ELECTRONIC HEALTH RECORD PHENOTYPING TO FACILITATE THE CATEGORIZATION OF GENETIC VARIANTS OF UNCERTAIN SIGNIFICANCE

By

Jennifer A. Pacheco

A CAPSTONE PROJECT

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

March 2020

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

Jennifer Allen Pacheco

"*Electronic Health Record Phenotyping to Facilitate the Categorization of Genetic Variants of Uncertain Significance*"

Has been approved

Capstone Advisor
Michael Mooney, PhD

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I acknowledge and sincerely thank the following groups and people:

The NUgene project, especially the Director Maureen Smith, and Washington University School of Medicine McDonnell Genome Institute, Tess Pottinger and Megan Puckelwartz for the WGS cohort genetic data, and for their moral support and feedback. Maureen inspired me to pursue the topic of this project and Tess gave me the data I needed.

The eMERGE project, Baylor College of Medicine Genetics Laboratories, Rex Chisholm, Maureen Smith, Laura J. Rasmussen-Torvik, Christie Hoell, Adam Gordon for the eMERGEseq cohort genetic data, and for their moral support and feedback. Also other eMERGE colleagues including Hana Zouk for her insight and expertise, Iftikhar Kullo and others for their inspiring work on familial hypercholesterolemia. Last but not least, Abel Kho who introduced me to the field of biomedical informatics and encouraged me to pursue this degree, and for his continued encouragement.

Beth Wilmot and Michael Mooney for their invaluable mentorship, advice, feedback, and patience as Capstone advisors. Without them this would not have been possible.

All other faculty and staff in the Department of Medical Informatics and Clinical Epidemiology at the Oregon Health & Science University, for their support and for their rigorous academic program in biomedical informatics which has taught me so much. Especially Bill Hersh as both the director and my first teacher in the program; Diane Doctor for the advice and taking care of countless details; and Vishnu Mohan and Judy Logan whom I had for multiple classes, not just because they taught some required classes, but because they were great teachers. In fact, all of the teachers I had in this program were good.

All other members of my department where I work at Northwestern University, the Center for Genetic Medicine, for moral support and insight.

My husband, Cesar Pacheco, for both his moral support and for taking care of everything from cooking to cleaning to being my chauffer all the years I was working on this degree and project. I have been a part-time student from the beginning of our courtship throughout our entire marriage. Also my other family members and friends, especially my parents and my in-laws, and my best friend Jin-hee Bae, for their moral and other support, and especially for listening to me go on and on about this project, and for cajoling me to keep going.

ABSTRACT

Objective

Many genetic variants are of unknown significance (VUS). Efficient and accurate electronic health record (EHR) phenotyping, having facilitated genome-wide association studies, could identify patients with VUSs who exhibit phenotypic features that might indicate pathogenicity of those variants. Identifying and following up with these patients could improve their healthcare, and assist in improving genetic variant categorization.

Methods

Subjects (N=3860) were recruited at Northwestern Medicine for 2 studies and genotyped on 2 separate platforms. Each study's platform genotyped the 3 genes containing variants that collectively explain ~40% of diagnosed cases of familial hypercholesterolemia (HC) (FH). Rare variants in these genes were queried for pathogenic/likely pathogenic (P/LP), conflicting interpretations of pathogenicity (CPV), and VUS classifications; and unreported non-synonymous variants (URV) were noted. Four EHR phenotype algorithms of varying complexity were implemented: 2 algorithms, for primary HC (PH) and FH; a subset of the PH algorithm: maximum low-density lipoprotein (LDL) without recurring high triglycerides (high LDL); and ICD diagnosis codes, grouped into phecodes for HC. The distribution of genetic variants, the distribution of the phenotypes found by the algorithms, and the overlap thereof, was examined. VUSs found in multiple subjects were further scrutinized to assess HC phenotypes in those subjects, and variant characteristics, to tentatively determine whether those VUSs lean toward being pathogenic or benign.

Summary of Results

Of the 24 patients with P/LP FH variants, 21 were found by any algorithm to have HC in the EHR. Furthermore, just over half of the patients with VUSs, CPVs, or URVs, but no P/LP variants, were found to have HC. As expected, the phecode algorithm (the simplest), found the most patients with HC with P/LP variants (21), or other queried variant types. Both the high LDL and PH algorithms found a similar number of HC patients with P/LP variants or other variants, although overall PH found more than the high LDL algorithm. As expected, the FH algorithm (the strictest), found the least number of patients with HC with P/LP variants (3) or other variants (15).   The phecode algorithm found all patients having P/LP variants with evidence of HC. However, both the phecode and PH algorithms were needed to identify all of the patients having VUSs, CPVs and/or URVs with evidence of HC. Thus, for patients with FH genetic variants, both of those algorithms are needed to identify patients for diagnostic evaluation. Twenty-five VUSs were identified in more than 1 subject. For 21 of these VUSs, the vast majority of subjects had evidence of HC in their EHR, indicating those VUSs may be pathogenic; thus, the 82 subjects who had those VUSs should be further investigated.

Conclusion

With further assessment, these methods, combined with other data, could be used to identify phenotypes in patients with VUSs, URVs, or CPVs, which in turn could facilitate the functional categorization of those variants as either pathogenic or benign.

Chapter 1 INTRODUCTION

Background

Variants of unknown significance (VUSs) in genes are often incidentally found with genetic testing (Kalia et al.; Safarova, Klee, et al.). These variants have unclear implications for patients and physicians, and the problem of interpreting the consequences of genetic variants is becoming more urgent as broader genetic testing is becoming more popular (Kalia et al.; Hooper et al.). We need more efficient ways to check for clinical manifestations of relevant diseases in large numbers of individuals with VUSs, especially in genes for which the American College of Medical Genetics and Genomics (ACMG) recommends returning secondary findings of pathogenic or likely pathogenic (P/LP) variants (Kalia et al.). Validated electronic health record (EHR) phenotyping algorithms could be used to discover clinical manifestations of VUSs efficiently in larger populations. Given the relative success of EHR phenotyping for extracting phenotypic data for genome-wide association studies (GWAS) (Rasmussen-Torvik et al.; Kho, Pacheco, et al.; Pathak et al.; Gottesman et al.; Klarin et al.), using published algorithms, or subsets thereof, could also be used to identify patients with phenotypes possibly associated with VUSs. In order to test these algorithms, we need to explore which offers the greatest yield in terms of detecting clinical manifestations of disease in those with a VUS, while minimizing time and effort to extract phenotypic data from the EHR, to determine which phenotypes patients have.

Genomic sequencing for clinical care is becoming more prevalent, and from this more VUSs are found, which are not known if clinically actionable (Kalia et al.; Iacocca et al.). Thus, we need an efficient way to check for any clinical significance of these VUSs. A pipeline using biomedical informatics, specifically data mining of clinical data (from the EHR), could be used to discover possible disease associations with VUSs in genes already known to have P/LP variants that cause the given disease. For example, some studies (Safarova, Klee, et al.; Chora et al.) have re-classified VUSs, in genes with variants that cause familial hypercholesterolemia (FH), as likely pathogenic, or likely benign. Another example is cancer studies that have revealed VUSs in BRCA1 and BRCA2 genes whose classification changed to be benign (Kast et al.). Similarly, for cardiac disease it has been difficult to determine pathogenicity of genetic variants associated with cardiomyopathies (Ackerman).

EHR data has been used to determine if patients have relevant phenotypes for GWAS (Rasmussen-Torvik et al.; Pacheco et al.; Jeff et al.; Kho, Hayes, et al.; Klarin et al.). Conversely, EHR data has been used for phenome-wide association studies (PheWAS) (Denny, Ritchie, et al.; Denny, Bastarache, et al.), to discover what other phenotypes patients have for genetic variants known to cause disease. Many genetic variants are VUS, and determining pathogenicity of genetic variants is important yet can be difficult; thus, efficient yet accurate EHR phenotyping could facilitate identifying patients for follow-up, and possibly subsequent genetic variant categorization. In particular if multiple patients have the same VUS and most either do, or do not, exhibit a phenotype associated with the gene in which the VUS occurs, then those VUSs would warrant further investigation for possible categorization of either likely pathogenic or likely benign.

Familial hypercholesterolemia (FH) was primarily selected for this pilot study because of the significantly larger number of patients in our study populations with the FH phenotype and/or pathogenic variants in genes known to cause FH. More importantly, FH confers a high risk of premature coronary artery disease (Wierzbicki et al.; Kramer et al.; Lan et al.; Akioyamen et al.) and studies have shown that a significant number of patients with FH go undiagnosed and/or untreated, with an estimated 1 in ~250 people having FH (Kramer et al.; Banda et al.; Myers et al.). In particular, a recent study found only about half of patients found to have FH via genetic testing were on a statin and even less were diagnosed with FH before the testing was done (Abul-Husn et al.), and suggested that analyzing EHR data could be used to uncover these un- or under-diagnosed patients for treatment. Also, phenotyping for FH can be done using mostly structured data in the EHR which can be abstracted without manual review of patients' charts, as the diagnosis is mostly based on high low-density lipoprotein levels (LDL), personal history of cardiovascular disease, plus sometimes also family history and/or specific physical symptoms (xanthomas and corneal arcus) (Humphries et al.; Séguro et al.).

In addition, genetic testing is recommended for patients with suspected FH (Sturm et al.; Stein et al.), yet, there are many VUSs in FH genes (Calandra et al.), ranging anywhere from 10% to over 40% of variants in FH genes reported to ClinVar, depending on the gene (Iacocca et al.). Furthermore, when pathogenic variants are found in genes known to cause FH, there is clear action to take, namely, cholesterol lowering drugs and in extreme cases, lipopheresis (Wierzbicki et al.). Sometimes, when genetic testing is done based on suspicion of FH, VUSs are considered to warrant such action by the clinician. Furthermore, with broader genetic testing being done without any specific diagnosis as an indication for the testing, VUSs cannot necessarily be interpreted in the same way (Ambry Genetics). For example, a recent study showed great "variability" (Safarova, Klee, et al.) in classifying FH variants, and another found many variants suspected of causing FH that do not have evidence of functional change to warrant pathogenic classification (Chora et al.). Furthermore, a PheWAS of genetic variants associated with FH found that phecodes related to lipid disorders were associated with those variants as expected, and were also associated with a non-lipid disorder (Safarova, Satterfield, et al.). A more recent study found that VUSs in genes associated with cardiac disorders, including FH, occurred more frequently in African-Americans (Pottinger Tess D. et al.), warranting further investigation of these VUSs.

Lastly, EHR phenotyping methods are evolving such that phenotype algorithms, usually created in laborious collaboration between informaticists and clinicians, can be instead created using more efficient methods such as by using phecodes (Bastarache et al.; Safarova, Satterfield, et al.; Denny, Bastarache, et al.; Denny, Ritchie, et al.) or machine learning (Pathak et al.; Hripcsak and Albers; Beaulieu-Jones and Greene; Liao et al.; Yu et al.). For example, recently a machine learning algorithm was developed to identify patients with possible FH in the EHR by training and testing a random forest classification algorithm against known cases, which successfully identified three-quarters of patients with known FH (Banda et al.).

Objective

The overall objective of this study is to evaluate phenotyping algorithms that use clinical data from EHRs to identify subjects with known genetic risk factors for hypercholesterolemia (i.e. P/LP variants in FH genes). Subsequently we aim to use those same algorithms to identify subjects with evidence of HC and VUSs in the FH genes. In addition, VUS that occur in more than one subject, where a majority of the subjects with the same VUS either have HC, or not, will be prioritized for follow up as these VUSs are more likely to be P/LP, or benign/likely benign (B/LB), variants. This will then allow us to meet the next objective which is to identify subjects with VUSs in FH genes and with HC for further investigation and potential follow-up to confirm their phenotype, and to improve their healthcare if they have an HC phenotype and are not adequately treated.

Subsequently, the phenotypic findings for these VUSs, especially in VUSs in > 1 subject, should be reported to ClinVar to assist in a final objective, a clearer categorization of the variants.

Thus, we will assess how to use more efficient phenotyping methods to find potential correlations between genotypes, especially LP/P variants and VUSs, and phenotypes. The purpose of this pilot study is to test different methods of extracting and analyzing diagnostic data (via database queries, i.e., not chart review) from the EHR to find diseases associated with VUSs, especially those in ACMG genes.

Our hypothesis is that there is a tradeoff between using common, easier to extract data versus more difficult to extract, less common, data from the EHR to characterize patients with P/LP or VUS variants in selected genes, as either having disease manifestations associated with those genes or not. Specifically, we expect to find a higher percentage of patients with P/LP variants, and possibly some VUS in genes, especially VUSs occurring in more than 1 subject, with evidence of disease, by using a broader phenotype algorithm that uses more data from the EHR. However, simply using phecodes (Denny, Ritchie, et al.; Denny, Ritchie, et al.) may still find a similar percentage of patients, which could be good enough. Also, it is also expected that if most of these patients are older, for there to be some evidence of FH, or at least HC, if they have FH. Lastly, given the previously observed penetrance of up to 96% (Kullo et al.; Kullo), the majority should have FH, or at least HC, if they have P/LP variants in FH genes.

Furthermore, some VUSs may be more likely to be pathogenic if seen in > 1 subject that manifested the expected phenotype. However, most P/LP variants, including those in FH genes, are not 100% penetrant (Shah et al.; Kullo et al.); therefore, not all subjects with P/LP variants are expected to have manifestations of the disease.

Assumptions

1. EHR data is sufficient to find at least some evidence of a given phenotype, including more structured but potentially less accurate data, such as billing codes, if the data is used appropriately (i.e., codes grouped into logical phenotype codes such as phecodes for PheWAS studies) (Denny, Bastarache, et al.). However, EHR data likely contains an incomplete record of a subject's health and healthcare, especially at an academic medical center like Northwestern Medicine (NM) where this study is being conducted, where subjects may only seek tertiary care. Thus, many subjects may only see the clinicians most likely to diagnosis FH, such as primary care and/or cardiology clinicians, outside of NM.

2. Genotypic assumption(s): Only variants with a significant minor allele frequency (MAF) have been studied (Pottinger Tess D. et al.; Zouk et al.), and each variant has been studied individually (negating the need for burden testing). Not all subjects with P/LP variants in FH genes will have the HC disease phenotype, as the currently known P/LP variants in FH genes are not 100% penetrant (Shah et al.; Kullo et al.; Kullo).

3. Phenotypic assumption(s): For data from the EHR (and not the phecodes derived from billing data), using existing algorithms, or parts thereof, will be faster than consulting with clinicians to determine phenotypic inclusion and exclusion criteria; however, we will likely still need to consult with clinicians to check if the results of any of the algorithms make sense clinically.

In summary, it is expected that there will be 1 or more phenotyping algorithms that are better at identifying HC in the EHR, and although not all subjects with P/LP variants in

FH genes will have HC, most will; thus, the same algorithms can be used to identify subjects with HC who also have VUSs. This would allow for the identification of subjects and VUSs for follow-up, for proper treatment for subjects with these VUSs that might be P/LP, and for eventual clearer categorization of these VUSs as either P/LP or B/LB.

Chapter 2 MATERIALS AND METHODS

Study population

Patients (N=3860) were recruited at NM, an academic tertiary hospital/healthcare system, for two studies. Participants were genotyped on two separate platforms (see details below for each study). DNA variants in three genes, *LDLR, APOB,* and *PCSK9*, across all participants, were examined, given the evidence that variants in these genes collectively explain ~40% of diagnosed cases of FH (Sharifi et al.).

The first study was a selection of 894 NM patients enrolled in the NUgene DNA biobank at Northwestern University (NU), who were selected by the NUgene team to have whole genome sequencing (WGS) (Pottinger Tess D. et al.); thus, this cohort will be referred to as the WGS cohort. They had to be NM patients seen in an NM clinic in the last few years at the time of the study, who did not already have broad genotyping. Furthermore, patients were selected such that approximately half were male, and such that the distribution of self-reported minority races/ethnicities approximated the following proportions: African American (~40%), Caucasian (~20%), and/or Hispanic/Latino (~40%) race/ethnicity, in order to conduct subsequent genetic studies of minorities who are typically under-represented in such studies (Pottinger Tess D. et al.). Approximately one-quarter were also selected for having one of 4 phenotypes: atopic dermatitis (AD, N=95), cardiomyopathy (N=56), cancer (N=118), using International Classification of Diseases (ICD) diagnosis codes, Current Procedural Terminology (CPT) codes, and/or medications; and chronic rhinosinusitis (CRS, N=180), using a previously published algorithm (Hsu et al.).

The second study, a part of the eMERGE (electronic Medical Records and Genomics) project's phase III Return of Results study (Zouk et al.), called the "Genetic Testing and your Health" study at NM, recruited ~3,000 NM patients for genetic testing to be returned to them and their clinical providers to study both patients' and providers' perceived utility of broader genetic testing. The genotyping panel called eMERGEseq was created by the eMERGE network, to contain the 59 aforementioned ACMG genes (Kalia et al.), plus ~41 other genes and some selected single nucleotide polymorphisms (SNPs) selected by the network to be of interest for the purposes of the study, including some for pharmacogenomics (Zouk et al.). Thus, this cohort will be referred to as the eMERGEseq cohort.

A subset of the patients in both cohorts were recruited by the respective NUgene and eMERGE teams, for specific indications and/or specialty clinics, including a lipid disorder clinic (N=283), of which a majority (N=246) were recruited for the eMERGEseq cohort. In this lipid clinic, some patients who already had hypercholesterolemia were selected, but also spouses and relatives of those patients without hypercholesterolemia were recruited. Other specific indications included a few other phenotypes, especially those related to the genotypes being extracted for eMERGEseq, such as cardiac diseases, and certain cancers, as mentioned previously.

Genotyping methods

Both studies consented patients and collected blood samples from which DNA was extracted, using NU Institutional Review Board approved protocols. In addition, only the eMERGEseq cohort used Clinical Laboratory Improvement Amendments (CLIA) standards in CLIA certified labs (Baylor Genetics Laboratories) (Zouk et al.), as it was the only study of the 2 to return the genetic testing results to the patients in the study. For the WGS cohort, rare variants in these genes were queried in ClinVar (National Center for Biotechnology Information, U.S., *ClinVar*) for (P/LP, which were grouped together) and VUS classifications, by Pottinger et. al. (Pottinger Tess D. et al.).

For the WGS cohort, the WGS was performed at McDonnell Genome Institute at Washington University, then researchers at NU used ClinVar designations to determine pathogenicity of genetic variants (Pottinger Tess D. et al.). These included VUSs, and variants with conflicting interpretations of pathogenicity (CPVs) in ClinVar, which although they are technically VUSs, they are also a separate category from VUS: CPVs are VUSs that have multiple conflicting interpretations reported to ClinVar which include reports of pathogenicity. Unreported variants (URVs, not reported to ClinVar) that were nonsynonymous substitutions (variants that cause a change in an amino acid which thus results in changes to the protein) were also collected by Pottinger et. al. (Pottinger Tess D. et al.).

For the eMERGEseq cohort, genotyping was done by Baylor Genetics Laboratories who determined variant designation by using ACMG guidelines, and by consulting ClinVar, and other sources, such as disease indication for patients with prior disease indicated as the reason for testing. Other sources expert curators also used to determine the classifications included published literature, and phenotypic information from the EHR. VUSs included the types of genetic mutations listed in Supplemental Table 1, and URVs included non-synonymous coding variants plus other mutation types, also listed in Supplemental Table 1. There were also no CPVs in eMERGEseq because the expert curators at Baylor lab investigated questionable variants by requesting and reviewing additional relevant EHR data, to resolve the conflict and assign CPVs, and sometimes

VUSs, as either P/LP, VUS, or B/LB. Lastly, Baylor also provided details on the exact mutations; thus, for this cohort we examined whether any VUSs occurred in more than one participant.

Phenotyping methods

The EHR was examined using four phenotype algorithms of varying complexity (Figure 1), to determine whether subjects had hypercholesterolemia (HC), and if so, which type they might have. Two algorithms were developed by Mayo Clinic (Safarova, Liu, et al.) and subsequently used by the eMERGE network for primary HC (PH) and FH (*Electronic Health Record-Based Phenotyping Algorithm for Familial Hypercholesterolemia | PheKB*). This algorithm defined FH cases as having “definite” or “probable” FH per Dutch Lipid Clinic Network (DLCN) criteria (Séguro et al.); thus, those that are “possible” FH (3-5 points using DLCN) (Séguro et al.) are categorized as just having PH, and grouped with those with “unlikely” FH (0-2 DLCN points) (Séguro et al.) into 1 PH category. Thus, those who had “possible” FH were also calculated separately for this pilot study. In addition, a subset of the PH algorithm was used to determine maximum low-density lipoprotein (LDL) per patient, specifically selecting the maximum LDL, without making any adjustments for lipid lowering treatments, for patients who did not have abnormally high triglycerides on more than 1 occasion, as these laboratory measures are more common and more easily extracted from the EHR.

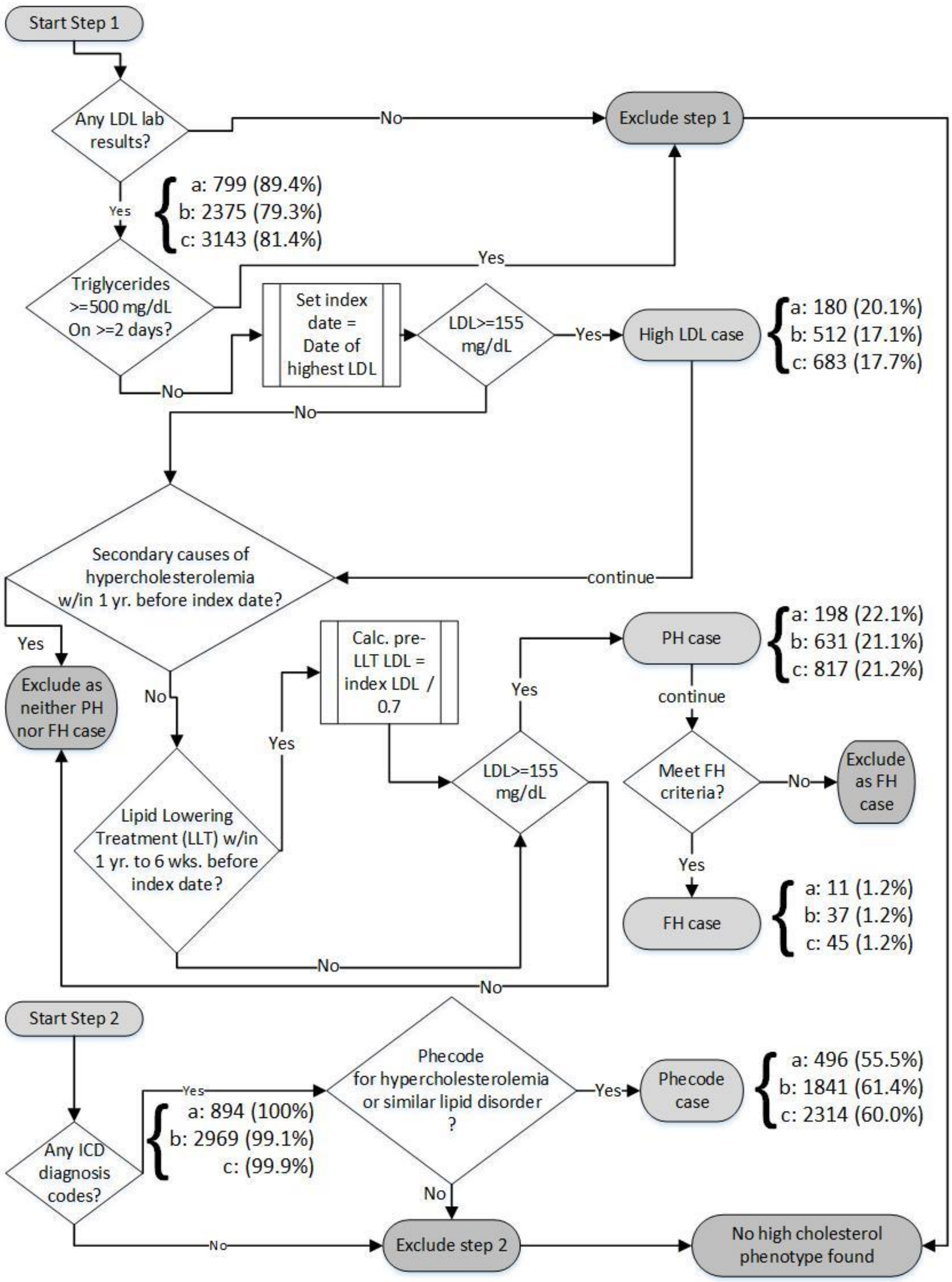
Start Step 1
Any LDL lab results?
No
Exclude step 1
Yes
a: 799 (89.4%)
b: 2375 (79.3%)
c: 3143 (81.4%)
Triglycerides >=500 mg/dL On >=2 days?
Yes
No
Set index date = Date of highest LDL
LDL>=155 mg/dL
Yes
High LDL case
a: 180 (20.1%)
b: 512 (17.1%)
c: 683 (17.7%)
No
continue
Secondary causes of hypercholesterolemia w/in 1 yr. before index date?
Yes
Exclude as neither PH nor FH case
No
Lipid Lowering Treatment (LLT) w/in 1 yr. to 6 wks. before index date?
Yes
Calc. pre-LLT LDL = index LDL / 0.7
No
LDL>=155 mg/dL
Yes
No
PH case
a: 198 (22.1%)
b: 631 (21.1%)
c: 817 (21.2%)
continue
Meet FH criteria?
No
Exclude as FH case
Yes
FH case
a: 11 (1.2%)
b: 37 (1.2%)
c: 45 (1.2%)
Start Step 2
Any ICD diagnosis codes?
Yes
a: 894 (100%)
b: 2969 (99.1%)
c: (99.9%)
Phecode for hypercholesterolemia or similar lipid disorder ?
Yes
Phecode case
a: 496 (55.5%)
b: 1841 (61.4%)
c: 2314 (60.0%)
No
No
Exclude step 2
No high cholesterol phenotype found

Figure 1. Phenotyping Algorithms

EHR phenotyping algorithms for simple high LDL without recurring high triglycerides, primary and familial hypercholesterolemia (PH and FH) (Safarova, Liu, et al.), and simply any hypercholesterolemia phecode. Note these case types are not mutually exclusive, i.e. a patient can be more than 1 type of hypercholesterolemia case. Number of subjects that meet each case criteria and percentage of the respective cohort are shown: a=WGS cohort, b=eMERGEseq cohort, c=both cohorts combined (note there is overlap between the cohorts, i.e., there are 31 subjects in both cohorts); LDL=Low-density lipoprotein.

Lastly, for the fourth algorithm, phecodes used for PheWAS were mapped to ICD diagnosis codes (using maps 1.2 and 1.2b1 from PheWAScatalog.org) (Wei et al.; P. Wu et al.) found in the EHR. Then phecodes (Table 1), which are groupings of ICD diagnosis codes into phenotypes for genetic association studies, shown to be associated with hypercholesterolemia or hyperlipidemia, were selected from a 2019 PheWAS of genetic variants in the 3 FH genes (Safarova, Satterfield, et al.), which map to the following ICD-codes shown in Table 1. For this algorithm, subjects were noted as have HC if they had at least 1 of these phecodes (Table 1).

If patients did not have evidence of HC in their EHR, it was assumed they are either healthy, or their EHR is lacking sufficient data to determine HC status. Thus, the number and percentage of patients with any LDL lab results was noted as LDL measurements are what is used to diagnosis HC, and if a patient does not have any LDL results, their HC phenotype status may not be able to be determined. Also, earliest age of diagnosis for any of these disorders was calculated based on the dates of diagnosis in the EHR, as shown in Table 1, or earliest LDL measure >= 155 mg/dL.

		corresponding ICD-9-CM codes		corresponding ICD-10-CM codes	
phecode	**phenotype**	**code**	**description**	**code**	**description**
272	Disorders of lipoid metabolism	272	Pure hypercholesterolemia		
		272.1	Pure hyperglyceridemia		
		272.2	Mixed hyperlipidemia		
		272.3	Hyperchylomicronemia		
		272.4	Other and unspecified hyperlipidemia		
		272.9	Unspecified disorder of lipoid metabolism		
272.1	Hyperlipidemia	272	Pure hypercholesterolemia	E78.4	Other hyperlipidemia
		272.1	Pure hyperglyceridemia	E78.5	Hyperlipidemia, unspecified
		272.2	Mixed hyperlipidemia		
		272.3	Hyperchylomicronemia		
		272.4	Other and unspecified hyperlipidemia		
272.11	Hypercholesterolemia	272.*	Pure hypercholesterolemia	E78.0	Pure hypercholesterolemia
				E78.00	Pure hypercholesterolemia , unspecified
				E78.01	Familial hypercholesterolemia
272.13	Mixed hyperlipidemia	272.2	Mixed hyperlipidemia	E78.2	Mixed hyperlipidemia

Table 1. Phecode to ICD diagnosis code map

Phecodes used to determine diagnosis of hypercholesterolemia, associated with variants in FH genes in PheWAS (Safarova, Satterfield, et al.)

Due to the clinical diagnosis of FH first requiring a patient to be diagnosed with PH, and due to the multiple types of data (lab test results, diagnosis codes) that can be used to determine patients' phenotypes, note that patients can have more than 1 hypercholesterolemia phenotype: in particular, a patient can have both PH, FH, and high LDLs, and phecodes for those disorders. Furthermore, patients can have several mutations or variants in a single gene or across the 3 FH genes. Thus, when comparing patients' phenotypes and genotypes, mutually exclusive groups of patients were created in a hierarchical fashion, such that the most severe phenotype (FH) and severe genotype (P/LP) were first selected, and then excluding those patients already selected with the most severe, patients with the next most potentially severe phenotype (PH without FH) and genotype (CPVs or VUSs without any P/LP variants) were grouped, followed lastly by the least severe phenotype and genotype (VUS or URVs).

Data Analysis and Validation

First, to assess if there were statistically significant differences between the demographic characteristics of the 2 cohorts, both chi-squared and Fisher's exact tests (using KNIME Analytics Platform, available from knime.com) were performed to compare gender, plus each race and ethnicity, between cohorts. In addition, Mann–Whitney–Wilcoxon tests (using KNIME Statistics Nodes (Labs), available from knime.com) were performed to test if the current ages, and ages at first diagnosis or first high LDL for subjects with phenotypic evidence of HC, differed significantly. Mann–Whitney–Wilcoxon tests were used as the distribution of those ages, especially in the eMERGEseq cohort, were skewed: most of subjects were older. Note that because the 31 subjects who are in both cohorts comprise only approximately 10% of the total number of subjects, those subjects were counted in both cohorts when comparing them, and the tests were also performed by excluding those 31 subjects from both cohorts, to make sure they did not affect the results.

Secondly, results of previous validations of the FH algorithm were gathered, and validation on the algorithms used in this study was performed where possible, to assess the accuracy of the algorithms. Specifically, when Mayo originally developed the FH algorithm, they reviewed the charts of approximately 100 cases and 100 controls, resulting in sensitivity, or recall, of 97%; specificity of 94%; positive predictive value (PPV), or precision, of 94%; and negative predictive value (NPV) of 97% (Safarova, Liu, et al.). Then, when Mayo subsequently validated the algorithm for use across the eMERGE network, by reviewing the charts of an additional 58 cases and 42 controls, the PPV of the algorithm was 100%, and the NPV was 98% (Mayo Clinic). In addition, another eMERGE site, Geisinger Health System, validated Mayo's FH algorithm on 25 cases and 25 controls, resulting in a PPV of 100% and NPV of 96% (Geisinger). Although the algorithms used were validated in previous studies, additional validation was performed where possible. Specifically, for the eMERGEseq cohort, de-identified outcomes forms were completed via chart review by the eMERGE network to assess patients' disease status before and after the return of P/LP variants, including both FH and PH. This included assessing if patients received lipid-lowering treatment for PH or FH before and/or after the return of those results. Thus, for additional validation, outcomes data, for patients with P/LP variants in FH genes, was searched to confirm if the subjects with penetrant disease (not all were expected to have FH given previously observed penetrance of < 100%) are found using the above phenotyping methods, to calculate accuracy statistics including precision (PPV) and recall (sensitivity).

The following analyses were performed to analyze if diagnostic data extracted from the EHR can be used to find possible correlations with genetic variants known to cause disease (P/LP genetic variants). Further analyses were done to subsequently test, if, using the same phenotyping methods of EHR data, we can find any genetic VUSs in patients with those diseases, or vice versa. First, we assessed the overlap of patients with P/LP FH variants and VUSs in the FH genes, and patients with HC according to the different algorithms, and then vice versa.

Specifically, first, descriptive statistics were calculated including percentages of patients that have any manifestation of disease using each of the phenotyping methods outlined above. Then, differences in percentages between the different types of EHR data extraction methods, as outlined above, were compared. Secondly, descriptive statistics were generated, for patients with P/LP genetic variants and VUSs (from eMERGE or Clinvar), and for the WGS cohort, also for CPVs and URVs, of which phenotypes they have, and vice versa, to assess overlap of the phenotypes and genotypes.

In addition, to verify patients had sufficient EHR data to be able to determine the presence or absence of hypercholesterolemia, whether patients had any non-zero LDL lab results were noted. As only ICD-10 (not ICD-9) has a specific diagnosis code for FH, and having more recent encounter(s) with diagnoses indicates recent health assessment, it was also noted whether patients had any ICD-10 diagnosis codes in any clinical encounters within the last ~5 years (specifically since October 1, 2015, when ICD-10 was mandated for use in EHRs the United Status in order to qualify for meaningful use reimbursement) (Bert et al.). Furthermore, subjects with the ICD-10 diagnosis code specific to FH (E78.01), were compared to subjects who had FH according to the algorithms, for further tentative confirmation of patients having FH.

Finally, for the eMERGEseq cohort which had detailed information on the VUSs in FH genes subjects had, those VUSs that occurred in more than 1 subject were further analyzed by examining the number of subjects with each of those VUSs, and the phenotypes they had from the EHR (and if they even had any relevant EHR data). In addition, the type of mutation by gene was also examined, as previous studies have shown penetrance of the FH phenotype to vary by gene and, of course, by type of mutation (i.e. the more severe the effect of the mutation the more likely FH will occur) (Shah et al.; Kullo et al.; Kullo). Lastly, the number of VUSs, CPVs, and URVs were compared across races and ethnicities, as previous studies have seen that minorities have a larger proportion of VUSs, compared to the majority Caucasian population, in both cancer and cardiac phenotypes (Pottinger Tess D. et al.; Landry and Rehm; Slavin et al.; Caswell-Jin et al.).

Chapter 3 RESULTS

Table 2 summarizes the demographic characteristics of the two study cohorts, including the ages when patients with H.C. were first found to have H.C. in the EHR. For those with no evidence of H.C. using the phenotyping methods described herein, current ages are shown. Subjects in the WGS cohort are slightly younger, and have more diverse races/ethnicities than those in the eMERGEseq cohort due to the targeted selection of minorities for the WGS cohort; in fact, all of the demographics are significantly different between the 2 cohorts (p-values <= 0.03 from chi-squared and Mann–Whitney–Wilcoxon tests).

	WGS		**eMERGEseq**		
	N	**%**	**N**	**%**	**p-values**
Total number of patients	894		2995		
Sex (Male)	307	34.3%	1156	38.6%	2.1103E-02
Hispanic/Latinx*	340	38.0%	180	6.0%	1.4655E-134
Race (Caucasian)*	610	68.2%	2406	80.3%	2.7379E-14
Race (African)*	343	38.4%	405	13.5%	1.8979E-61
Race (Asian)*	18	2.0%	152	5.1%	8.5205E-05
Race (Native American/Alaska Native)*	70	7.8%	31	1.0%	3.6428E-29
Race (Pacific Islander/Hawaiian)*	10	1.1%	14	0.5%	2.9142E-02
	mean	**SD**	**mean**	**SD**	
age at diagnosis for subjs w/ phenotype	47.2	10.5	58.1	11.7	0.0000E+00
current age for subjs w/o phenotype	48.4	10.3	51.4	15.8	3.7603E-03

Table 2. Demographics by cohort

Demographic characteristics of both cohorts in this study. *Some patients reported more than 1 race/ethnicity; thus, the total for all races/ethnicities is >100%. p-values are from chi-squared and Mann–Whitney–Wilcoxon test, as described in the methods.

There are coincidentally 31 subjects who are in both cohorts, and for those subjects the genotype results are generally the same between the 2 studies, with the following exceptions. First, all 31 have URVs in FH genes reported for the WGS study, that Baylor did not report in the eMERGEseq study, as eMERGE did not report URVs. In addition, because Baylor resolved CPVs to be either P/LP, B/LB, or VUSs, 1 out of 13 subjects with CPVs reported in the WGS study were classified in the eMERGEseq study by Baylor as VUS, and the rest (N=12) were classified as B/LB. Lastly, there is one subject who has a VUS in the WGS study that is not reported in the eMERGEseq study, again, because Baylor resolved CPVs and VUS variants in the eMERGEseq study where possible.

Figure 1 (in the Methods section above) shows the number of patients found by each of the algorithms, in addition to the number of subjects who had any LDL lab results or any ICD diagnosis codes (from an encounter or in the problem list, in the last 5 years) in their NM EHR. Approximately 80% of subjects in both cohorts had at least 1 LDL measure, and 17-20% had HC according to the high LDL algorithm. Furthermore, almost all had at least 1 ICD-10 diagnosis code from a recent (since 2015) encounter, and 55-61% had phecodes for HC. Lastly, just over 20% had evidence of PH, and just over 1% had evidence of FH.

For algorithm validation, the results of comparing the chart review results of the 18 subjects in the eMERGEseq cohort with P/LP variants, to the results of the phenotype algorithms for those subjects, are shown in Supplementary Tables 2a-d. Of the 18 subjects for whom there were outcomes forms in the eMERGEseq cohort, 17 were confirmed in the outcomes forms to have HC if not PH or FH, and 1 was confirmed to not have HC at all. Furthermore, all 17 subjects confirmed to have HC via chart review were found to have HC by at least 1 algorithm, and no algorithms found the 1 patient confirmed to not have to HC, to have HC. Only the phecode algorithm was 100% accurate, with 100% precision and recall; however, the PH algorithm also performed well with 82.4% recall and 100% precision, resulting in an overall accuracy of 83.3%. The FH algorithm had the worst performance with 55.6% accuracy overall, only 22.2% recall, and 66.7% precision. Lastly, the high LDL algorithm was more accurate than the FH algorithm, with 66.7% accuracy and 100% precision, but with 64.7% recall. Also, as part of that validation, it was found that of the 18 eMERGEseq cohort subjects with P/LP variants in FH genes, half (N=9) had an FH diagnosis before genetic testing, and of those that did not have an FH diagnosis before, 5 had an HC diagnosis; thus, the remaining 4 did not have a diagnosis of HC before testing, according to their abstracted outcomes data from eMERGE. Furthermore, 15 of these 18 subjects were on an LLT before genetic testing, 1 after, and thus 2 did not receive LLT before or after receiving the genetic test results.

In addition, across both cohorts, only 18 subjects had the ICD-10 code for FH and have both HC phenotype and FH genotype data, and of those half (N=9) have “definite,” “probable,” or “possible” FH and were also found by the high LDL algorithm; 5 (27.7%) have PH; and all had phecodes for HC. Furthermore, of the 18 subjects with the ICD-10 code for FH, 5 (27.7%) have P/LP variants in FH genes; 3 (16.7%) have VUSs in FH genes; and 4 (22.2%) have URVs in FH genes. In addition, as mentioned previously, 283 subjects across both cohorts were enrolled from a lipid disorders clinic, and of those, 270 (95.4%) have both an HC phenotype in the EHR and an FH genotype, of which 80

(29.6%) have "definite," "probable," or "possible" FH; 114 (42.2%) have PH; 90 (33.3%) have high LDL; and all but 1 subject had phecodes for HC (and that 1 subject had an FH genotype but no evidence of an HC phenotype).

Figures 2a-b summarize the number of patients found to have HC and/or FH via the 4 algorithms, stratified by genetic variant classification, and Figures 3a-b summarize the number of patients with the different types of genetic variants, stratified by phenotype algorithm classification, by showing the overlap of the algorithms as Euler or Venn diagrams. There are separate figures for each cohort, where a = WGS cohort and b=eMERGEseq cohort.

To determine whether the phenotyping algorithms were truly identifying the relevant subjects, we examined how many subjects in each phenotypic category had known risk variants (i.e. P/LP variants). From both cohorts, of the 24 patients with P/LP FH variants, 21 were found to have HC via any algorithm. All of the 3 that did not have HC did not have any LDL lab results in their NM EHR and one did not even have any problem list diagnoses in the EHR; thus, those patients may have HC that is not recorded in their NM EHR. Of those 21 found to have any HC, only 3 had "definite" or "probable" FH, but, of those that did not have ("definite" or "probable") FH, 9 had PH with "possible" FH per DLCN criteria, and all had phecodes for HC. As seen in Figure 2a, in the WGS cohort (N=894), only 4 of the 6 subjects with P/LP variants had HC and none of them were found to have FH. However, 2 of the 4 with HC actually had "possible" FH according to DLCN criteria, and of those 2 without HC evidence in their EHR, both had no LDL lab results in their EHR.

Given the number of subjects classified as FH or PH who did not have P/LP variants, but did have VUSs, we next examined which phenotypic criteria subjects with those variants met. As the analysis of the WGS study resulted in additional variant types (CPVs and URVs) that are not P/LP nor B/LB, other than VUSs, those were also examined separately. In particular, in the WGS cohort seen in Figure 2a, 324 patients with variants with conflicting interpretations of pathogenicity were found to have HC, and of the 246 not found to have HC, all but 57 did have LDL lab results; of those with HC, 7 had "definite" or "probable" FH while 44, of the 110 with PH, had "possible" FH, for a total of 51 with "possible," "probable," or "definite" FH; and 214 or approximately two-thirds of those found to have HC were only found using phecodes. Of the 9 patients with VUSs (but not VUSs that are CPVs), 8 patients with VUSs were found to have HC, 5 had PH, of which 2 had "possible" FH, and one-third (N=3) were found with phecodes only, and the 1 patient not found to have HC did have an LDL lab result and other diagnoses from recent encounters recorded in their NM EHR. Furthermore, of the remaining 308 subjects with only URVs, approximately one-third (N=130) had no HC phenotype, although 26 of those patients had other diagnoses but no LDL lab results; 4 had FH; approximately one-quarter (79) had PH, of which 33 had "possible" FH, for a total of 37 with "possible," "probable" or "definite" FH; and approximately one-third (N=98) were found to have HC only using phecodes.

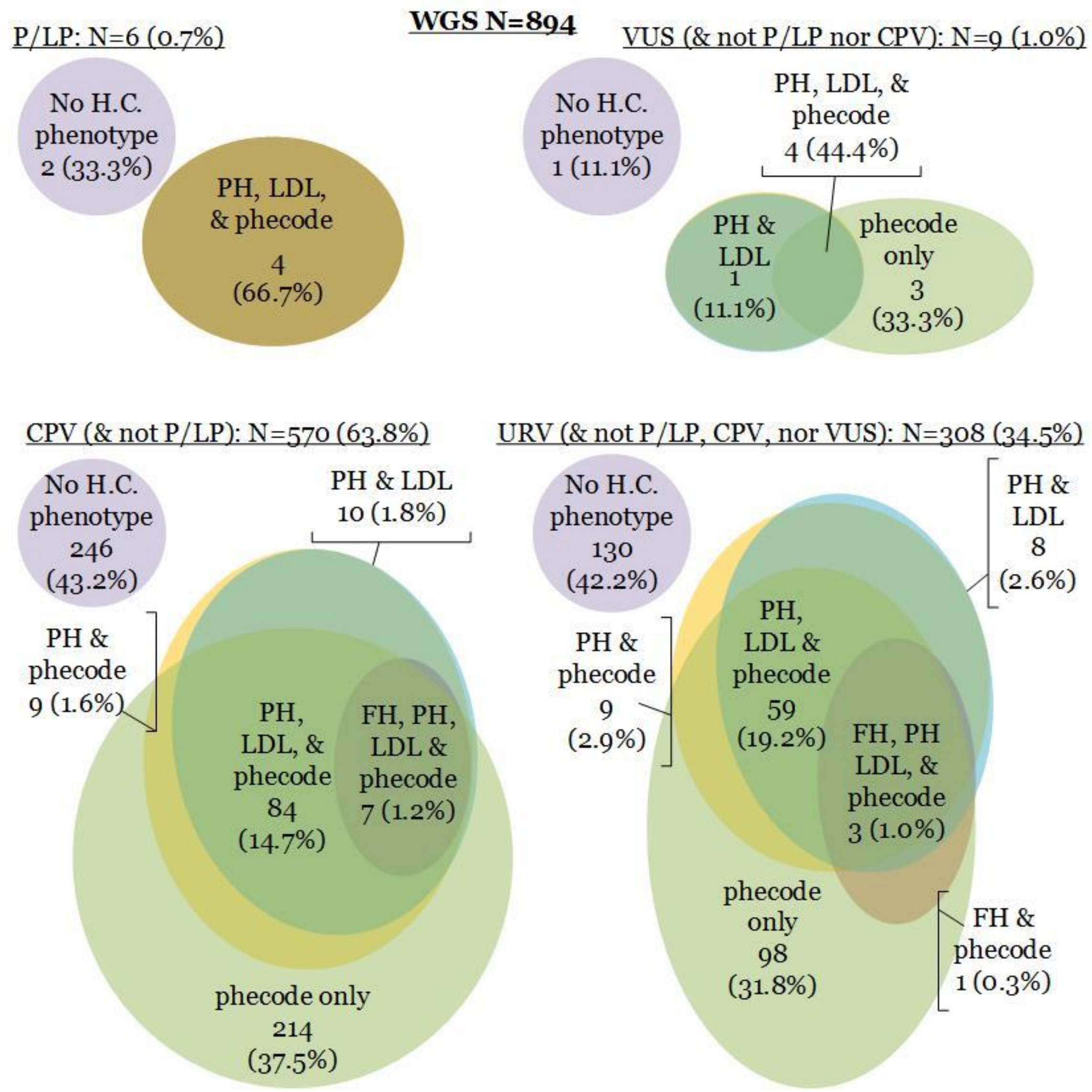

Figure 2a. Phenotype algorithm results by genetic variant, cohort a (WGS)

Counts and percentages of subjects with phenotypes by distinct genotypes in the WGS (a) cohort. P/LP = Pathogenic/Likely Pathogenic variant, CPV = Conflicting Pathogenic variant interpretations, VUS = Variant of Unknown Significance (but not any P/LP), URV = unreported variant (only). Phenotype algorithms: blue = LDL (high low-density lipoprotein without high triglycerides on >=2 days), yellow = PH (primary hypercholesterolemia (HC)), red = FH (familial

HC), green = phecode for HC or hyperlipidemia, purple = no H.C. (hypercholesterolemia) phenotype in EHR; note that overlapping areas are a mixture of the component colors.

Similarly, in Figure 2b, of the 18 eMERGEseq subjects with P/LP variants, all but 1 had evidence of HC, and although only 3 had FH, 7 the 13 with PH did have “possible” FH, for a total of 10 out of 18 with “possible,” “probable” or “definite” FH. Notably, phecodes were the only indication of HC for approximately one-quarter of the patients, i.e. none of the other 3 algorithms found 4, out of the 18 patients with P/LP variants in eMERGEseq, to have HC. For the 1 subject with no evidence of HC, there were no LDL lab results in the EHR. For the 165 eMERGEseq subjects with VUSs, approximately two-thirds had evidence of HC, and of the approximately one-third (N=52) who did not, only 33 had any LDL lab results in their EHR. Of those found to have HC, 4 had FH; 40 had PH, of which 12 had “possible” FH; and almost half (N=72 (43.6%)) only had evidence of HC found via phecodes.

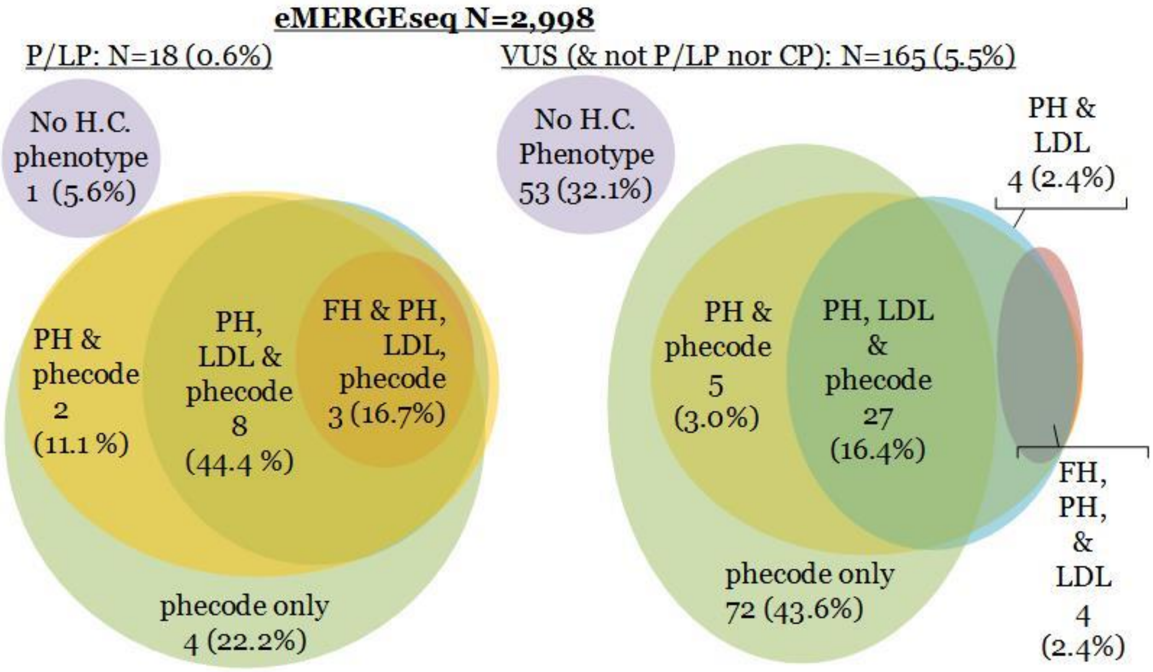

Figure 2b. Phenotype algorithm results by genetic variant, cohort b (eMERGEseq)

Counts and percentages of subjects with phenotypes by distinct genotypes in the eMERGEseq (b) cohort. Legend is the same as for Fig. 2a above. Note there were no CPVs nor URVs in the FH genes in the eMERGEseq cohort (b).

Next, to determine the number of subjects found by the phenotyping algorithms and how many of those had each variant type, we created similar figures by cohort. Figure 3a shows 11 WGS subjects with evidence of FH in their EHR as mostly having CPVs, with the remaining approximately one-third having URVs. Of those found to have PH without FH, all had a variant in the FH genes that was not known to be B/LB: more than half had CPVs, and of those, 44 had “possible” FH; 4 had P/LP variants of which 2 had “possible” FH, and of the remaining 188 with URVs, 33 had “possible” FH, for a total of 92 with “possible,” “probable,” or “definite” FH. Lastly of the 316 found to have HC via phecodes only, all but 1 had a possibly pathogenic variant: most had at least 1 URV, ~2/3 had at least 1 CPV, and only ~3% had a VUS.

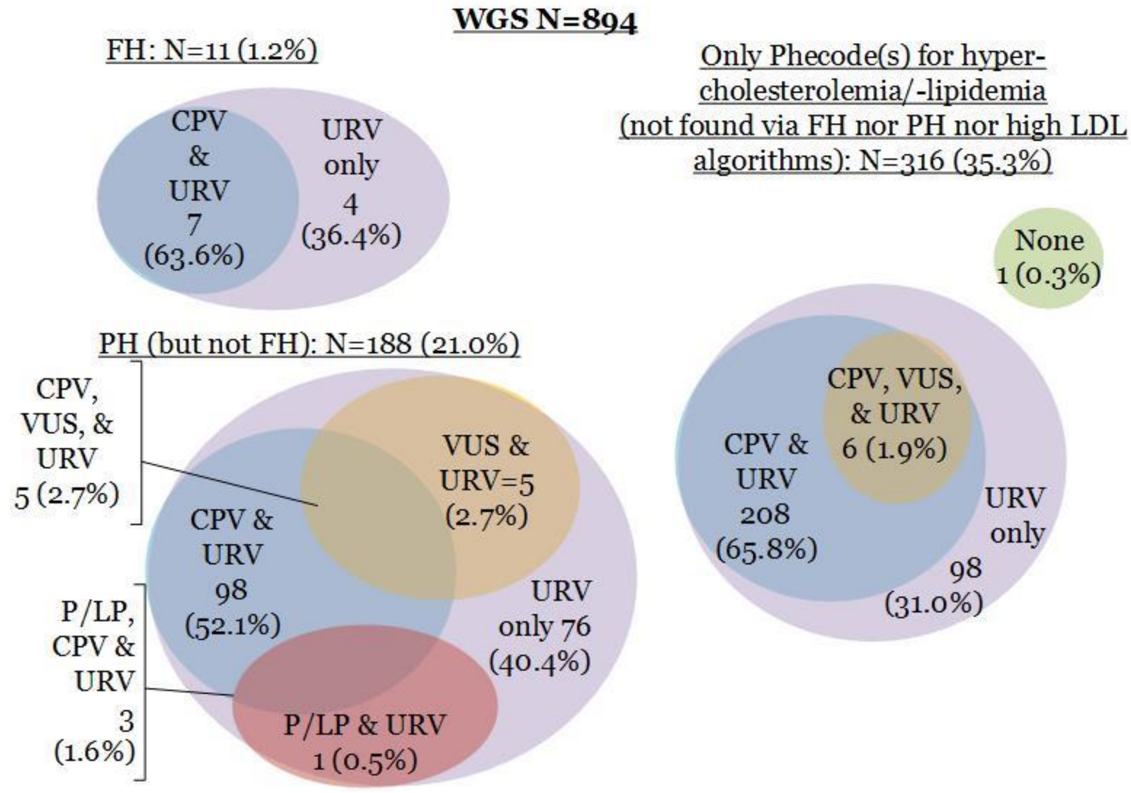

Figure 3a. Genetic variants by phenotype algorithm result, cohort a (WGS)

Counts and percentages of genotypes by distinct phenotypes in the WGS (a) cohort. FH = Familial Hypercholesterolemia (HC) algorithm, PH = Primary HC algorithm. Genetic variant types: red = P/LP (pathogenic/likely pathogenic), blue = CPV (conflicting pathogenic variant), yellow = VUS (variant of unknown significance), purple = URV (unreported variant), green = no variants; note that overlapping areas are a mixture of the component colors.

Finally, Figure 3b shows a majority of the eMERGEseq subjects found to have FH having no variants, with < 10 % each having P/LP variants or VUSs. Of those found to have PH, only a small percentage of those have P/LP variants, and of those, 7 (3.8%) had "possible" FH. Lastly, of the 1,238 subjects found to have HC only by using phecodes, only a small percentage had P/LP (0.3%) variants or VUSs (5.8%), and the remaining majority had no possibly pathogenic variants.

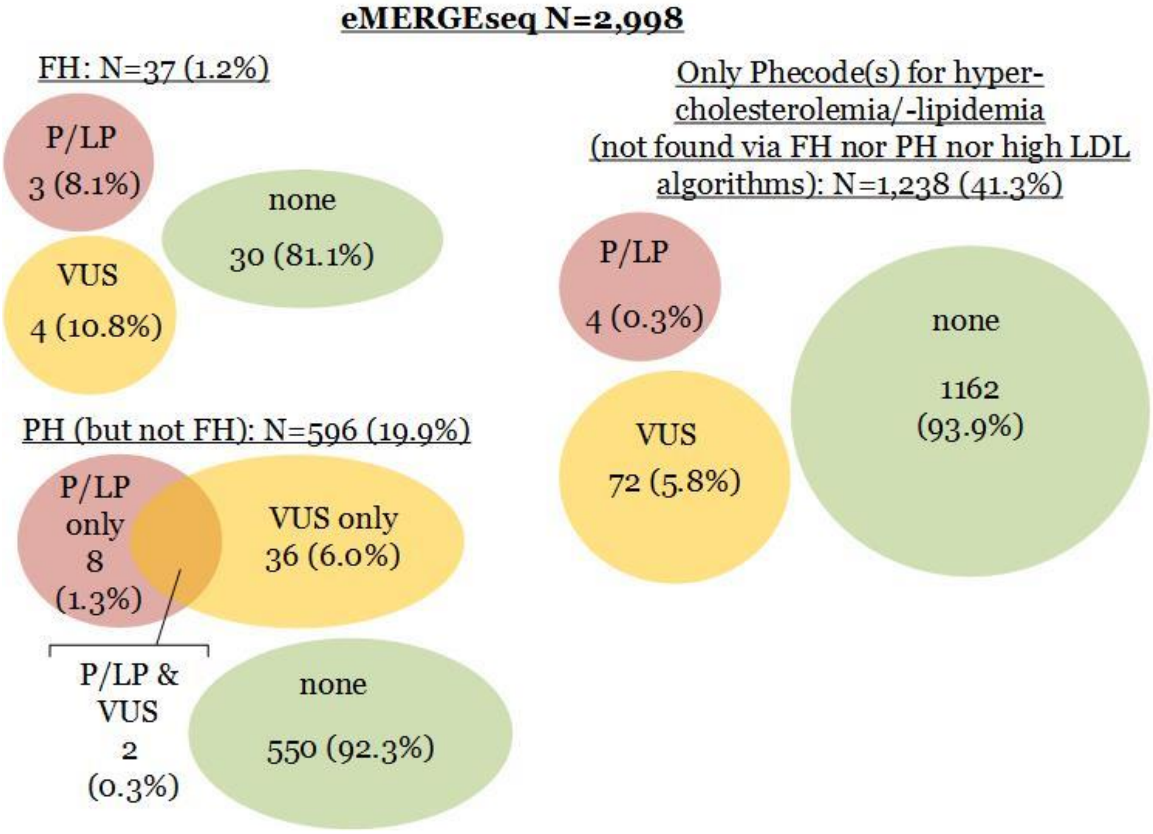

Figure 3b. Genetic variants by phenotype algorithm result, cohort b (eMERGEseq)

Counts and percentages of genotypes by distinct phenotypes in the eMERGEseq (b) cohort. Legend is the same as for Fig. 3a above.

There were no URVs in FH genes in the eMERGEseq cohort. Also, as stated in the methods, there were no CPVs reported in eMERGEseq. In particular, as mentioned previously, there was 1 subject who was in both cohorts, and from the WGS cohort, who had 7 CPVs in ClinVar; however, in the eMERGEseq study Baylor lab assigned 6 of those 7 variants as B/LB variants, and the remaining CPV as a VUS.

Next, to determine if some VUSs occurred in > 1 subject with evidence of HC, we examined the details of the VUSs, where possible. The results of comparing subjects with the exact same VUSs in FH genes is shown in Table 3, which list details for the 25 VUSs in FH genes in the WGS cohort that were seen in > 1 subject. Of those 25 VUSs, 21 (84.0%) had half or more (>=50%) of the subjects with that VUS exhibiting an HC phenotype. In other words, of 101 subjects with those 25 VUSs, 82 (81.1%) of those subjects had evidence of an HC phenotype in their EHR, for 21, collectively, of the VUSs. The VUSs in Table 3 are order in order from most to least likely to be pathogenic. In particular, the penetrance of FH from LDLR mutations is higher than those in APOB, which are higher than those in PCSK9, at least across the entire eMERGE network of ~25,000 subjects genotyped on the eMERGEseq platform, of which 128 had P/LP variants in FH genes and had outcomes forms filled out (Kullo). Furthermore, certain types of mutations, such as deletions and splices, are more severe than simple nonsynonymous mutations and thus also more likely to be pathogenic. Also VUSs that have more subjects with the VUS and a higher percentage of those subjects with HC, especially if FH or PH, might be more likely to be pathogenic. As noted in the legend for Table 3, evidence for each VUS to lean pathogenic vs. benign vs. still uncertain is highlighted in green, blue, and yellow, respectively. It is also highlighted if some of the subjects did not have LDL tested nor any other diagnoses in their EHR, as the lack of these types of data in the EHR make it difficult if not impossible to determine their HC status.

			Number of subjects in eMERGEseq cohort with:							
Gene	Mutation Type	VUS Genomic Info.	VUS	any HC (N (%))	definite or probable FH	PH & possible FH	PH (& not FH)	phecode(s) only	any LDL labs	no LDL labs nor any diagnosis in her
LDLR	Nonsynonymous	11240278 G>A	6	5 (83.3 %)	0	1	1	3	5	1
LDLR	Nonsynonymous	11231164 G>A	3	3 (100 %)	0	0	0	3	3	0
LDLR	Nonsynonymous	11224014 G>A	2	2 (100 %)	1	0	1	0	2	0
LDLR	Nonsynonymous	11200282 G>A	2	2 (100 %)	0	0	1	1	2	0
LDLR	Nonsynonymous	11233940 G>A	2	2 (100 %)	0	0	1	1	2	0
LDLR	Nonsynonymous	11224398 G>A	2	2 (100 %)	0	0	0	2	2	0
LDLR	Intronic	11216301 C>T	3	3 (100 %)	0	0	1	2	2	0
APOB	Deletion (nonframeshift)	21233099 _21233101 del	12	5 (41.7 %)	0	0	3	2	10	0
APOB	Splice region	21249840 A>T	6	6 (100 %)	0	0	0	6	6	0
APOB	Nonsynonymous	21238367 C>T	10	5 (50 %)	0	0	1	4	10	0
APOB	Nonsynonymous	21238323 G>A	9	7 (77.8 %)	0	0	3	4	9	0
APOB	Nonsynonymous	21225491 A>G	7	6 (85.7 %)	1	1	2	2	6	0
APOB	Nonsynonymous	21234674 C>T	7	4 (57.1 %)	0	1	2	1	7	0
APOB	Nonsynonymous	21232044 C>T	4	4 (100 %)	0	1	1	2	4	0
APOB	Nonsynonymous	21229032 G>A	4	4 (100 %)	0	0	1	3	4	0

			Number of subjects in eMERGEseq cohort with:							
Gene	Mutation Type	VUS Genomic Info.	VUS	any HC (N (%))	definite or probable FH	PH & possible FH	PH (& not FH)	phecode(s) only	any LDL labs	no LDL labs nor any diagnosis in her
APOB	Nonsynonymous	21229068 G>A	2	2 (100 %)	0	0	1	1	2	0
APOB	Nonsynonymous	21227979 C>T	2	2 (100 %)	0	0	0	2	1	0
APOB	Nonsynonymous	21260933 G>A	2	1 (50 %)	1	0	0	0	1	0
APOB	Nonsynonymous	21232455 A>T	2	1 (50 %)	0	0	0	1	2	0
APOB	Nonsynonymous	21231190 A>C	2	1 (50 %)	0	0	0	1	1	0
APOB	Nonsynonymous	21228437 A>G	2	0 (0 %)	0	0	0	0	1	0
PCSK9	Nonsynonymous	55505679 G>A	3	3 (100 %)	0	0	1	2	3	0
PCSK9	Nonsynonymous	55518374 C>T	3	1 (33.3 %)	0	0	0	1	1	0
PCSK9	Nonsynonymous	55523779 C>A	2	1 (50 %)	0	0	1	0	2	0
PCSK9	Nonsynonymous	55518422 C>T	2	0 (0 %)	0	0	0	0	1	0

Table 3. VUSs that occur in > 1 subject

Number of subjects in the WGS cohort for each VUS that occurs in >1 subject, with number & percentage of those who have HC by phenotype. Highlighted cells indicate evidence for each VUS to possibly be either pathogenic or benign as follows: green = favor pathogenic, yellow = maybe pathogenic, blue = favor benign.

Lastly, to determine if there were differences in the amount and type of VUSs between different races/ethnicities, as has been seen in previous cardiac disease studies in particular (Pottinger Tess D. et al.; Landry and Rehm), we examined the proportion of these types of variants for each race and ethnic group. Table 4 shows the number of subjects with VUSs, CPVs, and URVs across race and ethnicities, across both cohorts. Minorities, especially African-Americans, and Hispanics/Latinx, Native Americans/Alaska Natives, and Pacific Islanders/Hawaiians, have approximately twice as many VUSs, CPVs, or URVs than Caucasians, percentage-wise. Specifically: approximately one-third (34.3%) of Caucasians have VUSs, CPVs, or URVs; while approximately a half or more of African-Americans (58.6%), Native Americans (80.9%), Pacific Islanders (68.8%) and Hispanics (80.4%) have those types of variants. The only exception are Asians who have a similar percentage (32.6%) of subjects with VUSs, CPVs, or URVs compared to Caucasians. When breaking down by the individual classifications of these variants, VUSs (but no CPVs nor URVs), are lowest in Native Americans (3.4%) and Hispanics (4.8%), a little higher and about the same in Caucasians (6.5%) and Africans (7.1%), and highest in Asians (14%) and Pacific Islanders (12.4%). The difference is much higher when including CPVs with other VUSs: approximately one-quarter of Caucasians (21.5%) and Asians (24.4%) and 31.2% of Pacific Islanders have VUSs including CPVs variants; compared to, almost half (44.1%) of Hispanics, just over half (51.4%) of Africans, and approximately two-thirds of Native Americans (66.3%). Finally, the difference is the highest when comparing just URVs: Asians have the lowest at 20.9% and Caucasians are not much higher at 28.6%; yet, half of Africans (53.8%) and Pacific Islanders (56.2%) have URVs, and Hispanics (77.4%) and Native Americans (78.7%) have more than three-quarters with URVs.

		Number (%) of subjects with:			
Race/Ethnicity	Total across both cohorts	Total VUSs inc. CPVs, and/or URVs	VUSs inc. CPVs**	VUSs **	URVs**
Hispanic/Latinx*	438	352 (80.4%)	193 (44.1%)	21 (4.8%)	339 (77.4%)
Caucasian*	2135	732 (34.3%)	458 (21.5%)	139 (6.5%)	610 (28.6%)
African*	638	374 (58.6%)	328 (51.4%)	45 (7.1%)	343 (53.8%)
Asian*	86	28 (32.6%)	21 (24.4%)	12 (14%)	18 (20.9%)
Native American/Alaska Native*	89	72 (80.9%)	59 (66.3%)	3 (3.4%)	70 (78.7%)
Pacific Islander/Hawaiian*	16	11 (68.8%)	5 (31.2%)	2 (12.5%)	9 (56.2%)

Table 4. VUS proportions by race/ethnicity

Number & percentage of subjects across both cohorts, with the various types of variants of uncertain/unknown significance. VUS = variant of unknown significance, CPV = conflicting pathogenic variant, URV = unreported variant. Higher percentages of subjects in certain races/ethnicities, compared to other races/ethnicities, are highlighted. *Some patients reported more than 1 race/ethnicity; thus, the total for all races/ethnicities is >100%. **Similarly, some patients will have multiple types of uncertain variants, including VUSs, CPVs, and/or URVs; thus, the total across VUSs inc. CPV, VUSs, and URVs is >100%.

Chapter 4 DISCUSSION

It was expected that if there were significantly more patients found by 1 or more of the different phenotyping algorithms and the algorithms are found to be valid (most patients known to have disease found by the methods), then the algorithm(s) that identify more patients could be used to power future research, including genotype-phenotype association studies, for finding phenotypes associated with VUSs. Also, given the previously reported prevalence of up to 96% (Shah et al.; Kullo et al.; Kullo), and the older age of many of these patients (average age > 47), evidence in the EHR was expected of the manifestation of at least HC for the majority of those with P/LP variants and/or FH.

The validation results in other studies were sufficient to show that the FH algorithm performs well with >= 94% precision and 97% recall. In this study the validation results were similar for the PH and phecode algorithms, with precision at 100% for both, and recall at 82.4% and 100%, respectively. For this research, recall is more important, even at the sacrifice of lower precision, because it is more important not to miss any patients, so that more FH patients can be diagnosed and treated where possible. Also, the next step would be to follow-up with these patients to verify their phenotype, so lower precision is acceptable. Thus, the high LDL algorithm would not suffice even though the precision is 100%, as its recall is less than two-thirds. The only algorithm that did not perform well overall was the FH algorithm with less than one-quarter recall and only two-thirds precision. Inadequately documented family history in the EHR, either due to the history not being fully documented or being documented only in clinical narrative text, may be reason the FH algorithm did not perform as well (Safarova and Kullo; Mehrabi et al.). The results show that no single phenotyping algorithm found all relevant patients: even though the phecode algorithm did identify all patients with P/LP variants with HC, the phecode algorithm did not identify all patients with VUSs, CPVs, and/or URVs with evidence of HC. However, using both phecodes and the PH algorithm identified all (100%) of the patients VUSs, CPVs, and/or URVs with evidence of HC, as seen by the overlap of algorithms in the Euler diagrams in Figures 2a and 2b. These Euler diagrams show the high LDL and FH algorithms being almost completely covered by the PH algorithm, with a non-trivial percentage of patients being found by phecodes only; and is also illustrated by the higher number of patients with variants in the Euler diagrams in Figure 2 being found by the PH (but not FH) and phecode (but not FH, PH, or high LDL) algorithms. In hindsight, it seems almost obvious that PH would be the best algorithm compared to FH and high LDL algorithms, as the PH algorithm is a prerequisite for the FH algorithm and the high LDL algorithm is a subset of the PH algorithm. However, what was also found by this pilot study that might not be expected, is that there is a substantial proportion of the 643 patients across both cohorts with any type of variant, other than B/LB variants, in FH genes that were found to have HC via phecodes only (N=391 (60.1%)). In particular, 82% (N=2458) of those with an FH genotype and/or HC phenotype data (N=2795) had phecodes for HC.

In particular, summarizing across both cohorts, as seen in Figures 2 and 3, the phecode algorithm (the simplest algorithm), found the most patients with HC with P/LP variants, specifically 87.5% of those with P/LP variants (N=21), or VUSs including CPVs and URVs. Not all, but a significant proportion, from 22.2% with P/LP variants to 43.6% without P/LP variants but with VUSs, CPVs or URVs, were found only by using phecodes across both cohorts. Both the highest LDL and PH algorithms found a similar proportion of HC patients with P/LP variants or VUSs including CPVs and URVs, overall. Lastly, as expected, the FH algorithm (the strictest and most complex algorithm), found the least number of patients overall with HC with P/LP (N= 3) variants or VUSs including CPVs and URVs (N= 15), as illustrated in Figure 3. Note there were no subjects who only had high low-density lipoprotein labs >= 155 mg/dL (i.e., all of those with high LDLs had PH or FH).

Overall 99% of subjects did have recent ICD-10 diagnosis codes for any diagnosis, and over 80% had at least 1 LDL lab result; therefore, it is expected that the phenotyping algorithms would find evidence of HC if the subjects had HC. However, the eMERGEseq outcome forms indicated that not all of those with P/LP variants were diagnosed with FH or HC before genetic test results were given: only 77.7% (N=14) were, which indicates it might be hard for a phenotype algorithm to detect HC before genetic testing if almost a quarter of the patients weren't even diagnosed at that time. Thus, subjects that had P/LP variants but no evidence of HC were scrutinized further to ensure this was not due to lack of relevant EHR data. Of the 3 subjects with FH P/LP variants across both cohorts who did not have evidence of an HC phenotype in their EHR, all 3 never had a cholesterol or lipid panel lab test result in their NM EHR and 1 did not have any problem list diagnoses; thus, they may have FH but their EHR at NM appears to lack the necessary data to determine their FH status. Conversely, of the 24 subjects across both cohorts who had P/LP variants, half (N=12) had "definite", "probable," or "possible" FH, and 87.5% (N=21) had some evidence of HC, which parallels the previously observed penetrance of P/LP variants in FH genes (Kullo et al.; Kullo). Using phecodes was the only algorithm to find all 21 of the subjects across both cohorts with P/LP variants to have HC, while the PH algorithm also performed well, finding 70.8% (N=17) of those to have HC overall.

Most importantly, as illustrated by Figure 2, of those who had CPVs, VUSs, and/or URVs in both cohorts, and had evidence of HC, all were found via phecodes or the PH algorithm, yet neither of those individual algorithms found all of those with HC. Thus, using a combination of phecodes and the PH algorithm appears to be best way to identify patients with HC, especially as the PH algorithm includes "definite," "probable" and "possible" FH cases, and as seen in the results of this pilot study, some of the subjects with P/LP variants were classified as having "possible" FH indicating the need to not simply focus on "definite" and "probable" FH cases.

In addition, the fact that >95% of subjects enrolled from the lipid disorders clinic had evidence of HC, and all but 1 of those had phecodes for HC, confirms that the phecode algorithm works well to identify subjects with HC. The <5% of those who did not have HC could have been subjects who were not blood relatives of the lipid clinic patients as spouses and other relatives of patients were also enrolled in the studies. Lastly, the small proportion (20.8%, N=5) of subjects with P/LP variants who had the ICD-10 code for FH illustrates the need for using more than just ICD diagnosis codes to find patients with FH in particular.

Furthermore, overall, as seen in Figure 3, those who met criteria for "definite" or "probable" FH, most did not have P/LP but instead had VUSs, CPVs, or URVs, or no potentially pathogenic mutations in FH genes at all. Of those that did have variants, more had an URV, somewhat less had a CPV, and the least had VUSs, and this is also seen for the other phenotype algorithms; however, this could be a result of the eMERGEseq cohort receiving expert curation beyond ClinVar and thus not having any CPVs nor URVs, or, this could be due to the demographic differences between the cohorts. This illustrates the need for follow up on the individual patients for further investigation by experts.

For those subjects who met the phenotype criteria for HC, especially "definite," "probable," or "possible" FH, or PH, if they only had VUSs in the FH genes, those are patients whose EHR should be reviewed in more detail, and if warranted by the chart review, should be contacted to discuss and possibly conduct further confirmatory testing (such as further genetic testing and further lipid, esp. LDL, lab tests). In particular, for the 25 VUSs the eMERGEseq cohort seen in > 1 subject, over half (64%) had evidence that the variants were actually P/LP, 2 had evidence that the variants were actually B/LB, and another 28% had mixed evidence, which indicates the need for further investigation. Combining the phenotype algorithm results with genetic results can provide sufficient evidence for prioritizing subjects and variants for follow-up. Furthermore, at least 1 subject with P/LP variants was only put on an LLT after genetic testing, which illustrates the need to find undiagnosed patients who are not receiving treatment, as other studies have shown (Banda et al.; deGoma et al.). If the further investigation and/or follow-up reveals that the variant is likely pathogenic or likely benign, then this should be reported to ClinVar as evidence for assisting in the eventual proper categorization of that variant.

Lastly, the importance of finding patients with potential HC or FH for follow-up and doing further research to categorize genetic variants, especially in minorities, is highlighted by the results shown in Table 4. Even Asians, who have mostly the same proportion having VUSs, CPVs, or URVs, still have more than double the percentage of VUSs than Caucasians. Hispanics and Native Americans are affected the most, while African-Americans and Pacific Islanders also have a higher proportion of these variants which need clearer classification. The only categories in which Caucasians do not have the least, or close to the least, proportion is of VUSs only at 6.5%, where instead Hispanics (4.8%) and Native Americans (3.4%) have the least; and URVs with 28.6% which is slightly higher than Asians with the least proportion (20.9%) of URVs. These a relatively small differences compared to the much greater difference between Caucasians and minorities overall, which aligns with previous studies of VUSs in cardiac disease genes (Pottinger Tess D. et al.; Landry and Rehm). As previous studies have shown that the contribution of genotypes vs. the environment to the manifestation of FH varies sometimes significantly between races and ethnicities (Wright et al.), this should also be taken into account.

Caveats and Limitations

1. Given that genetic variants can be rare and genetic variation between humans is less than 1 percent among millions of nucleotide base pairs in our DNA, we might not have enough patients in our study to have enough power to find disease associations with any, or at least some of the rarer, genetic variations. However, in this study we are not performing association studies between the genetic variants found and the extracted phenotypic data, so we are not implying that we have found associations. Instead, we are simply filtering out patients with genetic variants, for further investigation, who also have some phenotypic evidence in their EHR of the disease caused by variants in the same genes which have known P/LP variants to cause the given disease.

2. The EHR usually does not contain a given patient's entire health history; in particular, data such as family history and environmental exposures are not well captured. Thus, the EHR may not have all the data that might be necessary to accurately determine all relevant diagnoses for the patients in our study.

3. No association studies are being performed, and correlation does not imply causation. This is a descriptive study and any validation being done will need to be replicated on a larger scale, and likely via review of patients' medical charts, to determine true accuracy.

Future work

Further investigation can be done in the WGS cohort in particular, specifically the details of the VUSs need to be extracted to determine which VUSs appear in >1 subject for investigating those VUSs further as was done in the eMERGEseq cohort. Furthermore, it would likely be worth further sub-dividing the CPVs into the 3 sub-types of conflicting pathogenic, in order from highest to lowest probability to become pathogenic/likely pathogenic, as follows:

1. P/LP & VUS
2. P/LP & B/LB
3. B/LB & VUS

For both cohorts, further investigation into the VUSs identified by this pilot study as likely pathogenic or likely benign is needed by genetic experts to determine if these variants are B/LB or P/LP. Testing should also be done on other genes that have genotypes associated with other disease phenotypes.

Further EHR phenotyping could also be conducted, first by using Observational Medical Outcomes Partnership (OMOP) common data model (*OMOP Common Data Model – OHDSI*) or other common data models (CDMs) to make the phenotype algorithms more portable to other sites for study, at least for sites that have these CDMs (Safarova and Kullo). Additional de-identified data elements, such as other common medications and/or common labs not already collected, could be more easily obtained via an OMOP or other CDM query, such that more phenotypes could be studied. Improving the documentation of family history in the EHR with integrated tools such as MeTree (R. R. Wu et al.), would enable these algorithms to more easily access this important piece of data for many phenotypes. Any clinical criteria used would need to be determined from existing algorithms or machine learning, and subsequently by consultation with clinicians who diagnosis the selected disease phenotypes to confirm feature selection.

In addition, more complex yet more efficient algorithms such as machine learning (ML), either classification or clustering or other appropriate technique mined from the literature (Liao et al.; Yu et al.; Myers et al.; Peissig et al.; Banda et al.), could also be performed, where the features used were the extracted commonly available discrete EHR data used by the algorithms described above. In particular, clustering of patients using extracted EHR data and P/LP variants, and VUSs, in selected genes; or association analysis using various types of logistic regression, might yield interesting results if enough patients had genetic testing for comparison and possible association. In a recent study whose goal was to find patients with undiagnosed FH in the EHR, machine learning successfully identified multiple cases of FH (Banda et al.); thus, it is likely possible to use a similar algorithm to identify case of FH in patients with VUSs in FH genes for follow-up.

Lastly, these phenotype algorithms could also potentially be used to identify patients without genetic testing who might have FH, to detect more undiagnosed cases of FH. As mentioned earlier, a machine learning algorithm was developed to do this, at Stanford, and results were replicated at Geisinger, with both institutions achieving a precision of 84% or higher at identifying FH cases (Banda et al.).

Chapter 5 Summary and Conclusions

Although further assessment is needed, these initial results demonstrate that EHR phenotyping can be used to identify phenotypes in patients with VUSs, and with other data, could be used to categorize VUSs as either P/LP or B/LB. The results also demonstrate that no single phenotyping algorithm, even those validated in previous studies, can necessarily extract the majority of patients with a given phenotype from EHR data. Finally, the phenotypic information is most useful when compared with VUSs that occur in more than 1 subject and can be used to prioritize patients, and VUSs, for further investigation.

Chapter 6 References

Abul-Husn, Noura S., et al. “Genetic Identification of Familial Hypercholesterolemia within a Single U.S. Health Care System.” Science, vol. 354, no. 6319, American Association for the Advancement of Science, Dec. 2016. science.sciencemag.org, doi:10.1126/science.aaf7000.

Ackerman, Michael J. “Genetic Purgatory and the Cardiac Channelopathies: Exposing the Variants of Uncertain/Unknown Significance Issue.” Heart Rhythm, vol. 12, no. 11, Nov. 2015, pp. 2325–31. ScienceDirect, doi:10.1016/j.hrthm.2015.07.002.

Akioyamen, Leo E., et al. “Estimating the Prevalence of Heterozygous Familial Hypercholesterolaemia: A Systematic Review and Meta-Analysis.” BMJ Open, vol. 7, no. 9, British Medical Journal Publishing Group, Sept. 2017, p. e016461. bmjopen.bmj.com, doi:10.1136/bmjopen-2017-016461.

Ambry Genetics. Understanding Your VUS Familial Hypercholesterolemia (FH) Genetic Test Result: Information for Patients with a Variant or Variants of Unknown Significance. https://www.ambrygen.com/file/material/view/1085/UYRVUS_FH_1018_Final.pdf . Accessed 17 Nov. 2019.

Banda, Juan M., et al. “Finding Missed Cases of Familial Hypercholesterolemia in Health Systems Using Machine Learning.” Npj Digital Medicine, vol. 2, no. 1, Apr. 2019, pp. 1–8. www.nature.com, doi:10.1038/s41746-019-0101-5.

Bastarache, Lisa, et al. “Phenotype Risk Scores Identify Patients with Unrecognized Mendelian Disease Patterns.” Science, vol. 359, no. 6381, American Association for the Advancement of Science, Mar. 2018, pp. 1233–39. science.sciencemag.org, doi:10.1126/science.aal4043.

Beaulieu-Jones, Brett K., and Casey S. Greene. “Semi-Supervised Learning of the Electronic Health Record for Phenotype Stratification.” Journal of Biomedical Informatics, vol. 64, Dec. 2016, pp. 168–78. ScienceDirect, doi:10.1016/j.jbi.2016.10.007.

Bert, Jm, et al. “Getting Ready for ICD-10 and Meaningful Use Stage 2.” Instructional Course Lectures, vol. 65, Jan. 2016, pp. 609–22.

Calandra, Sebastiano, et al. “Impact of Rare Variants in Autosomal Dominant Hypercholesterolemia Causing Genes.” Current Opinion in Lipidology, vol. 28, no. 3, June 2017, pp. 267–272. journals.lww.com, doi:10.1097/MOL.0000000000000414.

Caswell-Jin, Jennifer L., et al. “Racial/Ethnic Differences in Multiple-Gene Sequencing Results for Hereditary Cancer Risk.” Genetics in Medicine, vol. 20, no. 2, Feb. 2018, pp. 234–39. www.nature.com, doi:10.1038/gim.2017.96.

Chora, Joana Rita, et al. “Analysis of Publicly Available LDLR , APOB , and PCSK9 Variants Associated with Familial Hypercholesterolemia: Application of ACMG Guidelines and Implications for Familial Hypercholesterolemia Diagnosis.” Genetics in Medicine, vol. 20, no. 6, June 2018, pp. 591–98. www.nature.com, doi:10.1038/gim.2017.151.

deGoma, Emil M., et al. “Treatment Gaps in Adults With Heterozygous Familial Hypercholesterolemia in the United States: Data From the CASCADE-FH Registry.” Circulation. Cardiovascular Genetics, vol. 9, no. 3, June 2016, pp. 240–49. PubMed, doi:10.1161/CIRCGENETICS.116.001381.

Denny, Joshua C., Marylyn D. Ritchie, et al. “PheWAS: Demonstrating the Feasibility of a Phenome-Wide Scan to Discover Gene–Disease Associations.” Bioinformatics, vol. 26, no. 9, Oxford Academic, May 2010, pp. 1205–10. academic.oup.com, doi:10.1093/bioinformatics/btq126.

Denny, Joshua C., Lisa Bastarache, et al. “Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data.” Nature Biotechnology, vol. 31, no. 12, 12, Nature Publishing Group, Dec. 2013, pp. 1102–11. www.nature.com, doi:10.1038/nbt.2749.

Electronic Health Record-Based Phenotyping Algorithm for Familial Hypercholesterolemia | PheKB. https://phekb.org/phenotype/electronic-health-record-based-phenotyping-algorithm-familial-hypercholesterolemia. Accessed 7 Mar. 2020.

Geisinger. Implementation For Phase 3 EMERGE | PheKB. https://phekb.org/implementation/implementation-phase-3-emerge. Accessed 7 Mar. 2020.

Gottesman, Omri, et al. “The Electronic Medical Records and Genomics (EMERGE) Network: Past, Present, and Future.” Genetics in Medicine, vol. 15, no. 10, 10, Nature Publishing Group, Oct. 2013, pp. 761–71. www.nature.com, doi:10.1038/gim.2013.72.

Hooper, Amanda J., et al. “The Present and the Future of Genetic Testing in Familial Hypercholesterolemia: Opportunities and Caveats.” Current Atherosclerosis Reports, vol. 20, no. 6, May 2018, p. 31. Springer Link, doi:10.1007/s11883-018-0731-0.

Hripcsak, George, and David J. Albers. “Next-Generation Phenotyping of Electronic Health Records.” Journal of the American Medical Informatics Association, vol. 20, no. 1, Oxford Academic, Jan. 2013, pp. 117–21. academic.oup.com, doi:10.1136/amiajnl-2012-001145.

Hsu, Joy, et al. “Accuracy of Phenotyping Chronic Rhinosinusitis in the Electronic Health Record.” American Journal of Rhinology & Allergy, vol. 28, no. 2, SAGE

Publications Inc, Mar. 2014, pp. 140–44. SAGE Journals, doi:10.2500/ajra.2014.28.4012.

Humphries, S. E., et al. “Coronary Heart Disease Mortality in Treated Familial Hypercholesterolaemia: Update of the UK Simon Broome FH Register.” Atherosclerosis, vol. 274, July 2018, pp. 41–46. ScienceDirect, doi:10.1016/j.atherosclerosis.2018.04.040.

Iacocca, Michael A., et al. “ClinVar Database of Global Familial Hypercholesterolemia-Associated DNA Variants.” Human Mutation, vol. 39, no. 11, 2018, pp. 1631–40. Wiley Online Library, doi:10.1002/humu.23634.

Jeff, Janina M., et al. “Generalization of Variants Identified by Genome-Wide Association Studies for Electrocardiographic Traits in African Americans.” Annals of Human Genetics, vol. 77, no. 4, July 2013, pp. 321–32. PubMed, doi:10.1111/ahg.12023.

Kalia, Sarah S., et al. “Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing, 2016 Update (ACMG SF v2.0): A Policy Statement of the American College of Medical Genetics and Genomics.” Genetics in Medicine, vol. 19, no. 2, 2, Nature Publishing Group, Feb. 2017, pp. 249–55. www.nature.com, doi:10.1038/gim.2016.190.

Kast, Karin, et al. “Changes in Classification of Genetic Variants in BRCA1 and BRCA2.” Archives of Gynecology and Obstetrics, vol. 297, no. 2, Feb. 2018, pp. 279–80. Springer Link, doi:10.1007/s00404-017-4631-2.

Kho, Abel N., Jennifer A. Pacheco, et al. “Electronic Medical Records for Genetic Research: Results of the EMERGE Consortium.” Science Translational Medicine, vol. 3, no. 79, Apr. 2011, p. 79re1. PubMed, doi:10.1126/scitranslmed.3001807.

Kho, Abel N., M. Geoffrey Hayes, et al. “Use of Diverse Electronic Medical Record Systems to Identify Genetic Risk for Type 2 Diabetes within a Genome-Wide Association Study.” Journal of the American Medical Informatics Association:

JAMIA, vol. 19, no. 2, Apr. 2012, pp. 212–18. PubMed, doi:10.1136/amiajnl-2011-000439.

Klarin, Derek, et al. “Genetics of Blood Lipids among ~300,000 Multi-Ethnic Participants of the Million Veteran Program.” Nature Genetics, vol. 50, no. 11, 2018, pp. 1514–23. PubMed, doi:10.1038/s41588-018-0222-9.

Kramer, Adam I., et al. “Estimating the Prevalence of Familial Hypercholesterolemia in Acute Coronary Syndrome: A Systematic Review and Meta-Analysis.” Canadian Journal of Cardiology, vol. 35, no. 10, Oct. 2019, pp. 1322–31. ScienceDirect, doi:10.1016/j.cjca.2019.06.017.

Kullo, Iftikhar J. 6 Month Outcomes after ROR of FH Variants in 128 Adults: Results from Data Freeze 2. eMERGE Network Steering Committee Meeting, National Human Genome Research Institute, 20 Feb. 2020, Hyatt Regency Bethesda, MD. Presentation.

---. “The Return of Actionable Variants Empirical (RAVE) Study, a Mayo Clinic Genomic Medicine Implementation Study: Design and Initial Results.” Mayo Clinic Proceedings, vol. 93, no. 11, Nov. 2018, pp. 1600–10. ScienceDirect, doi:10.1016/j.mayocp.2018.06.026.

Lan, Nick S. R., et al. “Improving the Detection of Familial Hypercholesterolaemia.” Pathology, vol. 51, no. 2, Feb. 2019, pp. 213–21. ScienceDirect, doi:10.1016/j.pathol.2018.10.015.

Landry, Latrice G., and Heidi L. Rehm. “Association of Racial/Ethnic Categories With the Ability of Genetic Tests to Detect a Cause of Cardiomyopathy.” JAMA Cardiology, vol. 3, no. 4, 01 2018, pp. 341–45. PubMed, doi:10.1001/jamacardio.2017.5333.

Liao, Katherine P., et al. “Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic

Disease Cohorts." PLoS ONE, vol. 10, no. 8, Aug. 2015. PubMed Central, doi:10.1371/journal.pone.0136651.

Mayo Clinic. Mayo FH Implementation | PheKB. https://phekb.org/implementation/mayo-fh-implementation. Accessed 7 Mar. 2020.

Mehrabi, Saeed, et al. "Exploring Gaps of Family History Documentation in EHR for Precision Medicine -A Case Study of Familial Hypercholesterolemia Ascertainment." AMIA Summits on Translational Science Proceedings, vol. 2016, July 2016, pp. 160–66.

Myers, Kelly D., et al. "Precision Screening for Familial Hypercholesterolaemia: A Machine Learning Study Applied to Electronic Health Encounter Data." The Lancet Digital Health, vol. 1, no. 8, Elsevier, Dec. 2019, pp. e393–402. www.thelancet.com, doi:10.1016/S2589-7500(19)30150-5.

National Center for Biotechnology Information, U.S., National Library of Medicine. ClinVar. https://www.ncbi.nlm.nih.gov/clinvar/. Accessed 7 Mar. 2020.

---. Representation of Clinical Significance in ClinVar and Other Variation Resources at NCBI. https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/. Accessed 7 Mar. 2020.

OMOP Common Data Model – OHDSI. https://www.ohdsi.org/data-standardization/the-common-data-model/. Accessed 29 Feb. 2020.

Pacheco, Jennifer A., et al. "A Highly Specific Algorithm for Identifying Asthma Cases and Controls for Genome-Wide Association Studies." AMIA Annual Symposium Proceedings, vol. 2009, 2009, pp. 497–501.

Pathak, Jyotishman, et al. "Electronic Health Records-Driven Phenotyping: Challenges, Recent Advances, and Perspectives." Journal of the American Medical Informatics Association, vol. 20, no. e2, Oxford Academic, Dec. 2013, pp. e206–11. academic.oup.com, doi:10.1136/amiajnl-2013-002428.

Peissig, Peggy L., et al. “Relational Machine Learning for Electronic Health Record-Driven Phenotyping.” Journal of Biomedical Informatics, vol. 52, Dec. 2014, pp. 260–70. ScienceDirect, doi:10.1016/j.jbi.2014.07.007.

Pottinger Tess D., et al. “Pathogenic and Uncertain Genetic Variants Have Clinical Cardiac Correlates in Diverse Biobank Participants.” Journal of the American Heart Association, vol. 9, no. 3, American Heart Association, Feb. 2020, p. e013808. ahajournals.org (Atypon), doi:10.1161/JAHA.119.013808.

Rasmussen-Torvik, Laura J., et al. “High Density GWAS for LDL Cholesterol in African Americans Using Electronic Medical Records Reveals a Strong Protective Variant in APOE.” Clinical and Translational Science, vol. 5, no. 5, Oct. 2012, pp. 394–99. PubMed, doi:10.1111/j.1752-8062.2012.00446.x.

Safarova, Maya S., Benjamin A. Satterfield, et al. “A Phenome-Wide Association Study to Discover Pleiotropic Effects of PCSK9 , APOB , and LDLR.” Npj Genomic Medicine, vol. 4, no. 1, 1, Nature Publishing Group, Feb. 2019, pp. 1–9. www.nature.com, doi:10.1038/s41525-019-0078-7.

Safarova, Maya S., Hongfang Liu, et al. “Rapid Identification of Familial Hypercholesterolemia from Electronic Health Records: The SEARCH Study.” Journal of Clinical Lipidology, vol. 10, no. 5, Sept. 2016, pp. 1230–39. ScienceDirect, doi:10.1016/j.jacl.2016.08.001.

Safarova, Maya S., Eric W. Klee, et al. “Variability in Assigning Pathogenicity to Incidental Findings: Insights from LDLR Sequence Linked to the Electronic Health Record in 1013 Individuals.” European Journal of Human Genetics, vol. 25, no. 4, 4, Nature Publishing Group, Apr. 2017, pp. 410–15. www.nature.com, doi:10.1038/ejhg.2016.193.

Safarova, Maya S., and Iftikhar J. Kullo. “Using the Electronic Health Record for Genomics Research.” Current Opinion in Lipidology, vol. 31, no. 2, Apr. 2020, pp. 85–93. journals.lww.com, doi:10.1097/MOL.0000000000000662.

Séguro, Florent, et al. “Dutch Lipid Clinic Network Low-Density Lipoprotein Cholesterol Criteria Are Associated with Long-Term Mortality in the General Population.” Archives of Cardiovascular Diseases, vol. 108, no. 10, Oct. 2015, pp. 511–18. ScienceDirect, doi:10.1016/j.acvd.2015.04.003.

Shah, Naisha, et al. “Identification of Misclassified ClinVar Variants via Disease Population Prevalence.” The American Journal of Human Genetics, vol. 102, no. 4, Apr. 2018, pp. 609–19. ScienceDirect, doi:10.1016/j.ajhg.2018.02.019.

Sharifi, Mahtab, et al. “Genetic Architecture of Familial Hypercholesterolaemia.” Current Cardiology Reports, vol. 19, no. 5, Apr. 2017, p. 44. Springer Link, doi:10.1007/s11886-017-0848-8.

Slavin, Thomas P., et al. “Prospective Study of Cancer Genetic Variants: Variation in Rate of Reclassification by Ancestry.” Journal of the National Cancer Institute, vol. 110, no. 10, 01 2018, pp. 1059–66. PubMed, doi:10.1093/jnci/djy027.

Stein, Ricardo, et al. “Genetics, Dyslipidemia, and Cardiovascular Disease: New Insights.” Current Cardiology Reports, vol. 21, no. 8, June 2019, p. 68. Springer Link, doi:10.1007/s11886-019-1161-5.

Sturm, Amy C., et al. “Clinical Genetic Testing for Familial Hypercholesterolemia: JACC Scientific Expert Panel.” Journal of the American College of Cardiology, vol. 72, no. 6, Aug. 2018, pp. 662–80. ScienceDirect, doi:10.1016/j.jacc.2018.05.044.

Wei, Wei-Qi, et al. “Evaluating Phecodes, Clinical Classification Software, and ICD-9-CM Codes for Phenome-Wide Association Studies in the Electronic Health Record.” PloS One, vol. 12, no. 7, 2017, p. e0175508. PubMed, doi:10.1371/journal.pone.0175508.

Wierzbicki, Anthony S., et al. “Familial Hypercholesterolaemia: Summary of NICE Guidance.” BMJ, vol. 337, British Medical Journal Publishing Group, Aug. 2008. www.bmj.com, doi:10.1136/bmj.a1095.

Wright, Michelle L., et al. “A Perspective for Sequencing Familial Hypercholesterolaemia in African Americans.” Npj Genomic Medicine, vol. 1, no. 1, 1, Nature Publishing Group, May 2016, pp. 1–4. www.nature.com, doi:10.1038/npjgenmed.2016.12.

Wu, Patrick, et al. “Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation.” JMIR Medical Informatics, vol. 7, no. 4, Nov. 2019, p. e14325. PubMed, doi:10.2196/14325.

Wu, R. Ryanne, et al. “Implementation, Adoption, and Utility of Family Health History Risk Assessment in Diverse Care Settings: Evaluating Implementation Processes and Impact with an Implementation Framework.” Genetics in Medicine: Official Journal of the American College of Medical Genetics, vol. 21, no. 2, 2019, pp. 331–38. PubMed, doi:10.1038/s41436-018-0049-x.

Yu, Sheng, et al. “Toward High-Throughput Phenotyping: Unbiased Automated Feature Extraction and Selection from Knowledge Sources.” Journal of the American Medical Informatics Association, vol. 22, no. 5, Oxford Academic, Sept. 2015, pp. 993–1000. academic.oup.com, doi:10.1093/jamia/ocv034.

Zouk, Hana, et al. “Harmonizing Clinical Sequencing and Interpretation for the EMERGE III Network.” The American Journal of Human Genetics, vol. 105, no. 3, Sept. 2019, pp. 588–605. ScienceDirect, doi:10.1016/j.ajhg.2019.07.018.

Chapter 7 APPENDIX A: Supplemental Tables

Supplemental Table 1. Gene mutation types by variant type by cohort

Gene mutation types found in FH genes in the 2 patient cohorts and how used to classify variants as VUS and URVs

	eMERGEseq		WGS	
Mutation Type	**VUS**	**URV**	**VUS**	**URV**
Nonsynonymous	✓	✓	✓	✓
Synonymous	✓		✓	
Intronic	✓	✓	✓	
Cryptic Splice (Acceptor)	✓	✓	✓	
Cryptic Splice (Donor)	✓	✓	✓	
Splicing	✓	✓	✓	
Splice region	✓	✓	✓	
Insertion (nonframeshift)	✓		✓	
Deletion (nonframeshift)	✓	✓	✓	
Frameshift	✓	✓	✓	
Startloss	✓		✓	
Stoploss	✓		✓	
Stopgain	✓	✓	✓	
UTR3	✓		✓	
UTR5	✓		✓	
Upstream	✓		✓	

Supplemental Tables 2a-d. Accuracy of algorithms with confusion matrices

Confusion matrices with accuracy statistics comparing the chart reviewed results from the eMERGEseq outcome forms to the FH, PH, high LDL, and phecode algorithms. In yellow are lower precision and recall values that indicate the algorithm performance was not sufficient, and in green are higher precision and recall values that are more ideal. PPV = Positive Predictive Value (precision), NPV = Negative Predictive Value, Sens. = Sensitivity (recall), Spec. = Specificity.

		FH		**55.6%**	**accuracy**
		FH	NO FH		
FH algorithm	FH	2	1	**66.7%**	**precision/PPV**
	NO FH	7	8	53.3%	NPV
		22.2%	88.9%		
		recall/ sensitivity	specificity		
Supplementary Table 2a					

2a compares FH found via chart review with the FH algorithm

		PH		**83.3%**	**accuracy**
		PH	NO PH		
PH algorithm	PH	14	0	**100.0%**	**precision/PPV**
	NO PH	3	1	25.0%	NPV
		82.4%	100.0%		
		recall/ sensitivity	specificity		
Supplementary Table 2b					

2b compares PH (and FH, as patients with FH by definition have PH) found via chart review with the PH algorithm

		PH		**66.7%**	**accuracy**
		PH	NO PH		
high LDL algorithm	PH	11	0	**100.0%**	**precision/PPV**
	NO PH	6	1	14.3%	NPV
		64.7%	100.0%		
		recall/ sensitivity	specificity		

Supplementary Table 2c

2c compares PH found via chart review as for 2b,
but compares to the high LDL algorithm

		PH		**100.0%**	**accuracy**
		PH	NO PH		
phe-codes	PH	17	0	**100.0%**	**precision/PPV**
	NO PH	0	1	100.0%	NPV
		100%	100.0%		
		recall/ sensitivity	specificity		

Supplementary Table 2d

2d compares PH found via chart review as for 2b,
but compares to the phecode algorithm

Chapter 8 APPENDIX B: Definition of major terms

1. Phenotypes are diseases, disorders, or traits caused potentially, in part at least, by genetic variation.

2. Single Nucleotide Polymorphisms (SNPs) are single nucleotides, usually within a gene, that have multiple possible nucleotides (A, T, C, G) among the population, which can cause different phenotypes to be expressed.

3. Variant Classifications:
    a. Known Pathogenic (KP or just P) & Likely Pathogenic (LP) variants are genetic variants that are known to cause certain disease(s), or are likely to cause them, respectively.
    b. Variants of Unknown Significance (VUSs) are genetic variants that cause a change in the protein that the variant is coded to create, but it is not known if the variation or change causes disease or is benign (does not cause disease, is relatively harmless).
    c. Benign & Likely Benign (B/LB) variants are genetic variants that are known to NOT cause any known disease(s), i.e., are essentially harmless
    d. Variants with conflicting interpretations of pathogenicity (CPVs) are a ClinVar designation for VUSs that have multiple conflicting interpretations reported to ClinVar which include reports of pathogenicity. Thus, although they are technically VUSs, they are also a separate category from VUSs (National Center for Biotechnology Information, U.S., *Representation of Clinical Significance in ClinVar and Other Variation Resources at NCBI*).
    e. Unreported variants (URVs) are variants that were found during genotyping that are not reported to ClinVar, which are nonsynonymous substitutions (variants that cause a change in an amino acid which thus results in changes to the protein)

4. Hypercholesterolemia phenotypes and the algorithms:
    a. Hypercholesterolemia (HC) is high or elevated cholesterol levels in the blood which are a risk factor for heart disease.
    b. High low-density lipoprotein (LDL) lab results are an indication of HC. Thus the high LDL algorithm attempts to simply find these lab results, where the result of another lab test usually performed together with LDL as part of what's typically called a lipid panel, triglycerides, are not elevated more than once, as elevated triglycerides are an indication of a secondary cause of HC.
    c. Primary HC (PH) is HC not caused by secondary causes such as pregnancy. The PH algorithm thus looks for high LDL, where there are no

secondary causes, and also adjusts the LDL level when patients are also on lipid-lowering treatment at the time of the LDL measurement to estimate the LDL without treatment (Safarova, Liu, et al.).

d. Familial HC (FH) is PH inherited from family members and puts patients at a higher risk for heart disease (Akioyamen et al.; Lan et al.; Kramer et al.). The FH algorithm takes patients who meet the criteria for PH from the PH algorithm, and then looks for family and personal history of HC and heart disease, and physical symptoms of FH, per the Dutch Lipid Clinic Network (DLCN) criteria (Séguro et al.).
e. Phecodes are logical groupings of ICD-9 and ICD-10 diagnosis codes into phenotypes that do or might associate with genotypes. Each phecode represents a single phenotype. For example, there are multiple ICD codes for type 2 diabetes mellitus (T2DM), and particularly in ICD-9, they are not all grouped into 1 base code separate from type 1 diabetes mellitus (T1DM); thus, there is a phecode for T1DM and a phecode for T2DM that group these appropriately.

5. A nonsynonymous substitution is a substitution of one DNA nucleotide for another, such that the codon that the nucleotide is within now codes for a different amino acid. Therefore, the protein which is to be made from the exon region in which the mutation occurs will be different (specifically, its amino acid sequence will be different), potentially in a way that alters the protein's structure and function. Thus, this potential change in protein function could cause disease.