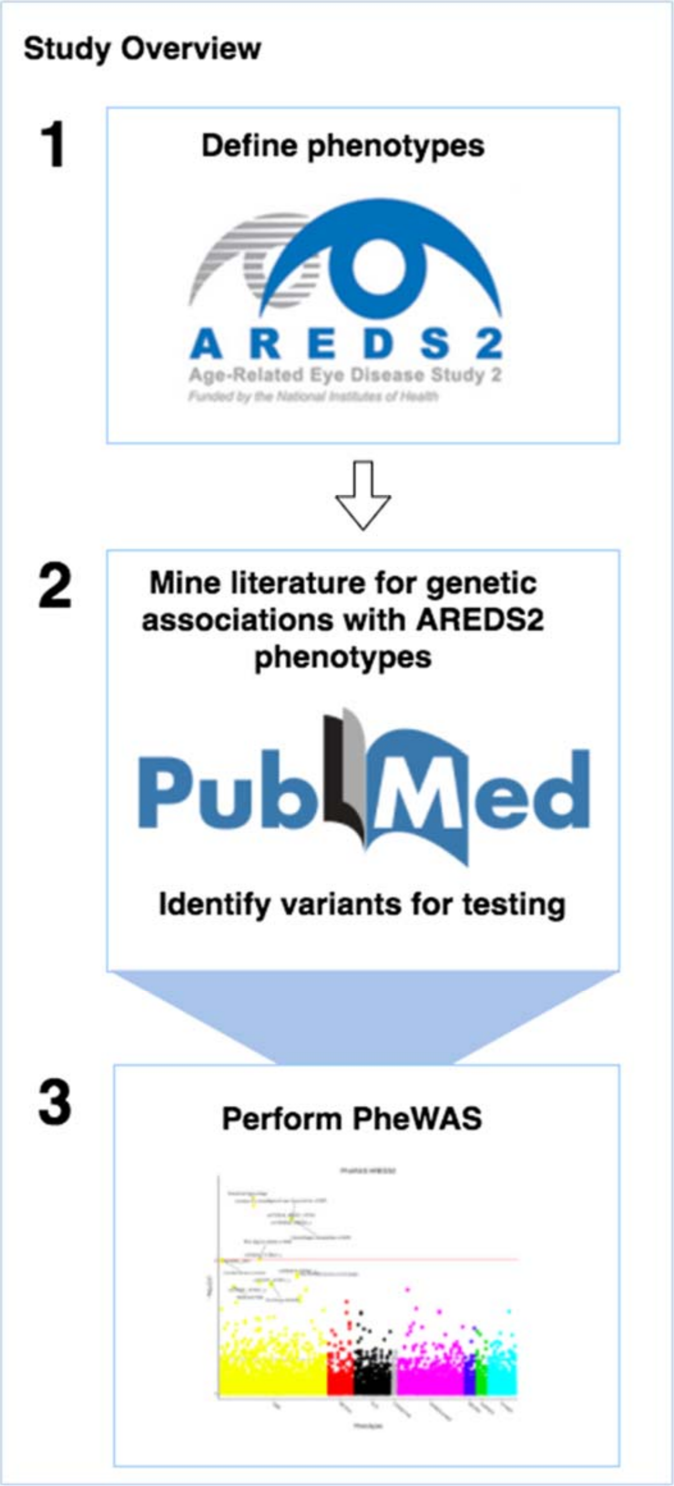
ABBREVIATION	DEFINITION
AMD	Age-related macular degeneration
AREDS	Age-Related Eye Disease Study
AREDS2	Age-Related Eye Disease Study 2
CNV	Choroidal neovascularization
DEEPAS	Deep Phenotype Association Study
GWAS	Genome-wide Association Study
LITWAS	Literature-wide Association Study
PCV	Polypoidal choroidal vasculopathy
PED	Pigment epithelial detachment
PHEWAS	Phenome-wide Association Study
RPE	Retinal pigment epithelium
RS#	Reference SNP number
SNP	Single-nucleotide polymorphism

Methods:

Overview

Figure 1 provides an overview of this study. Briefly, we first defined a set of deep and general phenotypes from AREDS2 study fields. We then mined the literature for genetic variants associated with those phenotypes to identify variants with novel, hypothetical pleiotropic relationships. We selected a number of such variants and tested these hypotheses with DeePAS using AREDS2 data.

Figure 1. Study Overview


Study Overview
1
Define phenotypes
AREDS2
Age-Related Eye Disease Study 2
Funded by the National Institutes of Health
2
Mine literature for genetic associations with AREDS2 phenotypes
PubMed
Identify variants for testing
3
Perform PheWAS

AREDS2 Study Design

The structure of the AREDS2 trial has been published previously [1112]. In brief, AREDS2 was a multicenter, phase III, randomized controlled clinical trial that investigated the efficacy and safety of lutein, zeaxanthin and ω-3 long-chain polyunsaturated fatty acid (LCPUFA) dietary supplements as a treatment for reducing the risk of progression of age-related macular degeneration [11]. The trial also investigated modifications of the initial AREDS supplement formulations, including a decrease in zinc dose and omission of β-carotene. The primary outcome measure for the trial was photographic progression to advanced AMD. All photos for the trial were read by a centralized reading center. A total of 4203 patients across 82 study sites within the United States participated in the trial. A number of these patients consented to contribute DNA samples to the AREDS2 biobank, and gave permission for their data to continue to be analyzed for research purposes.

AREDS Study Design

The structure of the Age-Related Eye Disease Study (AREDS) has also been published previously [13,14]. AREDS was double-masked clinical trial conducted across 11 centers that began in 1990 and concluded in 2001. The purpose of the trial was to evaluate the effect of antioxidant (vitamin C, vitamin E and beta carotene) and zinc supplements on the progression of AMD. Patients in the trial were randomized to receive daily oral vitamin tablets containing antioxidants, zinc, both antioxidants and zinc, or placebo. The primary outcomes were photographic progression to advanced AMD and at least moderate decrease in visual acuity from baseline. A total of 3640 participants were enrolled for an average of 6.3 years. A number of these patients also consented for genetic testing and for secondary analyses of their trial data.

Genotyping

The methods for DNA sampling, chip design, sequencing and gene annotation used in this study are described in detail elsewhere [2]. Briefly, the genetic sequence data for this study was generated as part of a large GWAS produced by the International AMD Genomics

Consortium. This study utilized a custom exome chip that was enriched for variants from known AMD loci. Actual sequencing was performed at a single center.

Phenotype definitions

Each “phenotype” for this study was created from one or more data fields from the AREDS2 trial. In the first step of defining phenotypes, manual reviewers selected study variables from the data collection forms of the AREDS2 trial that could correspond to potentially inheritable physical patient traits. Study-specific elements, such as logistical data about vitamin distribution and usage, were omitted. Two people then reviewed the list of field names and grouped redundant fields into a single phenotype, assigning the new phenotype group a unique identifier. For example, in AREDS2 it was possible for researchers to document a patient having coronary artery bypass grafting (CABG) in fields of three separate forms: the Cardiovascular Outcomes Study Report, the Hospitalization Report, and the Cardiovascular Adjudication Form. We grouped these three study fields together under a single identifier for “CABG”. Through this process, we generated a single list of potentially inheritable phenotypes for which data was available about AREDS2 trial participants.

After generating a list of phenotypes, two people normalized each field name where possible by identifying a corresponding phecode (using the Phecodes Database) and a corresponding SNOMED-CT term (using the SNOMED-CT online browser) for each phenotype. The final step of phenotype definition was then to organize the phenotype list into groups and hierarchies of related terms and to assign each phenotype a unique identifier.

Literature-wide Association Study

The core concept behind a LitWAS is that pleiotropic associations might be ‘hidden’ in the literature across multiple publications containing separate descriptions of associations of distinct phenotypes all of which share an association with a single, common polymorphism. We have previously described the creation of an algorithm for extracting relationships between genes, variations, and diseases from the literature [15,16]. To conduct a LitWAS, we applied this algorithm broadly across all abstracts in MEDLINE for each of the phenotypes that we crafted from the AREDS2 trial. We further refined our results through two additional steps: (1) normalization of different nomenclatures of variants and (2) identification of "negated" association descriptions or in other words, descriptions of associations that were found to *not* have a significant association. This section contains a brief description of our relationship extraction algorithm as well as an explanation of our methods for variant normalization and identification of negated descriptions.

Gene-variant-disease relationship extraction

The relationship extraction algorithm that we used in this LitWAS builds on the input of three open-source named-entity recognition and normalization tools: tmVar [17] (identifies variant descriptions), GNormPlus [18] (identifies gene descriptions), and DNorm [19] (identifies disease descriptions). Our algorithm takes the input of these tools and first utilizes a machine learning approach to identify relationships between descriptions of diseases and variants. The output of this step is a ranked list of references to MEDLINE abstracts that contain variant descriptions and are likely to be related to a target disease. The next step is to identify the names of the genes to which the variant belongs. We accomplished this by examining two sources of information outside the article under analysis: other abstracts in PubMed and the results of a Bing search query. This process produces a triplet that reflects a description in a given PubMed abstract of a gene, a variant of that gene, and a disease associated with that variant. As a final quality assurance step, our algorithm executes a sequence check and verifies that the identified variant does indeed fit in the sequence for the predicted gene. We applied this algorithm to all articles in English with an abstract that were returned by a query to PubMed with the phenotype term (e.g. "myocardial infarction", "reticular pseudodrusen", etc.). As our intent was to identify pleiotropic associations described across multiple abstracts, we included only variant-disease relationships where at least two distinct phenotypes shared a common SNP.

Normalization of variant descriptions

Mutation normalization refers to mapping various descriptions of a gene variant to a single standardized format. This is important because authors use many ways to describe genetic variants in the literature. For example, consider rs121913059, a variant in the Complement Factor H (CFH) gene, which has been shown across many studies to associate with the development of AMD. This variant results in an amino acid change from arginine to cysteine at amino acid position 1210. This variant may be described at protein level by this sequence change — p.R1210C — or at a DNA level by its DNA sequence alteration — g.100235C>T — or it may be described by its reference SNP number (rs#) — rs121913059. The Human Genome Variation Society (HGVS) has published recommendations that all variants should be described at the DNA level [20], but these recommendations are not uniformly followed, and even these may change over time as new builds of the reference human genomic sequence are released.

The mutation-tagging tool that we used — tmVar — in mining the literature can identify variant descriptions at all of the levels illustrated in the example above, but a major challenge that we faced in performing the LitWAS was to normalize these descriptions so that we could accurately detect pleiotropic associations across multiple articles. We chose to normalize or map all mutation descriptions to reference SNP numbers (rs#'s). To accomplish this, we used the publically available database, dbSNP as well as the local textual information within each PubMed abstract to assign the correct rs# for a mutation mentioned in the text.

In the first step of normalization, we execute a quality check to ensure that all extracted mutation descriptions contain all necessary components (e.g. a reference amino acid or nucleotide, the new AA or nucleotide or (in the event of a nonsense mutation) termination designation, and the position number of the mutation). Inconsistent mutations are not considered for normalization.

In the second step of normalization, we generate a list of possible variations of the extracted SNP and execute queries in dbSNP for these descriptions. For example, a protein-level variant, p|SUB|R|1210|C, extracted by tmVAR, would first be converted to its incomplete HGVS notation, p.Asp129Asn, and then the dbSNP database would be queried for all the rs# that contain this partial HGVS notation mentioned above. If the query returns an empty list (no search results) then we alter the SNP description by interchanging the position of the reference and mutant amino acids and execute a new query. The query process stops when a list of rs#'s is returned. The final step is then to match the associated gene from the query list with the gene from the initial extracted triplet.

We calculated separate precision values for our normalization of protein-level variant descriptions and our normalization of DNA-level variant descriptions. This involved creating a separate gold-standard dataset for evaluation of each type of variant. For each gold-standard test set, we utilized common syntactic patterns in abstracts for naming variants (e.g. protein variants are often named in PubMed abstracts using both their rs# and their protein sequence change) to construct the set.

Negation identification

Overview:

One source of frequent false positives in our gene-variant-disease relationship extraction algorithm is the presence in the literature of frequent descriptions of finding a lack of association between a disease and a given variant. For example, in PMID: 17305633, the authors investigate a possible association between the p.L7P polymorphism in the NPY gene and exudative age-related macular degeneration. They conclude, "There were no statistically significant differences in Leu7Pro polymorphism frequency between the exudative AMD and control cases." Our algorithm, which was designed to identify descriptions of relationships between gene variants and diseases correctly identifies that this abstract contains a discussion of the relationship between the NPY p.L7P polymorphism and exudative AMD, but it does not identify the sentiment of this discussion or the ultimate negative conclusions of the authors.

In order to identify and exclude negated association descriptions, we assembled and annotated a small corpus of abstracts containing negating phrases. We then built two dictionaries: one containing common negation words and a second containing negating phrases taken from the abstracts. We analyzed the phrases to identify syntactic patterns of negation. By screening abstracts for phrases that contain a negating term and follow common patterns of negation, we correctly identified abstracts where the authors reached negative conclusions and excluded them from the LitWAS.

Negation Corpus

The negation corpus for this task was curated by a fourth-year medical student who reviewed the results that were generated in validating our gene-variant-disease relationship extraction algorithm [15]. These results included all gene variants found in the literature for ten diseases: breast cancer, prostate cancer, pancreatic cancer, lung cancer, acute myeloid leukemia, Alzheimer's disease, hemochromatosis, age-related macular degeneration (AMD), diabetes mellitus, and cystic fibrosis. We selected only abstracts that contained negations without any assertions (i.e. we excluded abstracts that contained a mix of findings e.g. "variant x was not associated with disease y, but variant z was associated with disease y"). The final corpus contained a total of 103 abstracts. We then annotated each negating phrase in each abstract using the custom curation feature of the NCBI PubTator platform [21].

Negation phrase and term dictionaries:

We isolated all annotated negation phrases from the negation corpus and used a part-of-speech tagger to label each word (e.g. "noun", "verb", etc.) of each phrase. We then created separate dictionaries for each part-of-speech and added all the words from the negation phrases to their respective dictionaries (e.g. nouns were added to the "noun" dictionary, verbs to the "verb dictionary). At the same time, we collected a list of common negation words from several online forums and added these words to a separate "negation" dictionary.

To generate part-of-speech patterns, we evaluated each negation phrase. First, we executed a check to see if a given negation phrase contained one of the words from the negation dictionary. Negation words acted as "triggers" to initiate a part-of-speech pattern. Following each negation word, we sequentially aggregated the part of speech tags for the words following that negation until the first noun token was encountered or until the annotated negation phrase ended (whichever came first). These patterns of part-of-speech tokens went into a final dictionary of negation phrase patterns. As a final step, we removed all patterns that appeared only once in the negation corpus to isolate only the more generalizable patterns.

The process for evaluating abstracts for negated findings thus proceeds as follows: an abstract found to contain a gene-variant-disease triplet is first tokenized into sentences, and then each sentence is analyzed for the mutation under consideration. If a given sentence contains that mutation, it is next analyzed for negation trigger words. The presence of a negation word triggers a search for the presence of a strong negation pattern in the rest of the sentence. The words following the negation trigger are tokenized and tagged into their various parts of speech. If a strong negation pattern is found, then a check is executed to see if the words of the abstract that fit the negation pattern are contained in the various part-of-speech dictionaries assembled from the negation corpus. The final step in this process is a quality step to make sure that the identified negating phrase likely represents the variant-disease relationship initially identified by the relationship extraction algorithm. We check if the nearest disease to the variant in the sentence is the target disease name. If this is not true, then the negation is assumed to not refer to the variant-disease relationship in question and is ignored. If the sentence does not contain a disease name at all, we assume the negation phrase does refer to the variant-disease relationship under consideration and flag that relationship as negated.

Prioritization of SNPs for inclusion in the PheWAS

It is reasonable to expect that mining the literature for pleiotropic genotype-phenotype associations might return many more SNPs than could be tested via PheWAS without requiring large adjustments to significance thresholds to correct for multiple testing. We chose to build upon the work done by Fritsche et al in their recent publication of 34 genetic loci identified through a large GWAS to be independently associated with AMD [2]. In this study, we identified all AREDS2 phenotypes with published associations to these loci and tested these literature-defined pleiotropic associations. The remainder of the SNPs selected for testing were taken from a “convenience sample” of SNPs which were readily callable in our dataset. Six of these SNPs overlapped with SNPs identified through the LitWAS.

Phenotype Association Study

PheWAS consist of a series of statistical tests correlating the presence or absence of many phenotypes with the presence or absence of a specific genetic variant. The AREDS2 phenotypes in this study included both bivariate and continuous variables. We used the PheWAS statistical package in the R environment [22] to conduct logistic regression tests of bivariate variables and linear regression tests of the continuous variables. If an AREDS2 participant was ever recorded as having a given phenotype, we assumed the participant truly had that phenotype. We also assumed an autosomal dominant genetic model.

Results:

LitWAS

Normalization of variant descriptions

We assembled the evaluation dataset for protein-level variant descriptions by selecting all Medline abstracts posted between 1/1/2005 and 5/1/2015 that contained a specific pattern of variant description. In order to be included in the evaluation dataset, an abstract needed to contain a protein-level variant description with an rs# variant description immediately following it. We extracted only the first variant pair from each abstract. Since it is possible that two adjacent variants might not represent the same variant, a member of our team manually reviewed each protein-variant:rs# pair and discarded any incorrect pairs. Manual filtering also involved merging old rs#'s to the latest rs# as specified in dbSNP The final dataset contained 1672 unique pairings of a protein-level variant description with its accompanying rs#.

Construction of the DNA-level variant description evaluation dataset was similar to construction of the protein-level dataset, although it involved a different syntax patterns. We selected all PMIDs from the same date range above that contained two consecutive variant descriptions following one of two possible patterns: (1) "DNA-level_variant (rs#)"; or (2) "rs# (DNA-level_variant)". This selection process identified 1914 unique variant pairs. Manual, *ad hoc* analysis of these variant description pairs confirmed that each pair was a correct match so no further manual filtering was performed.

Table 2 displays the results from the evaluation of the performance our normalization algorithm on these two test sets.

Table 2. Variant Normalization Evaluation

Table 2. Variant Normalization Evaluation	
Abstracts in the protein-level variant evaluation dataset	1672
True positives (precision)	1588 (95%)
Abstracts in the DNA-level variant evaluation dataset	1914
True positives (precision)	1626 (85%)

Negation identification

We evaluated the performance of our negation identification algorithm using manual review of the output of a selection of results from evaluation of a test corpus. This test corpus included abstracts representing the same ten diseases in the negation training corpus. All abstracts used in this validation step were unique to the body of evaluation results that we reviewed in building the negation training corpus. The final negation evaluation corpus included one-hundred-ninety-six phrases that our algorithm predicted to contain a statement regarding a lack of association between a specific variant-disease pair.

One individual reviewed each phrase in the negation evaluation corpus and determined whether the phrase (1) contained a negation (a true positive); (2) contained an assertion (a false positive); (3) contained both a negation and an assertion (For example, in PMID: 21294239, the authors conclude that variants in the PPARGC1A gene are not associated with diabetes mellitus in Caucasian and East Asian populations, but are associated with Indian populations.); or (4) was not applicable (in every case, this was because of error in the variant tagger. For example, see PMID: 17680631, which describes the investigation of the effect of a drug, “S179D prolactin”, on prostate cancer cell growth. tmVAR incorrectly labeled this drug’s name as a protein variant description, i.e. “p.S179D”.) **Table 3** contains the results of this evaluation.

Table 3. Negation Identification Evaluation

Table 3. Negation Identification Evaluation	
Total number of phrases in the Negation Evaluation Corpus	196
Contains a negating phrase	160 (82%)
Contains an asserting phrase	11 (6%)
Contains both negating and asserting phrases	21 (11%)
Not applicable	3 (1%)

LitWAS

We identified 123 distinct phenotypes for which information was collected in the AREDS2 trial. A total of 9372 SNPs shared associations with at least two of these phenotypes. Considered individually, these pleiotropic SNPs shared, on average, 68% of the same references. In other words, for 68% of these pleiotropic sets, the authors were likely aware of the pleiotropic relationship that we identified via text mining. **Table 4** contains the results of the LitWAS and **Figure 2** contains a histogram displaying the number of pleiotropic associations per SNP. As expected, SNPs with fewer pleiotropic associations were far more common than SNPs with many associations. **Table 5** provides a snapshot of some of the pleiotropic associations identified for a single SNP, rs1040924, from the ARMS2 gene. The phenotypes in Table 4 were manually selected from 35 phenotypes identified through the LitWAS. Not all mined associations for this variant are presented. The OR, p-value, and Impact Phrase fields were manually extracted from the referenced articles.

Table 4. Literature-wide Association Study Results

Table 4. Literature-wide Association Study Results	
Total number of SNPs	9372
Total Genes	2884
Minimum number of pleiotropic associations shared by one SNP	2*
Average phenotypes per SNP	3.1
Maximum number of pleiotropic associations shared by one SNP	38

* This number was set as a requirement for inclusion in the LitWAS

Figure 2. Pleiotropic Associations per SNP

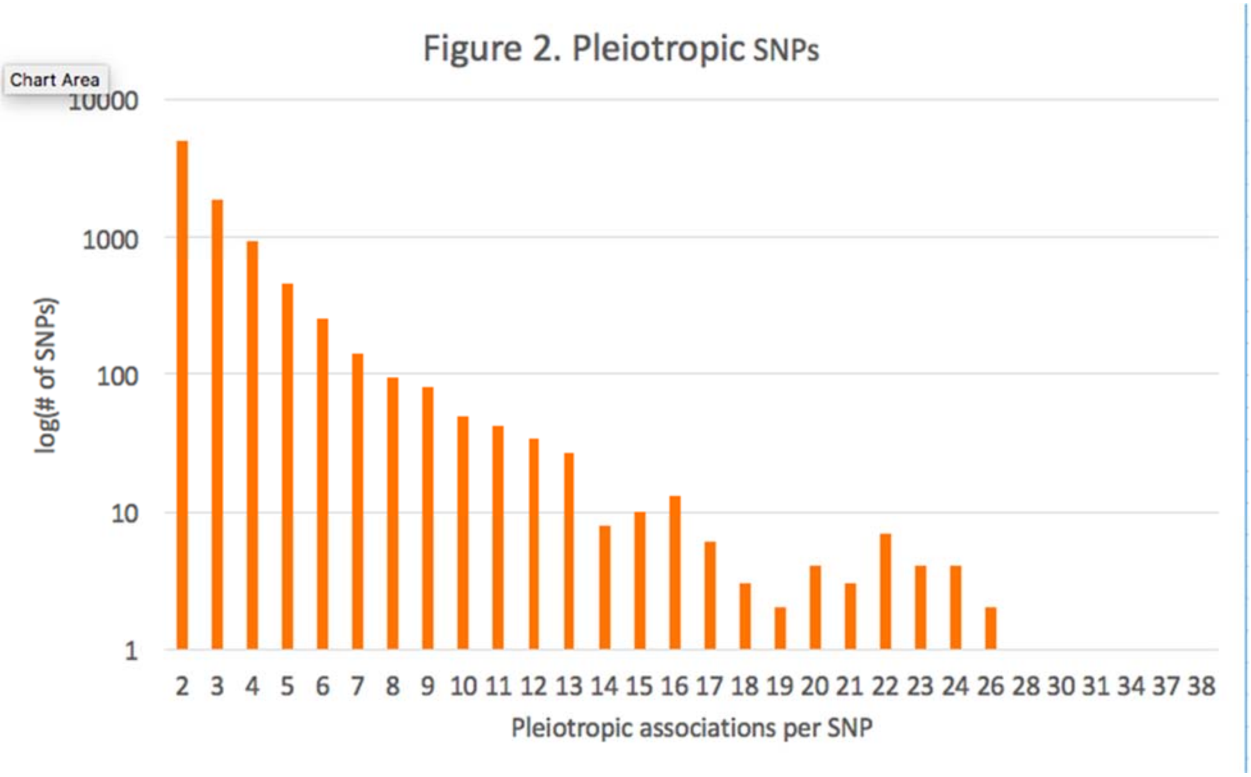


Table 5. Literature-mined associations with the ARMS2 variant, rs1040924

Table 5. Literature-mined Associations with the ARMS2 Variant, rs1040924				
Phenotype	PMID	OR	p-value	Impact Phrase
Exudative AMD	24362810	11.7	<0.0001	"ARMS2 … polymorphisms were associated with very high risk for exudative AMD"
Non-exudative AMD	20381870	6.01	<0.0001	"there was a significant association between genotype and presence of [geographic atrophy] for ... rs1040924"
Tobacco	26067391	8.33	---	"smoking synergistically increased the susceptibility of AMD for variants of [rs1040924]"
Vitreous hemorrhage	22509112	6.52	0.07	"Association of rs10490924 with … vitreous hemorrhage was reported in two studies"
Subretinal hemorrhage	21397333	12.4	0.0001	"[rs10494924] increased the likelihood for hemorrhagic PED by 12.4-fold [in polypoidal choroidal vasculopathy]"
Reticular pseudodrusen	24595987	3.23	0.008	"Logistic regression analysis revealed that … [the] T-allele at ARMS2 A69S … were risk factors for RPD"
Polypoidal Choroidal Vasculopathy	21397333	10.87	<0.0001	"The rs10490924 variant of the ARMS2 gene was shown to be associated with polypoidal CNV."
Serous retinal detachment	21397333	--	0.0092	"the at-risk allele of ARMS2 A69S was associated with a lower incidence of serous retinal detachment"

PheWAS

Patient characteristics

A total of 1826 AREDS2 participants gave consent for genetic testing and were included in this study. **Table 6** displays the characteristics of these study participants.

Table 6. AREDS2 Baseline Characteristics

Table 6. AREDS2 Baseline Characteristics	
With genetic testing	1826
Female	1071
White	1776
Never smoked	792
Some college education	881
Mean Age (standard deviation)	73.2 years (7.7 years)
Mean follow-up (standard deviation)	4.8 years (0.5 years)
AREDS2 treatment	
Control	446
Lutein/Zeaxanthin	449
DHA/EPA	481
Combination Therapy (L+Z+DHA+EPA)	450

Abbreviations: L, lutein; Z, zeaxanthin; DHA, docosahexanoic acid; EPA, eicosapentaenoic acid

PheWAS

Figure 3 displays a Manhattan plot of the results of the PheWAS conducted for all 33 loci included in this study. While many genotype-phenotype pairs had low p-values for significance of association, few remained significant after correcting for multiple testing. No pleiotropic pairs identified in the LitWAS achieved a significant association.

The ARMS2 variant identified in the study by Fritsche et al [2], rs3750846, was significantly associated with several phenotypes relating to retinal vascular integrity, the most significant of which was hemorrhage characteristic of AMD (p<0.0001). This variant is in complete linkage with the HTRA1-ARMS2 variant (rs1040924), which was identified in the LitWAS. We investigated the association between this variant and hemorrhage characteristic of AMD in data from AREDS2. **Table 7** contains the results of this investigation.

Figure 3. DeePAS results demonstrate significant associations with deep retinal phenotypes

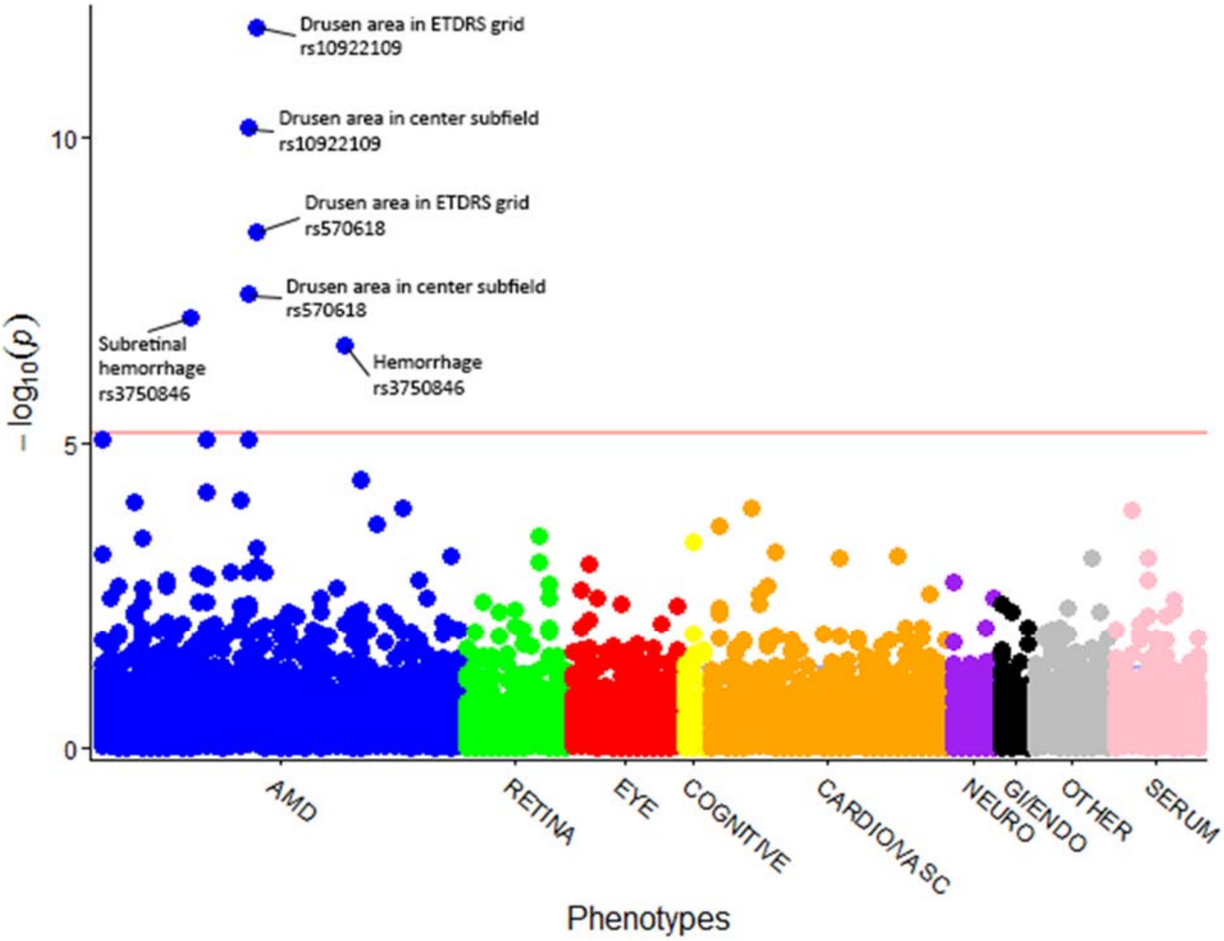


Table 7. HTRA1-ARMS2 (rs3750846) and subretinal hemorrhage in AREDS2

Table 7. HTRA1-ARMS2 (rs3750846) and subretinal hemorrhage in AREDS2			
AREDS2	ARMS2 +	ARMS2 -	Total
Hemorrhage (+)	303	120	423
Hemorrhage (-)	821	582	1403
Total	1124	702	1826
OR of Hemorrhage in the presence of ARMS2 variant (CI)			1.79 (1.41-2.27)
p-value			1.17 x 10^-6

The association between the HTRA1-ARMS2 variant and retinal hemorrhage in AMD was again highly significant. Taken in context, these values indicate that even in a cohort of patients who all had moderate to advanced AMD, the presence of the HTRA1-ARMS2 variant was associated with the development of hemorrhage.

Choroidal neovascularization (CNV) is a characteristic feature of exudative AMD, and its presence by itself might predispose to increased hemorrhage. We investigated the possibility that the association between the HTRA1-ARMS2 variant and hemorrhage was a byproduct of increased incidence of choroidal neovascularization by looking at a subpopulation of 855 patients in AREDS2 who had CNV. In this group, the increased predisposition to hemorrhage persisted (OR = 1.75 (1.3-2.3); p<0.001).

It is also possible that this association might have been due to a recruiting bias or other confounding characteristic of the AREDS2 study design. We investigated this possibility by examining a population of patients who all developed CNV during the course of the trial and comparing the odds ratios between those who had CNV at the outset of the trial and those who developed CNV only after the trial began. Among patients (n=249) whose neovascular condition only manifest after the recruitment process, the association between the HTRA1-ARMS2 variant remained constant in magnitude and significance (OR=1.94 (1.1-3.5); p=0.031). The odds ratio in *incident* populations who developed CNV (OR=1.94) was thus very similar to the odds ratio in *prevalent* populations who developed CNV (OR=1.75), giving strong support to our assumption of no bias in study recruitment.

The AREDS trial patient population was distinct from AREDS2, and it included patients with early to late stages of disease as well as controls. We defined similar subpopulations within AREDS2 and AREDS, and tested the association between rs3750846 and hemorrhage characteristic of AMD in each to compare its effects across these two distinct patient populations. The results of this investigation are displayed in **Table 8**.

Table 8. ARMS2 (rs3750846) and Hemorrhage Characteristic of AMD in AREDS and AREDS2

Table 8. ARMS2 (rs3750846) and Hemorrhage Characteristic of AMD in AREDS and AREDS2				
Patient Population	Hemorrhage	%	OR (CI)	p-value
AREDS (n = 2889)				
Large drusen at baseline and CNV at any time (n=507)	403	79.5	1.74 (1.2-2.6)	0.004
Large drusen at baseline and CNV developed during follow-up (n=308)	190	61.7	1.57 (1.0-2.5)	0.068
CNV at any time (n=748)	507	67.8	1.78 (1.3-2.5)	<0.001
CNV developed during follow-up (n=368)	226	61.4	1.69 (1.1-2.6)	0.019
AREDS2 (n= 1826)				
Large drusen at baseline and CNV at any time (n=286)	83	29.0	1.58 (0.9-2.7)	0.097
Large drusen at baseline and CNV developed during follow-up (n=224)	62	27.7	2.0 (1.0-3.8)	0.038
CNV at any time (n=855)	423	49.4	1.75 (1.3-2.3)	<0.001
CNV developed during follow-up (n=249)	70	28.1	1.94 (1.1-3.5)	0.031

As displayed in the table, the association between the HTRA1-ARMS2 variant and hemorrhage characteristic of AMD reached statistical significance ($p<0.05$) in AREDS and AREDS2 participants who developed CNV at any point or in follow up. In the subpopulations of both trials where the association between hemorrhage and the HTRA1-ARMS2 variant did not reach statistical significance, the sample sizes were relatively low and the direction of association was consistent with findings in the other subpopulations. Furthermore, the odds ratio for the effect of rs3750846 remained very consistent throughout all patient populations.

Discussion

The ARMS2 variant that we identified in our LitWAS (rs10490924) and the HTRA1-ARMS2 variant identified in the study by Fritsche et al [2] (rs3750846) are both part of the 10q26 locus, which overlies three genes: High Temperature Requirement A (HtrA) serine peptidase 1 (HTRA1), Age-related Maculopathy Susceptibility 2 (ARMS2), and Pleckstrin Homology Domain Containing, Family A Member 1 (PLEKHA1). This locus has a strong association signal with AMD [23], but because of the physical proximity of these genes, there has historically been difficulty determining which has the strongest association with AMD [24,25]. The recent large GWAS conducted by the International AMD Genomics Consortium excluded involvement of PLEKHA1 in risk predisposition to AMD [2].

The ARMS2 gene (which is sometimes designated more specifically by the gene name, LOC387715) has two exons. These encode a 16 kD protein that appears to localize to the outer mitochondrial membrane [26]. It has been proposed that the gene plays a role in protecting photoreceptors from oxidative damage so that variants of the gene may leave photoreceptors susceptible to degeneration over time [26]. The HTRA1 gene has 9 exons and encodes a serine protease that targets, among other things, the extracellular matrix [27]. Numerous studies have established a role of both genes in the development of AMD [26,28–30]. Some investigators have suggested that these two genes may interact in exerting their influence on the development of AMD [31].

Despite much research into the mechanism behind the association with these genes and AMD, great uncertainty remains. Several researchers have made connections between ARMS2 and the retinal vasculature. Hu et al recently published a meta-analysis of the association of the ARMS2 gene variant with positive response to anti-angiogenic treatment (e.g. response to anti-VEGF injections or photodynamic therapy) [32]. Yoneyama et al also described an association between the ARMS2 locus and the development of reticular pseudodrusen in exudative AMD [33]. To the best of our knowledge, the specific association between the ARMS2 gene and hemorrhagic phenotypes in AMD has not yet been described, although, as detected in our LitWAS, an association between ARMS2 and the development of subretinal hemorrhage, hemorrhagic pigment epithelial detachment (PED), and serous PED has been described in polypoidal choroidal vasculopathy (PCV) [34]. PCV is a condition involving the development of choroidal vascular abnormalities that typically develops in middle age. It is distinct from AMD in that it typically afflicts patients at younger ages and involves choroidal vascular abnormalities much earlier in the disease course than AMD. Nevertheless, the development of similar phenotypes in association with the ARMS2 variant is likely significant.

It was surprising that so few of the pleiotropic relationships that we identified across published abstracts in our LitWAS were replicated in the PheWAS. We had expected to see some of the literature-identified associations appear in our own data from AREDS2. Our algorithm did identify abstracts supporting an association between rs1040924 and subretinal hemorrhage (in PCV), and the PheWAS did identify a significance between drusen development and subretinal hemorrhage in people with this variant. Nevertheless, the majority of the phenotype associations with rs1040924 were not replicated. These negative results highlight some of the limitations of secondary data mining. It is difficult to say without further investigation whether the lack of significance of the literature defined pleiotropic relationships is due to the particular patient population in the AREDS2 trial, undetected biases in our data, inadequate sample sizes, or a true lack of pleiotropy.

With further refinement of our literature mining approach, it is possible that we might identify pleiotropic associations more accurately. For example, in our LitWAS, we made no attempt to characterize or identify the methodology by which authors established a specific association. Likewise, we did not account for the study population in the abstracts we mined, although we understood that the AREDS and AREDS2 trials were both conducted with largely Caucasian cohorts.

There were several other limitations to our study. Our LitWAS methods did not allow us to readily identify pleiotropic associations identified by authors in a single study. Such articles would be of particular interest because they would represent a stronger genotype-phenotype link than one simply created by overlapping gene variant mentions across abstracts. Our LitWAS methodologies completely ignore such references.

Another limitation inherent in mining relationships from published literature is that many published associations come with qualifications. For example, it is common for authors to establish a genotype-phenotype association only in certain populations or under certain conditions. We attempt to address this problem by identifying and excluding negations, and by favoring associations that are mentioned multiple times in our prioritization process Nevertheless, these measures are likely not adequate to distinguish strong, generalizable findings from those that are intended only for a narrow audience. Likewise, we did nothing in our LitWAS to account for the magnitude of association (i.e. by identifying odds ratios for specific associations) between genotypes and phenotypes in the literature.

Text mining in general has several potential applications in modern medicine with the primary text sources being the biomedical literature (as in this study) and electronic medical records [35]. This work represents one of the first applications of text mining the literature in ophthalmology epidemiological research.

Future Directions

The LitWAS work in this thesis could be expanded upon with the use of greater sophistication in normalizing and prioritizing variants. The previous discussion of the HTRA1-ARMS2 variant illustrates this avenue well. In our methods, we simply normalized variants with different descriptions to a single rs#. Although this process was necessary and we accomplished it with accuracy, the final output of our LitWAS did not acknowledge the strong linkage disequilibrium between ARMS2 variants and HTRA1 variants. Neither did we use any methods of lending more weight to genotype-phenotype associations that were shared between multiple variants of the same gene. Integrating a map of genetic linkage into our analysis or even simply weighting genotype-phenotype associations from the same gene differently could reasonably be expected to help improve the accuracy of the prioritization process. The output of our LitWAS could in fact be considered as a network of genetic variants and the phenotypes that connect them. Using network analysis algorithms to identify global or local similarity between sets of genes and phenotypes could be a more promising avenue to prioritize genetic variants most likely to replicate in AREDS2 data than the ad hoc methods that we used in this study [36,37].

Regarding the novel association between the HTRA1-ARMS2 locus and hemorrhagic phenotypes in AMD, it should be emphasized that we used secondary data analysis of AREDS2 data in establishing this association. Nevertheless, the fact that this association is replicated in the AREDS cohort as well lends confidence that this association is more than chance occurrence. Additional prospective research is merited to establish a role of ARMS2 in choroidal vascular integrity.

Conclusions

In this thesis, we explore genotype-phenotype relationships — particularly those relating to AMD and ocular diseases — through two related avenues: secondary analysis of clinical trial data and text mining biomedical literature. There are several major contributions of this work: (1) the integration of large-scale text mining of biomedical literature with clinical epidemiological research; (2) the application of the PheWAS construct to data from a large clinical trial, which allowed for evaluation of associations between candidate SNPs and deep phenotypes in AMD; and (3) the identification of a novel association between the HTRA1-ARMS2 locus and the deep phenotype of hemorrhage in AMD.

References

1. Smith W, Assink J, Klein R, et al. Risk factors for age-related macular degeneration: Pooled findings from three continents. *Ophthalmology*. 2001;108(4):697-704.

2. Fritsche LG, Igl W, Bailey JNC, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 2016;48(2):134-143.

3. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*. 2012;90(1):7-24.

4. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308(5720):385-389.

5. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol*. 2012;8(12):e1002823.

6. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26:1205-1210.

7. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31:1102-1111.

8. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet*. 2016;17(3):129-145.

9. Tony Burdett, Emma Hastings, Dani Welter, SPOT, EMBL-EBI, NHGRI. GWAS Catalog. http://www.ebi.ac.uk/gwas/search?query=age-related%20macular%20degeneration. Accessed June 10, 2016.

10. Moore CB, Verma A, Pendergrass S, et al. Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. *Open Forum Infect Dis*. 2015;2(1):ofu113.

11. AREDS2 Research Group, Chew EY, Clemons T, et al. The Age-Related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology*. 2012;119(11):2282-2289.

12. Age-Related Eye Disease Study 2 (AREDS2) Research Group, Chew EY, Clemons TE, et al. Secondary analyses of the effects of lutein/zeaxanthin on age-related macular degeneration progression: AREDS2 report No. 3. *JAMA Ophthalmol*. 2014;132(2):142-149.

13. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study

(AREDS): design implications. AREDS report no. 1. *Control Clin Trials*. 1999;20(6):573-600.

14. Age-Related Eye Disease Study Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Arch Ophthalmol*. 2001;119(10):1417-1436.

15. Singhal A, Simmons M, Lu Z. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Comput Biol*. 2016;12(11):e1005017.

16. Singhal A, Simmons M, Lu Z. Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. *J Am Med Inform Assoc*. 2016;23(4):766-772.

17. Wei C-H, Harris BR, Kao H-Y, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*. 2013;29(11):1433-1439.

18. Wei C-H, Kao H-Y, Lu Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *Biomed Res Int*. 2015;2015:918710.

19. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*. 2013;29(22):2909-2917.

20. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*. 2016;37(6):564-569.

21. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*. 2013;41(Web Server issue):W518-W522.

22. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014;30(16):2375-2376.

23. Jakobsdottir J, Conley YP, Weeks DE, Mah TS, Ferrell RE, Gorin MB. Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet*. 2005;77(3):389-407.

24. Kortvely E, Ueffing M. Gene Structure of the 10q26 Locus: A Clue to Cracking the ARMS2/HTRA1 Riddle? *Adv Exp Med Biol*. 2016;854:23-29.

25. Bergeron-Sawitzke J, Gold B, Olsh A, et al. Multilocus analysis of age-related macular degeneration. *Eur J Hum Genet*. 2009;17(9):1190-1199.

26. Kanda A, Chen W, Othman M, et al. A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration. *Proc Natl Acad Sci U S A*. 2007;104(41):16227-16232.

27. Grau S, Richards PJ, Kerr B, et al. The role of human HtrA1 in arthritic disease. *J Biol Chem*. 2006;281(10):6124-6129.

28. Dewan A, Liu M, Hartman S, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*. 2006;314(5801):989-992.

29. Yang Z, Camp NJ, Sun H, et al. A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. *Science*. 2006;314(5801):992-993.

30. Fritsche LG, Loenhardt T, Janssen A, et al. Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA. *Nat Genet*. 2008;40(7):892-896.

31. Yang Z, Tong Z, Chen Y, et al. Genetic and functional dissection of HTRA1 and LOC387715 in age-related macular degeneration. *PLoS Genet*. 2010;6(2):e1000836.

32. Hu Z, Xie P, Ding Y, Yuan D, Liu Q. Association between variants A69S in ARMS2 gene and response to treatment of exudative AMD: a meta-analysis. *Br J Ophthalmol*. 2015;99(5):593-598.

33. Yoneyama S, Sakurada Y, Mabuchi F, et al. Genetic and clinical factors associated with reticular pseudodrusen in exudative age-related macular degeneration. *Graefes Arch Clin Exp Ophthalmol*. 2014;252(9):1435-1441.

34. Sakurada Y, Kubota T, Imasawa M, et al. Role of complement factor H I62V and age-related maculopathy susceptibility 2 A69S variants in the clinical expression of polypoidal choroidal vasculopathy. *Ophthalmology*. 2011;118(7):1402-1407.

35. Simmons M, Singhal A, Lu Z. Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health. *Adv Exp Med Biol*. 2016;939:139-166.

36. Jia P, Zhao Z. Network.assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum Genet*. 2014;133(2):125-138.

37. Liu X, Yang Z, Lin H, Simmons M, Lu Z. DIGNiFI: Discovering Causative Genes for Orphan Diseases Using Protein-protein Interaction Networks. In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB ’16. New York, NY, USA: ACM; 2016:527-527.