Assessing the Feasibility of Predictive Modeling for HFE-Hereditary Hemochromatosis using Electronic Health Records

Krystal Slattery

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2018

Thesis Committee:

Peter Tarczy-Hornoch, Chair

Deborah Nickerson

Gang Luo

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Assessing the Feasibility of Predictive Modeling
for HFE-Hereditary Hemochromatosis using Electronic Health Records

Krystal Slattery

Chair of the Supervisory Committee:
Chair and Professor Peter Tarczy-Hornoch
Department of Biomedical Informatics and Medical Education

Secondary use of electronic health records allows researchers the opportunity to test hypotheses and gain new insights on complex disease phenotypes. Hereditary hemochromatosis is an inherited autosomal recessive disorder that causes excessive absorption of iron. Early diagnosis and disease management are critical, as iron accumulation in tissue leads to organ failure and eventually death. Diagnosis of hereditary hemochromatosis requires evidence of iron overload and a positive genetic test result. At the University of Washington there are no standard clinical guidelines for hemochromatosis genetic testing and only 7.5% of patients tested have a confirmed diagnosis.

We aimed to identify potential variables for additional screening criteria and inform clinical guidelines for hemochromatosis genetic testing. We found that using established recommendations for genetic testing of hemochromatosis from the American Association for the Study of Liver Diseases (AASLD) and the European Association for Study of the Liver (EASL) on patients screened by their physician for testing would have reduced the number of tested patients from 873 to 345 and maintained 92% of positive diagnoses.

Logistic regression and association rule mining both confirmed that high transferrin saturation is positively associated with HFE-hemochromatosis. It may not be possible to distinguish between hemochromatosis caused by HFE mutations and other genetic variants making a wider hemochromatosis gene panel necessary to identify all cases and discover novel variants.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS AND ABBREVIATIONS

AASLD: American Association for the Study of Liver Diseases

ACMG: American College of Medical Genetics and Genomics. Professional membership organization that establishes standard of care and laboratory policy for Genetic Medicine

AMALGA: Health IT platform produced by Microsoft used to integrate medical data from disparate sources

BMI: Body Mass Index

C282Y: Common HFE variant SNP: rs1800562 missense mutation that results in an amino acid change from Cysteine to Tyrosine at amino acid 282 in the HFE protein.

COMPOUND HETEROZYGOUS: Non-identical non-wild type alleles on homologous chromosomes

EASL: European Association for Study of the Liver

EHR: Electronic Health Record

EMERGE: Electronic Medical Records and Genomics Network. Consortium of U.S medical research institutions.

FER: UW Laboratory Medicine test code for serum ferritin

H63D: Common HFE variant SNP: rs1799945 missense mutation that results in an amino acid change from Histidine to Asparagine at amino acid 63.

HEMDNA: UW Laboratory Medicine test code for HFE genetic testing

HFE: Human Hemochromatosis Gene codes for Human Hemochromatosis protein. Also referred to as TFQTL2, MVCD7, HFE1, or HLAH. OMIM: 613609

HFE-HH: Hereditary Hemochromatosis confirmed to be caused by mutations in HFE either C282Y homozygous or C282Y/H63D compound heterozygous

HH: Hereditary Hemochromatosis

HOMOZYGOUS: Identical alleles on both homologous chromosomes

ICD-9: International Classification of Diseases 9th Revision 1979-1998. In use in Washington State until October 1, 2015 per Washington State Healthcare Authority.

ICD-10: International Classification of Diseases 10th Revision 1998-Present. In use in Washington State after October 1, 2015 per Washington State Healthcare Authority.

IRB: Institutional Review Board

ITHS: Institute of Translational Health Sciences

TIBC: Total Iron Binding Capacity. Measures as Serum Ferritin divided by Serum Transferrin.

TRSATD: UW Laboratory Medicine test code for Transferrin Saturation

UWMC: University of Washington Medical Center

ACKNOWLEDGMENTS

First and foremost, I wish to express my sincere gratitude to the University of Washington for providing access to amazing faculty and world-class research facilities/resources including ITHS.

This work would not be possible without the collective knowledge and boundless support of many people:

- The members of my committee, Dr. Peter Tarczy-Hornoch, Dr. Debbie Nickerson and Dr. Gang Luo. Thank you for your generosity with your time and expertise.

- The members of the Precision Medicine Informatics Working Group. I am humbled by the passion and talent of this group of faculty and fellows. Thank you for providing a forum for discussion and critique. My work benefitted greatly from your intellectual contributions and emotional support.

- My fellow graduate students from the 2016 and 2017 cohorts.There is no other group of people I would rather have cheering me on or debating the merits of Oxford commas with than you. Thank you for taking this journey with me.

Finally, I would like to personally acknowledge some incredible people: my husband Sean, my parents Richard and Shirlin, my sister Heather, my in-laws Peggy and John, my friend and "work mom" Monica Tackett, and all the wonderful people in my life past and present. You all make me who I am everyday. You have supported my passions wherever they took me and gave me the strength to challenge myself. This accomplishment belongs to you as much as myself.

DEDICATION

to Timtom and Mortimer

Chapter 1
INTRODUCTION

Patient electronic health records (EHR) collected through the course of normal medical visits and treatment are a rich source of data for secondary researchers looking for non-invasive methods of hypothesis testing (Shekelle, Morton, & Keeler, 2006). Use of EHR data has been proposed for a wide variety of applications such as public health surveillance (Birkhead, Klompas, & Shah, 2015), quality improvement (Chassin, Loeb, Schmaltz, & Wachter, 2010), automated medical phenotyping (Yu et al., 2015), and development of predictive models (Miotto, Li, Kidd, & Dudley, 2016).

HFE-Hereditary hemochromatosis (HFE-HH) is a common genetic disorder with an autosomal-recessive inheritance pattern where the pathogenic variant of interest occurs in 6% of individuals with European ancestry (Hanson, Imperatore, & Burke, 2001). Despite the relatively high prevalence of homozygous individuals in the population, there is variable penetrance of symptoms and morbidity outcomes (Powell et al., 2006). Diagnosing HFE-HH is challenging because clinical manifestations resemble other disorders, there is a spectrum of severity, presentation is different between genders, and definitive diagnosis requires a positive genetic test (Alexander & Kowdley, 2009; Powell et al., 2006).

The University of Washington department of Laboratory Medicine performs hundreds of genetic tests for HFE-Hemochromatosis each year. Less than 8% of individuals tested have genotypes that confer diagnosis. The goal of this research was to access the feasibility of using EHR data to develop predictive models for HFE-Hemochromatosis.

Chapter 2

BACKGROUND

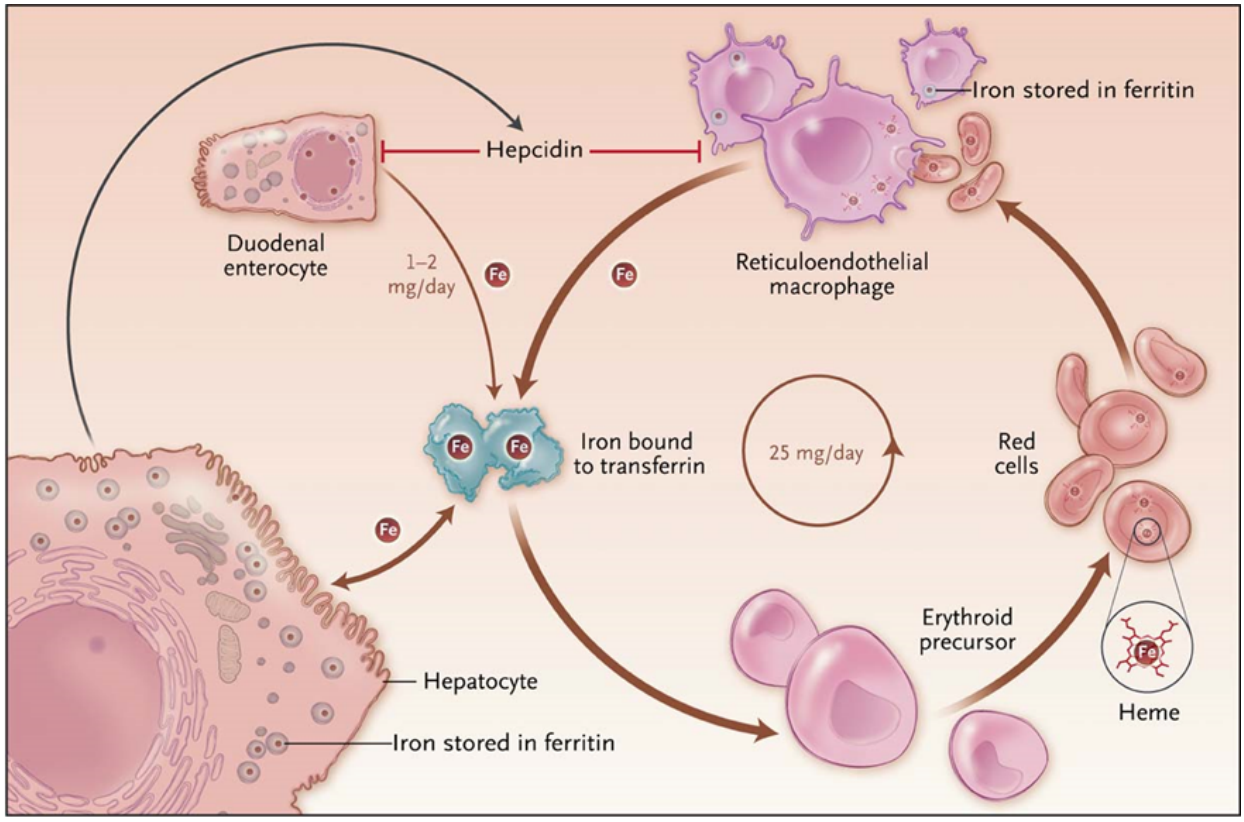
2.1 Heredity Hemochromatosis

Hemochromatosis has been the subject of medical inquiry for well over a hundred years. The condition was formally named in 1889 to describe bronze-stained organs. It was recognized as hereditary in 1935 with an autosomal recessive inheritance pattern in the 1970s (Pietrangelo, 2004). In 1996, the causal gene HFE (OMIM:613619), was first identified. The gene encodes for a MC-I class cell surface protein believed to be involved in cellular iron uptake and regulation (Fleming & Ponka, 2012). Mutations in HFE account for the majority of hereditary hemochromatosis cases, 90% of HH patients are homozygous for the C282Y mutation (Porto et al., 2016).

Hereditary hemochromatosis (HH) is characterized by increased absorption of dietary iron. Healthy individuals absorb 10% of the iron they consume (1-2 mg per day). Individuals with hemochromatosis absorb up to five times more iron from their diet. Iron is necessary for many body processes most notably, it is the central component of the heme molecule in red blood cells. Like many integral chemical components used in the body, there are systems in place to store excess iron for later use but there are no means to excrete excess iron. Iron is only lost through bleeding injury and menstruation in reproductive-age women. Figure 2.1 depicts the iron cycle in the body from absorption via duodenal cells located in the intestine, iron in the red blood cell cycle and storage of excess iron as ferritin in the liver.

Excess iron accumulates preferentially in parenchymal cells of the liver, pancreas and other organs. When left untreated HH can lead to diabetes, heart disease, organ failure and even

Figure 2.1: Depiction of the Iron Cycle



Reprinted from Iron overload in Human Disease.The New England Journal of Medicine, 366(4), 348–359 by Fleming and Ponka (2012).

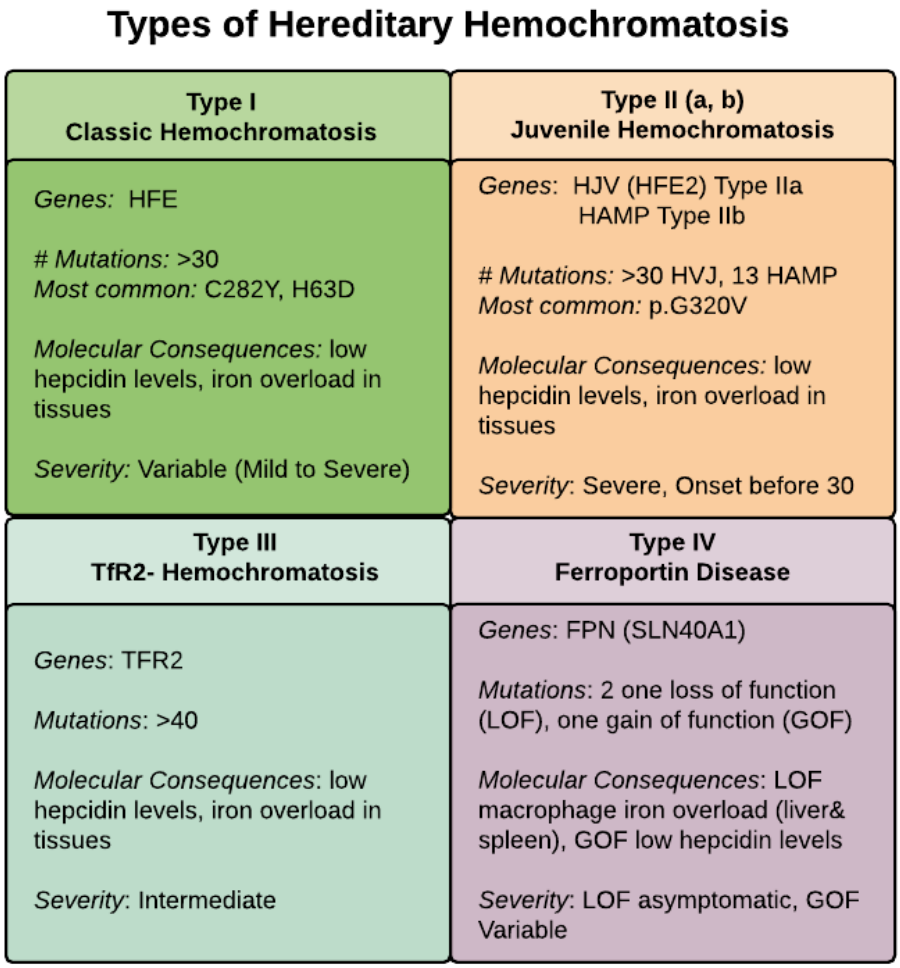
death. Hepcidin is an iron-regulating hormone that binds to and degrades ferroportin to inhibit the release of cellular iron into the bloodstream (Hollerer, Bachmann, & Muckenthaler, 2017). There are a number of genetic mutations found in several cell-surface proteins in the liver that impact hepcidin expression and can manifest as iron overload (Figure 2.2).

2.2 Common HFE Mutations

C282Y is the most common disease causing mutation in the general population with 6% of Caucasians having one allele (heterozygous) and 0.4% having the mutation at both alleles (homozygous) (Merryweather-Clarke, Pointon, Shearman, & Robson, 1997). This mutation is defined as a 845G polymorphism that results in an amino acid substitution from Cysteine to Tyrosine at position 282 in the HFE protein. The amino acid change impacts the protein structure rendering it unable to bind at the cell surface (where it is found in healthy individuals). Instead, the HFE protein circulates intracellularly reducing its signaling capabilities in the Hepcidin iron regulation pathway (Hollerer et al., 2017).

Population geneticists hypothesize the C282Y mutation originated in Celtic populations 5000 years ago and may confer increased fitness in heterozygotes (much like malaria resistance for Hb S in Sickle Cell). Studies of C282Y heterozygous carriers found they are taller, have longer life expectancies, are more athletic, and have higher rates of fertility (Balistreri et al., 2002; Bulaj, Griffen, Jorde, Edwards, & Kushner, 1996; Cippà & Krayenbuehl, 2013). It has been speculated that the timing of this mutation may coincide with the transition in early human populations from a high protein and iron rich diet from hunting and gathering to an iron-poor grain based diet as farmers. The ability to absorb more of what little iron was available would have proven advantageous and caused the mutation to be perpetuated in this population. Given the large number of identified mutations in iron absorption pathway genes, it is likely there was significant selective pressure for these changes to occur.

Figure 2.2: Types of Hereditary Hemochromatosis



Adapted from Figure 1 Pathophysiological consequences and benefits of mutations: 20 years of research. Haematologica, 102(5), 809–817.by Hollerer et al. (2017).

H63D is another prevalent mutation in HFE (15% of Caucasians are heterozygous). The impact of homo- and heterozygosity of this mutation on HH is highly debated (Paul C. Adams, 2014; Hollerer et al., 2017). When H63D is found with C282Y in an individual, the genotype is referred to as a compound heterozygote (two non-wild type alleles) and is considered to a genotype consistent with HH during testing.

2.3 Diagnosis

Diagnosis of HFE-HH is challenging, as clinical manifestations resemble other disorders. The ubiquity of serum iron in biological processes means that clinical symptoms of HH are often mosaic and can impact any of the following systems: neurological, gastrointestinal, musculoskeletal, dermatological, endocrine, and cardiovascular. Known symptoms include chronic fatigue, skin pigmentation, stiffness or pain in joints, diabetes, impotence, and liver disease (including fibrosis, cirrhosis, and cancer). HFE-HH diagnosis is more prevalent in men and postmenopausal women. Regular menstruation in premenopausal woman removes excess iron, reducing the risk of iron overload, and masks increased iron absorption.

Penetrance of HFE-HH among individuals with the homozygous C282Y genotype is difficult to determine as diagnosis requires both confirmed genetic testing and either onset of symptoms or blood tests indicating iron overload (P. C. Adams, 2015). Meta-analysis of 16 studies have suggested penetrance around 14% (European Association For The Study Of The Liver, 2010). Low penetrance poses a challenge to genetic testing as the majority of people (86%) who are homozygous for C282Y appear symptom free. Homozygous individuals can range from symptom free to severe. This range makes it likely many mild cases are not captured.

Laboratory blood tests that reveal elevated liver enzymes, ferritin and/or transferrin saturation can occur before clinical symptoms develop, making preventive treatment possible with early detection. While blood tests for these markers are inexpensive, they suffer from low specificity and/or low sensitivity to detect HH. Elevated ferritin most often indicates

acute or chronic inflammation, chronic alcohol consumption, liver disease, renal failure, or metabolic syndrome, rather than iron overload (Koperdanova & Cullis, 2015). Therefore elevated ferritin exhibits low specificity for HH screening. Transferrin Saturation is reported as a percentage and is calculated by taking the total serum ferritin and dividing it by the total iron binding capacity (determined by serum levels of transferrin).

Iron uptake is only capable in the body when ferritin is bound to a transport molecule such as transferrin. Serum ferritin reports the amount of iron circulating in blood as ferritin but the transferrin saturation is a better measure for ferritin intended for storage in tissue or reuse in red blood cell generation. When transferrin molecules are saturated, ferritin binds to other molecules such as citrate with lower molecular weight. The ferritin bound to lower molecular weight transporters have increased rate of uptake by certain cell types (Fleming & Ponka, 2012). Most prevailing genetic testing algorithms for HH use transferrin saturation as a key benchmark. However, transferrin saturation has low sensitivity in premenopausal women and is prone to high variability based on time of date and fasting conditions.

Early detection of HH greatly improves clinical outcomes. Treatment for iron overload is simple and inexpensive. Historically treatment has consisted of phlebotomy (frequency dependent on level of iron overload with serum ferritin >1000 being referred for liver biopsy). Recently, medications that act as iron chelators have also been introduced as an alternative for individuals who do not tolerate phlebotomy well.

A recently published study found that 1 in 10 males with HH will develop severe liver disease in their lifetime (Grosse, Gurrin, Bertalli, & Allen, 2018). Early diagnosis and treatment can prevent irreparable organ damage. HFE-HH patients are significantly more likely to suffer cardiac myopathy (Allen et al., 2008) and 9 times more likely to develop cirrhosis of the liver if they consume excess alcohol (Fletcher, Dixon, Purdie, Powell, & Crawford, 2002) than individuals without this condition. Comorbidity of hemochromatosis with diabetes has shown

to confer a 7-fold increased risk of death from diabetes (Niederau et al., 1985) and a 3-fold increased risk in all-cause mortality for C282Y homozygotes (Ellervik, Mandrup-Poulsen, Tybjærg-Hansen, & Nordestgaard, 2014). Diabetes is found in 8.5% of adults worldwide, further increasing the need for early detection of HH to prevent additional mortality burden (Mathers & Loncar, 2006).

Several studies have evaluated the need for population-based screening for hemochromatosis (Burke et al., 1998; Motulsky & Beutler, 2000; Pardo, 2000; Sánchez et al., 2003) however all concluded that the cost and uncertain penetrance makes population screening unfeasible. Due to low and/or uncertain penetrance, none of the mutations in genes associated with hemochromatosis appear on the ACMG list of medically actionable variants requiring secondary return of results. This means that if a patient undergoes genetic testing for research purposes or another genetic condition, any findings about their HFE genotype will not be returned to the patient. Some authors have advocated for targeted population screening for HFE-HH such as HFE testing for all adults with type-II diabetes (Barton & Acton, 2017) or to include addition of C282Y as medically actionable (Grosse et al., 2018). Nonetheless, current genetic testing requires physicians to have a clinical suspicion of primary (hereditary) iron overload and to recommend patients for confirmatory genetic testing.

2.4 HFE-Testing Algorithms

The American Association for the Study of Liver Diseases (AASLD) and the European Association for Study of the Liver (EASL) recommend genetic testing for HFE-HH for all patients with abnormal organ findings (such as elevated liver enzymes) who present with transferrin saturation greater than 45% regardless of manifestation of symptoms and for all first degree relatives of patients diagnosed with HFE-HH (Vanclooster et al., 2015). These recommendations are based on the desire to identify all cases of HFE-HH and reduce the need for potentially more invasive tests such as liver biopsies.

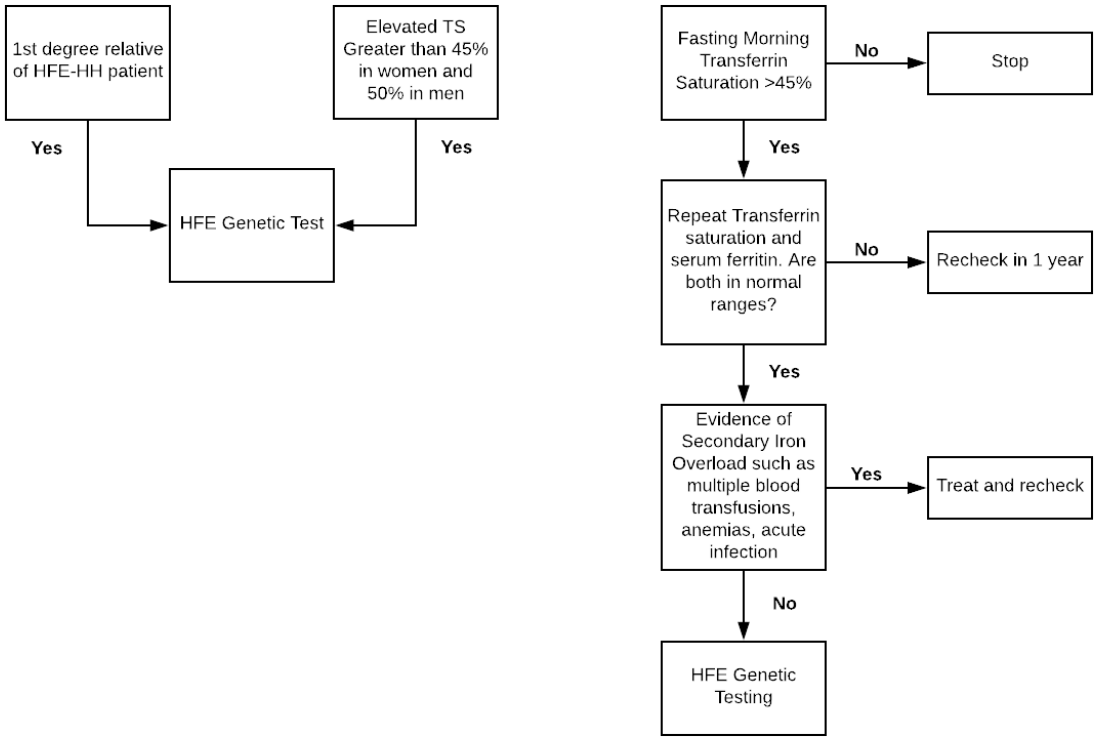
Prior to the development of genetic testing, the only way to diagnosis HH was through symptom manifestation. The development of genetic testing algorithms highlight the desire to take advantage of potential early detection of HH prior to irreversible organ damage balanced with an evidence-medicine based approach to resource management. Some older algorithms have also listed abnormal ferritin levels ($>$ 200 ng/mL for females and $>$300 ng/mL for males) as criteria for testing. However recent research studies have consistently shown abnormal ferritin to be a poor predictor of HH (O'Toole, Romeril, & Bromhead, 2017). Figure 2.3 shows two potential testing algorithms. The one on the left is a simplified model based on the recommendations from AASLD and EASL. The diagram on the left is the algorithm currently in use for genetic testing at the Mayo clinic. It acknowledges that abnormal transferrin saturation requires additional follow-up but accounts for the variability found in this blood test by requiring two fasting values that are abnormal a year apart.

2.5 EHR Data

The widespread adoption and expansion of EHRs have allowed patient data once contained in individual charts to be quickly aggregated for secondary use research. For the first time in medical history it is possible to use aggregate patient data to identify individuals or cohorts, classify patients or conditions, and potentially predict patient outcomes (Blumenthal & Tavenner, 2010). The benefits of using EHR data is that it is already routinely collected as part of medical visits, it contains much of the data historically used for phenotyping, and it is relatively accessible to researchers.

EHR data could be used to improve phenotyping of HFE-HH. Previous attempts at medical phenotyping using EHR data have identified challenges with data extraction, missing data, and low signal-to-noise ratio, depending on the research question (Denny, 2012; Hripcsak & Albers, 2013; Jiang et al., 2011). Acknowledging the breadth of challenges in using EHR data, it is necessary to adapt traditional methodologies and develop novel hybrid workflows. Machine learning techniques have proven robust in light of missing data and more effective

Figure 2.3: Examples of HFE-testing Algorithms



Left current HFE-testing recommendation from AASLD and EASL. Right, current HFE-testing algorithm employed at Mayo Clinic Medical laboratories.

at extracting signals than traditional statistical methods (Dekel & Shamir, 2008; Kotsiantis, Zaharakis, & Pintelas, 2006; Schlimmer & Granger, 1986).

Machine learning techniques are widely used when analyzing large datasets. Much like human physicians looking for common symptoms or traits to phenotype a disease, machine learning algorithms are capable of looking for patterns in complicated, noisy datasets. Using a training set of data, an algorithm can learn to classify a particular state and then identify it in a new unseen dataset. There have been numerous studies that have employed machine learning techniques on EHR datasets to classify patients or traits associated with a wide-variety of diseases such as diabetes (Kho et al., 2012), arrhythmia (Ritchie et al., 2013), and bipolar disorder ((Ritchie et al., 2013; Shivade et al., 2014).

Machine learning techniques have also been utilized when looking for associations between genotype and phenotype however these studies used genetic data in an attempt to identify phenotypes (Kohane, 2011);(Gallego et al., 2015). To date no machine learning studies have attempted the reverse, using machine learning and phenotype data to look for underlying genotype.

The most relevant study in the space of EHR genetic phenotyping was conducted in 2013 through the eMERGE network, a consortium of research centers that performed microarray genotyping assays and sequencing on large cohorts of patients with linked EHR records. The eMERGE database includes 39,000 patients with genetic data and EHR of whom 100 were homozygous for HFE. The study attempted to use ICD-9 codes and genotype data to determine the percentage of individuals in the database who were homozygous for HFE mutations and had been diagnosed for HFE-HH (Gallego et al., 2015). The study found that only 20% of patients with the homozygous HFE genotype had been diagnosed with HFE-HH, meaning that they met the clinical criteria for testing and tested positive prior to their inclusion in the eMERGE project.

In addition to providing more evidence on incomplete penetrance of HFE-HH, the study also produced a list of features associated with C282Y-homozygous individuals regardless of their clinical diagnosis. These features, combined with previous literature describing the medical phenotype and comorbidities of hemochromatosis, informed the data obtained and features produced for analysis. While the study found classic hemochromatosis symptoms such as liver disease and arthritis associated with undiagnosed C282Y homozygous individuals, the study did not return genetic results in accordance with the ACMG recommendations for return of results.

Chapter 3

METHODS

3.1 Selection of Study Subjects

Initial identification of a cohort of University of Washington patients with HFE-Genetic Testing was performed using the University of Washington Medicine LEAF tool in the IRB approved de-identified preparatory research mode. This mode provides a de-identified aggregate database of patients that is searchable by condition or test result. Using the laboratory medicine code HEMDNA for the Hemochromatosis genetic test, the LEAF tool identified 901 potential patients with genetic test results (Fig 3.1) and additionally another approximately 6,000 patients with transferrin saturation greater than 45%.

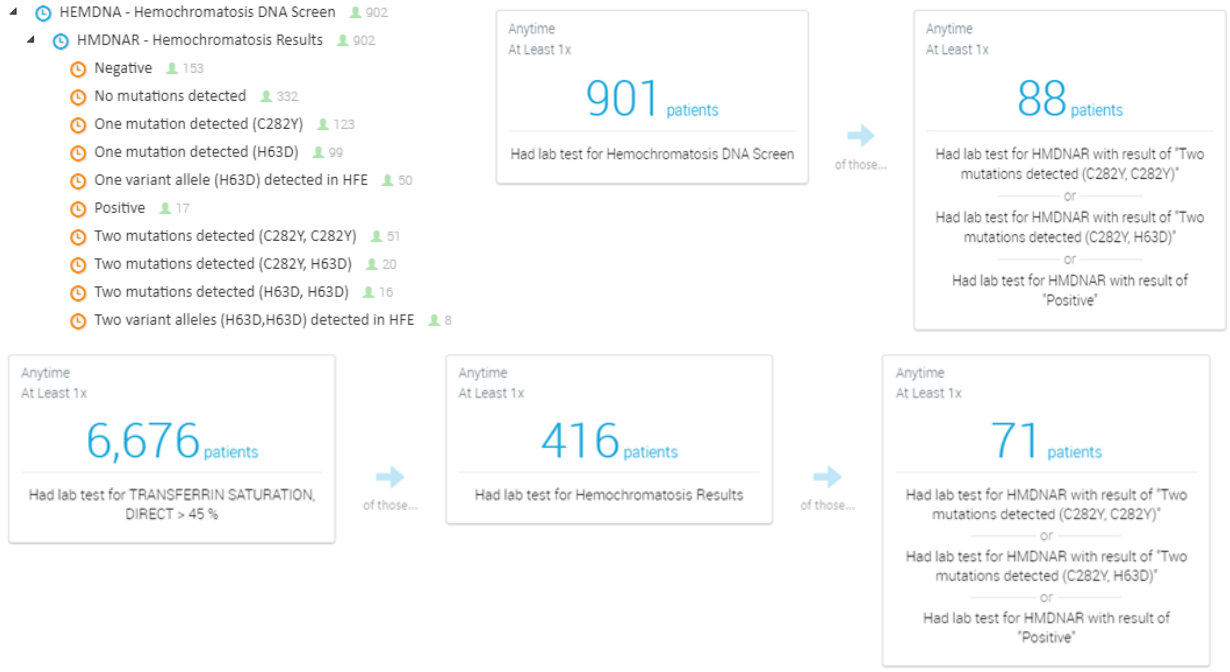
3.2 IRB Exemption

An IRB exemption (University of Washington, Seattle,WA) was obtained for the de-identified demographic, vital, laboratory, diagnosis and medication data on 7,766 UW medicine patients: 873 who had undergone genetic testing for hemochromatosis and 6,893 who had laboratory testing that included total iron binding capacity (TIBC) with transferrin saturation values. All patients were between the ages of 18-84. None were members of protected classes as outlined in by IRB. Data extraction from AMALGA database was performed by ITHS. Table 1 shows descriptive statistics for all study subjects.

3.3 Data Processing

Patient data was received from ITHS in relational matrices by data type (demographics, vital signs, laboratory results, diagnosis codes, and medications). For each individual: height in cm and calculated BMI (body mass index) were extracted from vitals; results for FER (Fer-

Figure 3.1: Cohort Identification with LEAF



Screenshots of HEMDNA and Transferrin Saturation Results using ITHS LEAF tool

Table 3.1: Descriptive Statistics of Cohorts

	N	HFE-HH Positive	HFE-HH Negative	Abnormal TIBC	Normal TIBC	Combined
		$N = 66$	$N = 807$	$N = 3010$	$N = 3883$	$N = 7766$
Gender : Female	7765	42% (28)	40% (325)	47% (1494)	50% (1869)	48% (3716)
Age	7766	46 56 64	44 55 64	46 59 67	40 56 67	43 57 67
Race : White	7766	86% (57)	71% (571)	69% (2173)	68% (2528)	69% (5329)
Black		3% (2)	5% (42)	8% (264)	11% (404)	9% (712)
Asian		0% (0)	11% (91)	11% (332)	9% (335)	10% (758)
Hispanic		0% (0)	1% (6)	1% (33)	1% (29)	1% (68)
Other		0% (0)	4% (30)	5% (150)	3% (120)	4% (300)
Unknown		11% (7)	8% (67)	7% (209)	8% (316)	8% (599)
Abnormal_Iron	7249	68% (43)	67% (452)	97% (2824)	8% (299)	50% (3618)
FER_max	7249	192 425 926	155 460 1077	566 1140 2241	28 65 140	59 248 1026
Abnormal_TransferrinSat	7608	92% (56)	51% (331)	100% (3161)	0% (0)	47% (3548)
TRSATD_max	7608	61 86 92	30 49 81	59 77 89	13 21 29	20 40 74
Height_CM_Mean	7385	167 173 180	163 171 178	162 169 177	162 170 178	162 170 177
BMI_mean	7279	24.7 26.7 31.1	24.1 27.1 31.9	23.1 26.4 30.8	23.4 27.0 31.7	23.3 26.8 31.3
Appt_total	7765	17.5 48.5 96.5	16.0 40.0 98.5	31.0 89.0 174.0	18.0 50.0 113.0	21.0 61.0 139.0
Record_Len_Days	7765	678 1587 2499	404 1376 2877	475 1400 2616	612 1690 2977	521 1556 2822
ICD9_total	7765	60 155 344	62 162 399	153 421 847	70 191 454	88 255 612
Infection	7766	41% (27)	59% (478)	75% (2374)	53% (1992)	63% (4871)
Cancer	7766	48% (32)	44% (355)	65% (2042)	44% (1655)	53% (4084)
Diabetes	7766	21% (14)	18% (145)	26% (812)	20% (728)	22% (1699)
Anemia	7766	26% (17)	48% (386)	84% (2664)	66% (2469)	71% (5536)
SexOrgan_Disorder	7766	36% (24)	32% (261)	36% (1153)	41% (1545)	38% (2983)
Arthritis	7766	68% (45)	64% (513)	72% (2284)	69% (2562)	70% (5404)
DegenerativeCNS	7766	6% (4)	11% (87)	15% (477)	15% (545)	14% (1113)
Mental_Health_Disorders	7766	64% (42)	56% (451)	60% (1902)	56% (2076)	58% (4471)

$_a\ b\ _c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. N is the number of non–missing values. Numbers after percents are frequencies.

ritin), TRSATD (Transferrin Saturation) and HEMDNA (Hemochromatosis genetic testing) from laboratory results; age, gender, race and unique de-identified patient identifier from demographics; and date and diagnosis code from diagnoses. Repeated laboratory results were summarized into count, minimum value, mean value, maximum value.

The number of appointments per patient was extrapolated as the number of unique dates for which there was a diagnosis code. The record length in days was determined to be the length of time between the first diagnosis date and the last diagnosis date for each patient. Diagnosis codes in the dataset included both ICD-9 and ICD-10 diagnosis codes (as the data spanned from 2009 to the present). ICD-9 and ICD-10 codes were converted from their decimal forms to short forms using R package icd and then ICD-10 codes were backwards crosswalked to ICD-9 codes using the R package icdcoder.

12,285 unique ICD-10 codes were present in the initial dataset and 95% were successfully mapped to ICD-9 codes. The 587 unmapped ICD-10 codes were removed from the dataset. Including converted ICD-10 codes, the dataset included 9,171 unique ICD-9 codes. ICD-9 E and V codes (supplemental codes for external injury and health status) were removed from the dataset leaving 7,027 unique ICD-9 codes.

3.4 Feature Selection

Data features were informed by previous literature based on known symptoms, positive correlation, negative correlation, and risk factors. Table 3.2 shows the broad category of features selected prior to additional analysis.Continuous variables such as age, height, BMI, appointment number and record length were assessed using descriptive statistics and then binned to make categorical variables based on frequency (age, appointment number, record length) or biologically relevance (Ex: BMI; Underweight, Normal, Obese, Morbidly Obese). Table 3.3 shows methodology for each continuous variable.

Table 3.2: Features Selected from Literature Review

Variable	Source	Rationale	Citation
Age	Demographics	Onset of Diagnosis	Adams et al.,1997
Sex	Demographics	Differential Prevalence	Allen et al., 2008
Race	Demographics	Differential Prevalence	Adams et al.,2005
BMI	Vital Signs	Positive Correlation	Hollerer et al., 2017
Height	Vital Signs	Positive Correlation	Cippà, P. E., & Krayenbuehl, P.-A., 2013
Serum Ferritin	Laboratory result	Symptom of HH	Fleming, R. E., & Ponka, P., 2012
Transferrin Saturation	Laboratory result	Symptom of HH	Fleming, R. E., & Ponka, P., 2012
Arthritis	ICD-9 Codes	Symptom of HH	Hollerer et al., 2017
Sexual Organ Disfunction	ICD-9 Codes	Symptom of HH	Hollerer et al., 2017
Depression/ Mental Health	ICD-9 Codes	Symptom of HH	Hollerer et al., 2017
Diabetes	ICD-9 Codes	Symptom of HH	Hollerer et al., 2017
Cancer (Any type)	ICD-9 Codes	Increased Risk	Hollerer et al., 2017
Infection	ICD-9 Codes	Increased Risk	Hollerer et al., 2017
Neurodegenerative diseases	ICD-9 Codes	Decreased Risk	Fleming, R. E., & Ponka, P., 2012
Anemia	ICD-9 Codes, Laboratory Results	Neg correlation	Fleming, R. E., & Ponka, P., 2012

Table 3.3: Variable Transformation Methodology

Variable	Original Format	Chosen Format	Methodology
Age	Continuous	Categorical	10 years per bin except the ends
BMI	Continuous	Categorical	Underweight, Normal, Overweight, Obese
Height	Continuous	Categorical	3 inches per Bin
Serum Ferritin	Continuous	Categorical	Anemic, Normal, High, Very High, Extreme
Transferrin Saturation	Continuous	Categorical	Low, Normal, Abnormal, Extreme
Arthritis	Count	Binary	At least 1 ICD-9 Code in binned category
Sexual Organ Disfunction	Count	Binary	At least 1 ICD-9 Code in binned category
Depression/ Mental Health	Count	Binary	At least 1 ICD-9 Code in binned category
Diabetes	Count	Binary	At least 1 ICD-9 Code in binned category
Cancer (Any type)	Count	Binary	At least 1 ICD-9 Code in binned category
Infection	Count	Binary	At least 1 ICD-9 Code in binned category
Neurodegenerative diseases	Count	Binary	At least 1 ICD-9 Code in binned category
Anemia	Mixed	Binary	Mean Ferritin value less than 20ng/mL or at least 1 ICD-9

Table 3.4: Diagnosis Variables

		HFE-HH Pos vs. Neg		HFE-Tested vs. Non-tested	
Variable	Definition	ChiSq	p-value	ChiSq	p-value
Infection	ICD-9 codes 0-139.999	$\chi_1^2 = 8.4$	P=0.0041	$\chi_1^2 = 10$	P=0.002
Cancer (Any type)	ICD-9 codes 140-239.999	$\chi_1^2 = 0.5$	P=0.481	$\chi_1^2 = 26.9$	P<0.0011
Diabetes	ICD-9 codes 250-250.999	$\chi_1^2 = 0.43$	P=0.5111	$\chi_1^2 = 7.73$	P=0.005
Anemia	ICD-9 codes 280-285.999 or Mean ferritin less than 20 ng/mL	$\chi_1^2 = 11.96$	P<0.0011	$\chi_1^2 = 303.26$	P<0.0011
Sexual Organ Disfunction	Males: ICD-9 600-608.999, Females: 614-629.999	$\chi_1^2 = 0.45$	P=0.5031	$\chi_1^2 = 13.82$	P<0.0011
Arthritis	ICD-9 codes 710-729.999	$\chi_1^2 = 0.56$	P=0.4531	$\chi_1^2 = 14.93$	P<0.0011
Neurodegenerative diseases	ICD-9 codes 330-337.999	$\chi_1^2 = 1.46$	P=0.2281	$\chi_1^2 = 12.23$	P<0.0011
Depression/ Mental Health	ICD-9 codes 295-316.999	$\chi_1^2 = 1.49$	P=0.2221	$\chi_1^2 = 0.49$	P=0.485

ICD-9 codes were coarsely binned to match features outlined in Table 2 based on ICD-9 hierarchy. Fisher exact calculations were performed to evaluate differences between the population of patients who tested positive and those who tested negative, as well as between the total population tested and those with abnormal TIBC laboratory results. Table 3.4 outlines the ICD-9 codes used for each variable and the P value for the two fisher exact calculations performed.

3.5 Logistic Regression

Logistic regression models the odds ratio of a binary outcome. It is possible to rewrite a logistic regression equation to calculate probability of a binary outcome from 0 to 1.

$$\frac{1}{\left(1 + e^{-(\beta_0 + \beta_i x_i + \beta_j x_j + ... \beta_n x_n)}\right)}$$

In the above equation, a logistic regression model will fit data to produce the coefficient betas for each variable x. The logistic regression assumes a linear combination of variables but is estimated using maximum likelihood. Due to the small sample size, logistic regression was performed using a penalized likelihood method designed to reduce bias in maximum likelihood estimates (Firth, 1993; Heinze, Ploner, & Beyea, 2013). The Firth method was developed to handle separation that can occur then the outcome of interest is a rare event.

Historically it has been used in rare cancer datasets as well as gene variant analysis. Additionally it allows for continuous variables.

Previous studies have attempted to predict C282Y homozygosity among individuals with iron overload using simple logistic regression with continuous variables serum ferritin and transferrin saturation, however the resulting model reported very wide confidence intervals. Using the published regression, a patient with 50% transferrin saturation and 500 ng/mL serum ferritin had a 1.3% (95% CI 1.1% to 8.8%) of being C282Y homozygous (Lim, Speechley, & Adams, 2014). This study benefited from a large sample size but the equation is likely to underestimate the probability of C282Y homozygosity in patients who are recommended to receive TIBC testing due to abnormal liver enzymes. TIBC is not a routine laboratory blood test and is generally only recommended in cases of suspected anemia or iron overload. The published equation and accompanying calculator were based on population screening which may not be clinically relevant.

Two logistic regression models were built: one to capture the probability of C282Y homozygosity among those recommended for HFE-testing and the second to capture the probability of being recommended for testing among patients who have received TIBC blood tests. The descriptive statistics revealed significant differences between the HFE-tested population and the TIBC-tested population as a whole (Table 3.4). The purpose of two tests would be to identify individuals in the untested cohort who are more similar to the tested cohort and then to determine their probability of having C282Y homozygosity.

The models were built using a combination of features including demographics, laboratory results, vitals and diagnosis information. The model was developed iteratively using anova to compare the fit between models. All logistic regression was conducted using the R package logistf.

3.6 Machine Learning

Traditional machine learning approaches operate best with large datasets. They are robust to handle missing and sparse data, but they require a lot of individuals in the sample to perform any sort of classification. This dataset contained over 7,000 individuals with rich data, but there was only 66 gold-standard confirmed HFE-HH cases. Ideally a classifier would be built to distinguish between confirmed HFE-HH and all other patients. But this poses a problem of unlabeled data. Of our 7,000 individuals, there is not HFE genotype data for 6,000 of them. While there have been some methodologies developed to classify with unknown labels, these techniques are very recent and require a significant number of samples.

Additionally, prediction or classification using machine learning algorithms require gold standard training sets and a separate unseen test set for validation. There are methods designed to cross-validate, that rely on splitting the data such that each sample may be for both training and validation (but not in the same set). The very small number of confirmed cases did not allow cross-validation as it sacrifices power to detect signal for ability to evaluate. With a sufficiently large dataset this trade-off may reduce the fit slightly but in this case it rendered the task impossible.

3.7 Association Rule Mining

Association rule mining is a machine learning technique originally developed in 1993 to draw inferences from supermarket transactions. Each supermarket transaction consisted of one or more items purchased together by an individual and the methods were developed to identify patterns of items that were frequently bought together. It is a "bottom-up" approach where no starting insight is provided and instead frequency of items is used to guide associations (Agrawal, Imieliński, & Swami, 1993; Ordonez, 2006).

The apriori algorithm starts by identifying frequent items in the dataset that meet a minimum support standard. The minimum is generally 1% but can be as low as 0.5%. This means that 1% support for a HFE-Case would correspond to at least six individuals sharing a trait in a dataset of 600. Confidence is a parameter to determine the proportion of individuals meeting the criteria within a frequent set. A confidence level of 0.5 is generally used for medical applications (Ordonez, 2006). These two parameters mean identifying variables with a minimum of six cases and no more than an equal number of non-cases. These sets are built for all variables and the combined iteratively one at a time. Each combination or variables must meet the parameters until all combinations have been identified.

The apriori algorithm is exhaustive, which can be computationally expensive for large datasets, but was not a limitation given the number of individual patients in this study. Association rule mining on EHR data has the potential for rules to be validated by medical knowledge (Li, Simon, Chute, & Pathak, 2013; M.Kang'ethe, Kang'ethe, & Wagacha, 2014). Valid patterns in the data should mimic underlying biological mechanisms and elucidate which features in combination are necessary to confer a phenotype. This is particularly crucial for HH as there is not a clear distinct phenotype established for this condition. Association rules were mined using the apriori function of the R package arules with minimum support=0.01 and confidence=0.5.

Chapter 4

RESULTS

4.1 Transferrin Saturaion

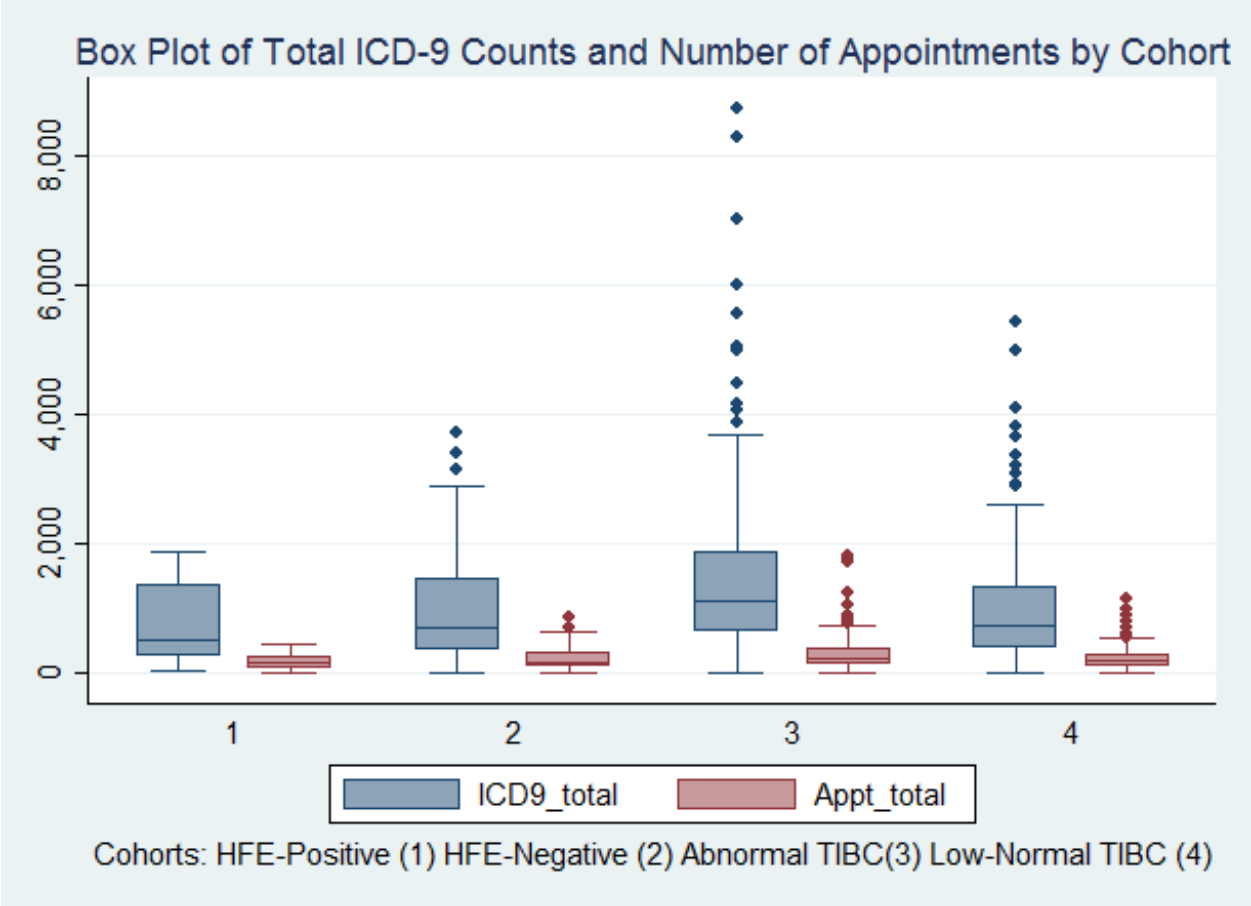
Transferrin saturation was the most prominent variable associated with HFE-HH diagnosis. Despite clear recommendations for abnormal transferrin saturation prior to genetic testing referral, only 715 out of 873 patients had at least one transferrin saturation laboratory finding in the dataset. Of these 715 only 48% (345/715) had values above the established threshold. 92% of confirmed HFE-HH cases had abnormal transferrin saturation.

Descriptive statistics, logistic regression and association rule mining all confirmed that transferrin saturation is a key variable for HFE-HH prediction. If abnormal transferrin saturation was required as secondary screening prior to genetic testing, it would have reduced the total population tested to 345 and captured 56 out of 61 positive cases. This additional screening has a specificity of 91% but a sensitivity of 46%. It is not possible to compare this to the current protocol as we don't have a value for true positives in the general population. Current positive predictive power is 7.5% but this additional screening would increase the positive predictive power to 12.5%.

4.2 HFE-tested Cohort

HFE-HH cases and those who tested negative were similar across most features selected for analysis. The only statistically significant differences between those who tested positive and negative were self-reported race, transferrin saturation, infection and anemia (Table 3.4). The HFE-test cohort however was quite different from the untested cohort in nearly all diagnosis code categories (Table 3.4) as well as self-reported race, gender, height and

Figure 4.1: Comparisons of Total ICD-9 and Appointments by Cohort



laboratory results. The cohort of patients with abnormal TIBC results who were not HFE-tested had significantly higher number of diagnosis codes and appointments suggesting a higher overall disease burden (Figure 4.1).

4.3 Logistic Regression Models

Penalized logistic regression on the cohort of HFE-HH tested patients revealed abnormal transferrin saturation was the best predictor of HFE-HH diagnosis. Anova comparison between penalized logistic regressions with a single variable, transferrin saturation, and the linear combination of transferrin saturation and ferritin value found transferrin saturation

alone was a better fit (p=0.04).

Applying transferrin saturation alone on our cohort of HFE-HH tested patients found a transferrin value of 50 corresponded to a probability of 5.7% of being HFE-HH positive (95% CI: 1.6-17.6%). Binning maximum transferrin saturation into a binary value of abnormal (>45% for females, >50% for males) slightly increased the likelihood ratio and may better reflect the reality that difference in abnormal values do not accurately reflect differences in probabilities.

While the best fit came from transferrin saturation alone, this does not provide a tool to identify new candidates for screening from the abnormal TIBC cohort and therefore additional variables were evaluated (Table 4.1). Table 4.1 depicts all variables tested in logistic regression models and their corresponding p-values. Seven variables were (bolded) were shown to be statistically significant with a threshold of 0.05.

The logistic regression model to predict HFE-testing status identified 7 variables: gender, abnormal ferritin, abnormal transferrin saturation, total ICD-9 diagnoses, cancer diagnosis, and anemia. Table 4.2 displays the coefficients for the model along with their standard errors, confidence intervals and p-values. While Firth penalized logistic regression attempts to eliminate separation, two variables and the intercept still exhibit separation (chisq values approaching infinity). The logistic regression model to identify the HFE-testing cohort had sensitivity of 70% and specificity of 60% when using a probability cut-off of 0.1 or higher.

4.4 Association Rule

Nine total rules, each with between four and six variables, were found to be associated with HFE-HH patients among those with genetic test results (Table 4.3). The majority of variables included in the rules appeared in multiple rules and had biological rationale (Table 4.4). Abnormal transferrin saturation (High TRSATD) appeared in 8 out of the 9 rules. The combination of high transferrin saturation, male, white, and age 51-60 match the canonical HH

Table 4.1: Logistic Regression Variables Tested using HFE-Tested Cohort

Variable	ChiSq	df	P-value
Abnormal_TransferrinSat	**49.39**	**1**	**2.10E-12**
Anemia	**13.28**	**1**	**2.68E-04**
Infection	**8.69**	**1**	**3.20E-03**
Diabetes	**5.53**	**1**	**1.87E-02**
BMI_mean	**5.25**	**1**	**2.19E-02**
ICD9_total	**4.87**	**1**	**2.73E-02**
Appt_total	**4.72**	**1**	**2.99E-02**
Height_CM_Mean	3.34	1	6.74E-02
Gender	2.29	1	1.30E-01
Age	1.81	1	1.79E-01
Race	6.86	5	2.31E-01
Mental_Health_Disorders	1.25	1	2.63E-01
Abnormal_Iron	0.24	1	6.26E-01
Record_Len_Days	0.21	1	6.47E-01
Cancer	0.11	1	7.42E-01
Arthritis	0.02	1	8.84E-01
DegenerativeCNS	0.01	1	9.01E-01

Table 4.2: Likelihood Penalized Logistic Regression Coefficients for Predicting HFE-Testing

Variable	Coef	SE(coef)	Lower 0.95	Upper 0.95	Chisq	p-value
(Intercept)	-2.18	0.10	-2.39	-1.98	Inf	0.00E+00
MaleTRUE	0.25	0.09	0.09	0.42	8.78	3.05E-03
Abnormal_IronTRUE	1.77	0.14	1.50	2.03	Inf	0.00E+00
Abnormal_TransferrinSatTRUE	-0.55	0.13	-0.80	-0.30	18.14	2.05E-05
ICD9_total	-3.00E-04	0.00	0.00	0.00	10.30	1.33E-03
CancerTRUE	-0.24	0.09	-0.42	-0.06	6.62	1.01E-02
AnemiaTRUE	-1.48	0.10	-1.68	-1.29	Inf	0.00E+00
Mental_Health_DisordersTRUE	0.36	0.09	0.18	0.53	15.57	7.94E-05

Likelihood Ratio Test=474.1 on 7 df, p-0, n=7179

Wald test=463.4 on 7 df, p=0

Table 4.3: Association Rules Generated from HFE-HH Tested Patients

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	support	confidence	lift	count
High TRSATD	High FER_Max	ICD9_25-100	Arthritis			0.01	0.5	5.29	8
High TRSATD	Age_51-60	Male	Record_36-60 months			0.01	0.5	5.29	7
High TRSATD	Age_51-60	Male	Tall_STD1			0.01	0.54	5.69	7
High TRSATD	Age_51-60	Race_White	Tall_STD1			0.01	0.5	5.29	7
High TRSATD	**Age_51-60**	**Male**	**Race_White**	**Tall_STD1**		**0.01**	**0.58**	**6.17**	**7**
High TRSATD	High FER_Max	BMI<18.5	ICD9_25-100	Arthritis		0.01	0.5	5.29	8
High TRSATD	High FER_Max	Race_White	ICD9_25-100	Arthritis		0.01	0.54	5.69	7
High TRSATD	High FER_Max	Race_White	ICD9_25-100	BMI<18.5	Arthritis	0.01	0.54	5.69	7
Age_51-60	Race_White	High FER_Max	Appt_25-100	BMI<18.5	Arthritis	0.01	0.5	5.29	7

phenotype described in historic literature. The rule with the highest lift, a measurement to describe the combination of support and confidence, has five variables, highlighting the need for a combination of variables. The only variable that was a diagnosis type was arthritis, which is a symptom of HH.

Applying the nine association rules on a subset of untested patients who had 10% or greater probability of being tested yielded 179 individuals including two who met the criteria for all nine rules. 42 cases out of 66 total also had 10% or greater probability of being tested using the logistic regression model and the nine association rules captures 21 of them (50%). This corresponds roughly to 50% specificity and 80% sensitivity using a hybrid approach to identify HFE-cases.

Table 4.4: Prevalence and Rationale for Association Rule Variables

Variable	# Rules	Variable Type	Biological Rationale
High TRSATD	8	Laboratory Result	Iron Overload
High FER_Max	5	Laboratory Result	Iron Overload
Arthritis	5	Diagnosis	HH Symptom
Age_51-60	5	Demographic	Typical Age of Onset
Race_White	5	Demographic	Increased Prevalence
ICD9_25-100	4	Morbidity Level	
Male	3	Demographic	Increased Prevalence
Tall_STD1	3	Vital	Positive Correlation
Underweight_BMI<18.5	3	Vital	
Record_36-60 months	1	Record Type	
Appt_25-100	1	Morbidity Level	

Chapter 5

DISCUSSION

The challenges of identifying HH through medical phenotyping have been previously reported. The lack of available datasets that include both rich medical phenotype data such as electronic health records and gold standard genetic test results have compounded this issue. While hereditary hemochromatosis is one of the most common genetic conditions, its mosaic phenotype and unclear penetrance makes it difficult to identify. Phenotypically there was not significant variation between those who tested positive for HFE-HH and those who did not in our dataset. The entire cohort of tested individuals had similar number of appointments, ICD-9 codes and diagnoses. The HFE-positive subset was generally more likely to be white, had more individuals with abnormal transferrin level and were taller on average.

5.1 Transferrin Saturation

Unfortunately only abnormal transferrin level is easily viable for additional screening criteria. While previous studies and our findings clearly support that C282Y is predominantly found in Caucasians, self-reported race in the EHR may not reflect underlying ancestry. Self-reported race has been identified as a particularly error-prone data element in EHR, as it is not always self-reported but is instead assigned. As we move forward into the era of precision medicine and whole genome sequencing, it may be possible to use population genetics and principal component analysis to inform the race listed in the EHR at which point it would be worthy to note that European ancestry greatly increases the likelihood of C282Y homozygosity.

While abnormal transferrin saturation was observed in the majority of HFE-HH cases, five patients did not have abnormal transferrin saturation (and five did not have TIBC laboratory results in the dataset). It is unclear whether any of these cases had previous laboratory results outside the UWMC system or other explanations. It is also possible that the transferrin saturation was lower due to any number of external factors (such as injury with bleeding, accumulation of iron in tissue and therefore reduced iron levels in serum, or reduced dietary iron).

5.2 Testing and Clinical Decision Support

In evidence-based medicine it always challenging to draw the line for sensitivity and specificity but as long as the screening criteria does not overrule clinical judgement and there is a process for allowing exception the University of Washington would benefit from adopting high transferrin saturation as a metric for genetic testing. This would reduce the number of tests and increase sensitivity. Adoption of this type of screening criteria is in line with what other major research universities with specialty clinics such as Mayo have instituted.

Additionally, recommended testing guidelines appear to apply to a large number of untested individuals in the UWMC system. TIBC tests are generally only given in cases of abnormal liver enzymes or iron-related symptoms. So unless these cases exclusively fall into patients with confirmed hepatitis infection or alcohol-related liver disease, it is likely that a fair number of patients in this untested cohort would benefit from genetic testing. The logistic regression model to fit the probability of testing could be used as an automated first pass to identify patients who look more like the tested cohort. From this narrowed list, the association rules can further identify promising candidates for chart review. 179 individuals were identified in this study, including four who met eight or nine of the rules. Genetic testing in this identified cohort could provide validation on whether having additional measures to identify HFE-HH patients is beneficial.

The same association rules could be used to inform a decision support system. If abnormal TIBC results come up, then it could check for ICD-10 codes related to positive hepatitis diagnosis, alcohol-related liver disease and other secondary iron overload. If patients have abnormal TIBC, no clear secondary iron overload, and meet one or more of the association rules, then it might be worth letting physicians know to consider HFE-HH testing.

The data also supports the findings of previous literature that ferritin levels are less indicative of HFE-HH than transferrin saturation and additionally that HFE-HH cases do not have hyper-ferritin levels (greater than 5,000 ng/mL). Laboratory results in this range are more likely due to acute infection than gradual iron-accumulation (Sackett, Cunderlik, Sahni, Killeen, & Olson, 2016). These kind of findings could be used to inform clinical decision support systems or laboratory medicine quality assurance measures.

5.3 Non-HFE Hemochromatosis

While this genetic test is considered the gold standard for HFE-HH, it is difficult to rule out HH all together as there are other variants in known iron absorption pathways that lead to the same phenotypic condition but with different molecular causes. Interestingly the tested population included a substantial portion of individuals who self-identified as Asian (91 individuals). A recent large scale study of iron overload found high levels of iron overload in Asian populations despite having a C282Y prevalence of 3.9 in 10 million (compared to 6-10 per 1,000 in Caucasians) (Paul C. Adams et al., 2005).

A recent unpublished study found 80% of Asians with iron overload had a nonsynonymous mutation in a different iron absorption gene that is potentially causal (Zhang, W., Lv, T., Xu, A., You, H., Jia, J., Ou, X., & Huang, J., 2017). While HFE-HH is the most common type, there are four other genes in iron absorption pathways with variants known to cause

iron overload (Lok et al., 2009). It is possible that many of the individuals in the dataset who tested negative for C282Y mutations HFE-HH in fact have other variants responsible and because multiple genes can impact the same pathway it would be phenotypically impossible to distinguish between them making true positive and negatives difficult to evaluate.

UWMC's HEMDNA test is part of a panel of three genes that are sequenced together. They use custom targeted capture probes to sequence only the areas of the gene around the known clinical variant (UWMC Laboratory Medicine Internal Protocol, 2016). The laboratory results extracted from the EHR contain interpretation on the raw data but not all the findings. It is not possible to identify new potentially pathogenic variants in HFE or in other iron pathway genes from this type of sequencing. With advances in next-generation sequencing technology it could be possible to create custom panels that include HFE, HJV, HAMP, TFR2 and SLC40A1 to identify known pathogenic variants as well as look for novel disease-causing mutations.

5.4 Logistic Regression

Logistic regression was chosen for modeling because the outcome was binary. Logistic regression models are more difficult to interpret than linear models and do not provide insight that can be easily translated to clinical support systems or guidelines. While logistic regression using abnormal transferrin saturation was the best fit for identifying HFE-cases among those tested, it had lower sensitivity and specificity than applying a clear threshold. However, using logistic regression to identify patients who are similar to the testing cohort could prove to be a valuable automated way to screen more potential cases. Early detection is key to improving outcomes and because HFE-HH increases the risk of morbidity when coupled with diabetes or alcoholism, wider-detection mechanisms are crucial even as population screening is not recommended.

5.5 Association Rule Mining

Association rule mining provided additional interpretability to features and allowed for soft validation against medical knowledge. Association rules have the advantage of combining a collection of terms that are not necessarily linear but instead collectively correlated. While individual variables that make up the rules have been previously reported characteristics of HFE-HH patients, their combination is the canonical HFE-HH phenotype established before genetic testing. The association rule mining from a proof of concept standpoint has unequivocally shown that this technique can be applied to EHR data to retrieve medically relevant sets. Without further validation and analysis it is unclear if the features not listed in previous literation (such as number of ICD-9 codes) and features counter to existing knowledge (such as low BMI), represent novel insights or viable patient subsets, or if they merely fit this specific dataset.

A hybrid approach using logistic regression and association rules can provide insight when there is an issue of unlabeled positives in a retrospective study. It allows for the untested cohort to be subsetted and applies the association rules to only those individuals who look the most like the tested cohort (where the association rules were mined). Validation of this approach in a wider dataset may help provide better clinical guidelines than the overly broad abnormal TIBC laboratory results while increasing the rate of early detection.

5.6 Sensitivity and Specificity

Sensitivity and specificity for HFE-HH screening with serum ferritin and transferrin saturation have not been published in previous studies likely due to the difficulty in confirming HH phenotype in C282Y homozygotes. The sensitivity and specificities listed in this paper assume that clinicians have successfully screened for HH phenotype and are confirming the genetic component. This means it is not possible to easily compare the values found in this

study to previous literature. The best comparison is between the previous positive predictive value (PPV) of clinician genetic testing recommendation alone (7.5%) to the proposed combination of clinician screening + transferrin saturation (12.5%). The goal of this research was not to establish these values but future research may want to re-evaluate this question in terms of diagnostic screening for HFE-HH as most papers refer to a study conducted in 1984 prior to HFE-genotype testing (Borwein, Ghent, & Valberg, 1984)

5.7 Limitations

Sample size was the primary limitation of this study. The ratio of cases to "controls" made it infeasible to use traditional machine learning methods that can handle missing and noisy data more appropriately. It also prevented direct validation of findings from this research by subsetting the data into training and test portions. It is unclear whether the models developed through this study are valid in other datasets or generalizable across systems but this is certainly an area for future work.

Chapter 6

FUTURE WORK

The results of this study require validation. This could be achieved by seeking additional electronic health record data for HFE-HH confirmed cases from other medical systems to use as a test set. Within the UW medical system, it may be prudent to work with physicians in the department of Laboratory Medicine to review charts of the 179 patients identified as candidates for testing to further analyze whether they are good candidates and perform HFE-HH testing on those that look promising.

The confirmation of previous literature guidelines for genetic testing of this condition could lead to additional quality assurance work in the UWMC laboratory medicine department in terms of analyzing the referring physician and identifying physicians who may be referring patients with low likelihood of hemochromatosis. The results of this research coupled with genetic testing algorithms in place at other institutions could be used to develop internal guidelines for HFE-HH testing or clinical decision support systems. It may be feasible to pilot test an alert system that would remind physicians that high transferrin saturation is key to having a diagnosis of HFE-HH.

The eMERGE project has one of the largest collections of exome data coupled with electronic health record. While their past examination only found 20% of C282Y homozygous individuals had an ICD-9 diagnosis of HFE-HH, 33.5% shared the comorbidity of arthritis which is known to be associated with hereditary hemochromatosis. Out of their 98 C282Y homozygous patients, only 18 had transferrin saturation testing. Status as C282Y homozygous is not currently reported back to physicians nor patients in the eMERGE project but it

is arguable that many of these patients would benefit from transferrin saturation screening and follow-up. This could be an excellent validation opportunity to look at the hybrid logistic regression and association rules to identify other potential candidates in their system and because they have the genotypes already it may be enough to tip the scale on return of results for this variant.

This study strictly used structured EHR data. Concurrent work with UWMC Genetic Medicine Clinic highlighted the breadth of patient information contained in pdfs and supplemental information not easily extracted from the EHR without prior knowledge and natural language processing. There is still a lot more data available in the EHR as a whole and natural language processing could extract pertinent information from physician notes and family histories. This could be particularly valuable when considering European ancestry as a viable screening criteria for HFE-HH.

The similarity between HFE-tested individuals has highlighted the need for an evaluation of wider sequencing panels that includes all iron transport genes with tools for interpretation and novel mutation discovery. Particularly for the cohort of 91 Asian patients with iron overload, recent literature suggests that there could be a missense mutation responsible and this would be an excellent cohort to use to validate that finding.

Chapter 7

CONCLUSION

Electronic health record data is notoriously difficult to work with and requires substantial medical knowledge, patience and expertise to curate a viable dataset. Viable signals and key features for even difficult and mosaic conditions such as HFE-HH can be extracted from the data. Predictive modeling of common genetic conditions using EHR data is feasible provided there is well documented phenotype information and a validated confirmation set. It is difficult to distinguish HFE-HH from other genetic causes of hemochromatosis but as whole genome sequencing increases in prevalence and techniques for using EHR data improve, it may be possible to expand the definitions of hereditary hemochromatosis and develop a comprehensive clinical gene panel that includes all relevant variants. In the meantime, developing clinical testing guidelines that fit our current genetic testing system would reduce unnecessary testing among referrals and improve identification of candidates for treatment from individuals with abnormal TIBC values.

Chapter 8

REFERENCES

Adams, P. C. (2014). H63D Genotying for Hemochromatosis: Helper or Hindrance? Canadian Journal of Gastroenterology and Hepatology, 28(4), 179–180.

Adams, P. C. (2015). Epidemiology and diagnostic testing for hemochromatosis and iron overload. International Journal of Laboratory Hematology, 37 Suppl 1, 25–30.

Adams, P. C., Reboussin, D. M., Barton, J. C., McLaren, C. E., Eckfeldt, J. H., McLaren, G. D., ... Sholinsky, P. (2005). Hemochromatosis and Iron-Overload Screening in a Racially Diverse Population. The New England Journal of Medicine, 352(17), 1769–1778.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. https://doi.org/10.1145/170035.170072

Alexander, J., & Kowdley, K. V. (2009). HFE-associated hereditary hemochromatosis. Genetics in Medicine: Official Journal of the American College of Medical Genetics, 11(5), 307–313.

Allen, K. J., Gurrin, L. C., Constantine, C. C., Osborne, N. J., Delatycki, M. B., Nicoll, A. J., ... Gertig, D. M. (2008). Iron-Overload–Related Disease inHFEHereditary Hemochromatosis. The New England Journal of Medicine, 358(3), 221–230.

Balistreri, C. R., Candore, G., Almasio, P., Di Marco, V., Colonna-Romano, G., Craxi, A., ... Caruso, C. (2002). Analysis of hemochromatosis gene mutations in the Sicilian population: implications for survival and longevity. Archives of Gerontology and Geriatrics. Supplement, 8, 35–42.

Barton, J. C., & Acton, R. T. (2017). Diabetes in Hemochromatosis. Journal of Diabetes Research, 2017, 9826930.

Birkhead, G. S., Klompas, M., & Shah, N. R. (2015). Uses of electronic health records for public health surveillance to advance public health. Annual Review of Public Health, 36, 345–359.

Blumenthal, D., & Tavenner, M. (2010). The "meaningful use" regulation for electronic health records. The New England Journal of Medicine, 363(6), 501–504.

Borwein, S., Ghent, C. N., & Valberg, L. S. (1984). Diagnostic efficacy of screening tests for hereditary hemochromatosis. Canadian Medical Association Journal, 131(8), 895–901.

Bulaj, Z. J., Griffen, L. M., Jorde, L. B., Edwards, C. Q., & Kushner, J. P. (1996). Clinical and Biochemical Abnormalities in People Heterozygous for Hemochromatosis. The New England Journal of Medicine, 335(24), 1799–1805.

Burke, W., Thomson, E., Khoury, M. J., McDonnell, S. M., Press, N., Adams, P. C., ... Collins, F. S. (1998). Hereditary hemochromatosis: gene discovery and its implications for population-based screening. JAMA: The Journal of the American Medical Association, 280(2), 172–178.

Chassin, M. R., Loeb, J. M., Schmaltz, S. P., & Wachter, R. M. (2010). Accountability

measures–using measurement to promote quality improvement. The New England Journal of Medicine, 363(7), 683–688.

Cippà, P. E., & Krayenbuehl, P.-A. (2013). Increased height in HFE hemochromatosis. The New England Journal of Medicine, 369(8), 785–786.

Dekel, O., & Shamir, O. (2008). Learning to classify with missing and corrupted features. In Proceedings of the 25th international conference on Machine learning - ICML '08. https://doi.org/10.1145/1390156.1390184

Denny, J. C. (2012). Chapter 13: Mining electronic health records in the genomics era. PLoS Computational Biology, 8(12), e1002823.

Ellervik, C., Mandrup-Poulsen, T., Tybjærg-Hansen, A., & Nordestgaard, B. G. (2014). Total and cause-specific mortality by elevated transferrin saturation and hemochromatosis genotype in individuals with diabetes: two general population studies. Diabetes Care, 37(2), 444–452.

European Association For The Study Of The Liver. (2010). EASL clinical practice guidelines for HFE hemochromatosis. Journal of Hepatology, 53(1), 3–22.

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. Biometrika, 80(1), 27.

Fleming, R. E., & Ponka, P. (2012). Iron overload in human disease. The New England Journal of Medicine, 366(4), 348–359.

Fletcher, L. M., Dixon, J. L., Purdie, D. M., Powell, L. W., & Crawford, D. H. G. (2002). Excess alcohol greatly increases the prevalence of cirrhosis in hereditary hemochromatosis.

Gastroenterology, 122(2), 281–289.

Gallego, C. J., Burt, A., Sundaresan, A. S., Ye, Z., Shaw, C., Crosslin, D. R., ... Jarvik, G. P. (2015). Penetrance of Hemochromatosis in HFE Genotypes Resulting in p.Cys282Tyr and p.[Cys282Tyr];[His63Asp] in the eMERGE Network. American Journal of Human Genetics, 97(4), 512–520.

Grosse, S. D., Gurrin, L. C., Bertalli, N. A., & Allen, K. J. (2018). Clinical penetrance in hereditary hemochromatosis: estimates of the cumulative incidence of severe liver disease among HFE C282Y homozygotes. Genetics in Medicine: Official Journal of the American College of Medical Genetics, 20(4), 383–389.

Hanson, E. H., Imperatore, G., & Burke, W. (2001). HFE gene and hereditary hemochromatosis: a HuGE review. Human Genome Epidemiology. American Journal of Epidemiology, 154(3), 193–206.

Heinze, G., Ploner, M., & Beyea, J. (2013). Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions. Statistics in Medicine, 32(29), 5062–5076.

Hollerer, I., Bachmann, A., & Muckenthaler, M. U. (2017). Pathophysiological consequences and benefits of mutations: 20 years of research. Haematologica, 102(5), 809–817.

Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association: JAMIA, 20(1), 117–121.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their asser-

tions from discharge summaries. Journal of the American Medical Informatics Association: JAMIA, 18(5), 601–606.

Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., Pacheco, J. A., Thompson, W. K., Armstrong, L. L., ... Lowe, W. L. (2012). 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, 19(2), 212–218.

Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. Nature Reviews. Genetics, 12(6), 417–428.

Koperdanova, M., & Cullis, J. O. (2015). Interpreting raised serum ferritin levels. BMJ , 351, h3692.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159–190.

Li, D., Simon, G., Chute, C. G., & Pathak, J. (2013). Using association rule mining for phenotype extraction from electronic health records. AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, 2013, 142–146.

Lim, A., Speechley, M., & Adams, P. C. (2014). Predicting C282Y homozygote genotype for hemochromatosis using serum ferritin and transferrin saturation values from 44,809 participants of the HEIRS study. Canadian Journal of Gastroenterology & Hepatology, 28(9), 502–504.

Lok, C. Y., Merryweather-Clarke, A. T., Viprakasit, V., Chinthammitr, Y., Srichairatanakool, S., Limwongse, C., ... Robson, K. J. H. (2009). Iron overload in the Asian community.

Blood, 114(1), 20–25.

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS Medicine, 3(11), e442.

Merryweather-Clarke, A. T., Pointon, J. J., Shearman, J. D., & Robson, K. J. (1997). Global prevalence of putative haemochromatosis mutations. Journal of Medical Genetics, 34(4), 275–278.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6, 26094.

M.Kang'ethe, S., Kang'ethe, S. M., & Wagacha, P. W. (2014). Extracting Diagnosis Patterns in Electronic Medical Records using Association Rule Mining. International Journal of Computer Applications in Technology, 108(15), 19–26.

Motulsky, A. G., & Beutler, E. (2000). Population Screening in Hereditary Hemochromatosis. Annual Review of Public Health, 21(1), 65–79.

Niederau, C., Fischer, R., Sonnenberg, A., Stremmel, W., Trampisch, H. J., & Strohmeyer, G. (1985). Survival and causes of death in cirrhotic and in noncirrhotic patients with primary hemochromatosis. The New England Journal of Medicine, 313(20), 1256–1262.

Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society, 10(2), 334–343.

O'Toole, R., Romeril, K., & Bromhead, C. (2017). Using iron studies to predict HFE mutations in New Zealand: implications for laboratory testing. Internal Medicine Journal, 47(4), 447–454.

Pardo, A. (2000). Cost-effectiveness of genetic diagnosis for Hereditary Hemochromatosis screening. Gastroenterology, 118(4), A1405.

Pietrangelo, A. (2004). Hereditary Hemochromatosis — A New Look at an Old Disease. The New England Journal of Medicine, 350(23), 2383–2397
Porto, G., Brissot, P., Swinkels, D. W., Zoller, H., Kamarainen, O., Patton, S., ... Keeney, S. (2016). EMQN best practice guidelines for the molecular genetic diagnosis of hereditary hemochromatosis (HH). European Journal of Human Genetics: EJHG, 24(4), 479–495.

Powell, L. W., Dixon, J. L., Ramm, G. A., Purdie, D. M., Lincoln, D. J., Anderson, G. J., ... Bassett, M. L. (2006). Screening for hemochromatosis in asymptomatic subjects with or without a family history. Archives of Internal Medicine, 166(3), 294–301.

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., ... on Behalf of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) QRS Group. (2013). Genome- and Phenome-Wide Analyses of Cardiac Conduction Identifies Markers of Arrhythmia Risk. Circulation, 127(13), 1377–1385.

Sackett, K., Cunderlik, M., Sahni, N., Killeen, A. A., & Olson, A. P. J. (2016). Extreme Hyperferritinemia: Causes and Impact on Diagnostic Reasoning. American Journal of Clinical Pathology, 145(5), 646–650.

Sánchez, M., Villa, M., Ingelmo, M., Sanz, C., Bruguera, M., Ascaso, C., & Oliva, R. (2003). Population screening for hemochromatosis: a study in 5370 Spanish blood donors. Journal

of Hepatology, 38(6), 745–750.

Schlimmer, J. C., & Granger, R. H. (1986). Incremental learning from noisy data. Machine Learning, 1(3), 317–354.

Shekelle, P. G., Morton, S. C., & Keeler, E. B. (2006). Costs and Benefits of Health Information Technology. https://doi.org/10.23970/ahrqepcerta132

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association: JAMIA, 21(2), 221–230.

Vanclooster, A., Cassiman, D., Van Steenbergen, W., Swinkels, D. W., Janssen, M. C. H., Drenth, J. P. H., ... Wollersheim, H. (2015). The quality of hereditary haemochromatosis guidelines: a comparative analysis. Clinics and Research in Hepatology and Gastroenterology, 39(2), 205–214.

Yu, S., Liao, K. P., Shaw, S. Y., Gainer, V. S., Churchill, S. E., Szolovits, P., ... Cai, T. (2015). Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. Journal of the American Medical Informatics Association: JAMIA, 22(5), 993–1000.

Zhang, W., Lv, T., Xu, A., You, H., Jia, J., Ou, X., & Huang, J. (2017). Non-HFE variation in Chinese patients with primary iron overload: recurrent HJV variants in the signal peptide region. Hepatology , 66, 431–432.