USE OF PHENOME-WIDE AND GENOME-WIDE APPROACHES TO IDENTIFY PATTERNS OF DISEASE

By

Jamie Rene Robinson

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

June 30, 2020

Nashville, Tennessee

Approved:
Joshua Denny, M.D., M.S.
Robert Carroll, Ph.D.
Gretchen Purcell Jackson, M.D., Ph.D.
Dan Roden, M.D.
Naji Abumrad, M.D.

ACKNOWLEDGMENTS

This work would not be possible without the financial support of the National Library of Medicine Fellowship in Biomedical Informatics through the training grant 5T15LM007450-12.

I am grateful for the mentorship and support of my Thesis Committee, including Drs. Joshua Denny, Robert Carroll, Gretchen Jackson, Dan Roden, and Naji Abumrad who have guided and contributed invaluable insights into this work. I would especially like to thank the chair of my committee Dr. Joshua Denny for his guidance in research direction and producing the quality manuscripts within this thesis.

I am also indebted to the participants of Vanderbilt’s BioVU, the eMERGE network, and the UK Biobank, without which the data to perform this research would have not been available. I am also grateful to everyone in the Departments of Biomedical Informatics and General Surgery who have helped me through my fellowship and completion of this research throughout residency.

Finally, none of this would have been possible without the support of my husband, Sina Iranmanesh, as I pursued this goal as we started and have continued to grow our family.

TABLE OF CONTENTS

LIST OF TABLES

Chapter II

Table Page

Chapter IV

Table Page

Chapter V

Table Page

Chapter VI

Table Page

Supplemental Table

LIST OF FIGURES

Chapter II

Chapter III

Chapter IV

Chapter V

Figure Page

Chapter VI

Figure Page

Supplemental Figure

CHAPTER I

Introduction

Research Motivation

The rise in available longitudinal patient information in electronic health records (EHRs) and their coupling to DNA biobanks has resulted in a dramatic increase in genomic research using EHR data for phenotypic information. Simultaneously, high-throughput methods for genotyping have considerably decreased the cost of genetic discovery while also improving in accuracy and availability. The benefit of leveraging EHRs for genomic research as opposed to prospective cohort-based studies is the ability to obtain large sample sizes with relatively less time or expense. This approach has allowed for the accumulation of unprecedented cohorts of genotyped individuals with well-defined phenotypes to drive genomic discovery.

Genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS) have provided powerful methods for investigating the impact of genetic variation on phenotypes. GWAS and PheWAS are modern genetic tools for the exploration of datasets to efficiently identify genomic risk factors for disease. PheWAS is a reverse genetics approach that provides a systematic methodology for the analysis of many phenotypes, often derived from EHR data, against a specific independent variable, such as a genotype.(1) PheWAS has shown the feasibility of analyzing genomic associations with thousands of phenotypes across a cohort of individuals and finding novel associations (2–5). This approach can also be applied using various predictive attributes in the PheWAS analysis, including genetic risk scores, a set of SNPs aggregated into a single continuous score.

Specific Aims

This thesis describes four research projects, presented in five manuscripts, that address important gaps in the scientific evidence on the use of genotyping and phenotyping to characterize clinical diseases.

Specific Aim #1: Perform a Comprehensive Review of Approaches to High-throughput Methods of Phenotyping and Genotyping

The first aim of this thesis was to determine the current state of the evidence about the use of phenotyping and genotyping to drive research. This aim comprises two manuscripts which detailed a comprehensive review examining the current knowledge on using clinical data within EHR to derive phenotypes that can further genomic research and drug discovery. To utilize genomic data to drive discovery and improvements in clinical care, accurate and efficient phenotyping methods must be utilized. These phenotypes include specific diseases or observable traits and are used to decipher the genetic determinants of human diseases, physiologic attributes, and medication response.

We highlighted in these manuscripts the recent advances in phenotyping methods, biobanking, and drug development and repurposing accelerated by applying genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS) to longitudinal health data information, along with limitations of these methods. GWAS and PheWAS do not only provide insight into biology of diseases, but also provide opportunities for drug targeting, development, and identification of populations at risk for drug-related adverse events. Knowledge of the genetic mechanisms that drive phenotypic and drug response variation can help guide diagnosis and the tailoring of medication therapy. In addition to summarizing the current data, in this aim we focus on opportunities for future applications of phenotyping that can provide linkages between disease-gene associations and therapeutic approaches.

Specific Aim #2: Characterize a Rare Clinical Disease, Loxoscelism, using Electronic Health Records and Phenotyping Methods

The second research project described the use of the EHR to characterize a rare disease process, loxoscelism, that can arise following a bite from a brown recluse spider. Systemic loxoscelism in its mild form consists of nausea, vomiting, fever, chills, or arthralgia. In its more severe form, brown recluse bites may cause massive hemolysis, hemoglobinuria, acute renal failure, disseminated intravascular coagulation, and rarely death.(6–10) In this study, we described clinical characteristics and outcomes of the largest known cohort of individuals with systemic loxoscelism to date, leveraging our large de-identified electronic clinical data warehouse. We then performed a phenome-wide association study (PheWAS) of these individuals matched to a control population to identify key differences in ~1800 phenotypes between individuals who develop systemic loxoscelism and those who do not. We aimed to demonstrate how high-throughput phenotyping methods can provide insight into diseases. In doing so, we highlighted clinical characteristics of this rare and potentially lethal illness and uncovered previously undocumented phenotypic associations.

Specific Aim #3: Evaluate the Association of a Common Disease, Obesity, and Obesity Genetic Risk with Postoperative Complications

The third research project used genotyping to derive polygenic risk scores coupled with phenotyping to evaluate for associations of obesity with postoperative complications. Obesity, defined as a body-mass index (BMI) of 30.0 kg/m^2 or greater, is known to be a strong predictor of cardiovascular morbidity and mortality. Over two-thirds of the adult population in the United States have an overweight or obese BMI.(11–14) However, the extent to which obesity and genetics determine post-operative complications is incompletely understood. We aimed to determine the influence obesity and genetic risk for obesity has on postoperative outcomes using high-throughput methods of both phenotyping and genotyping. We leveraged a large EHR population to identify specific postoperative complications, including postoperative infection, incisional hernia, and small bowel obstruction, associated with BMI. In a separate cohort, we then used a polygenic risk score for BMI to investigate the relationship between genetic risk for obesity and these postoperative complications.(15) We demonstrated that both clinical and genomic risk of obesity is associated with the development of postoperative incisional hernia and infection.

Specific Aim #4: Evaluate the Association of Obesity and Genome-wide Obesity Genetic Risk with Healthcare Disease Burden

Obesity is known to have a strong influence on the development of comorbidities and increased mortality risk. However, the extent of the role obesity has on comorbid conditions across the phenome is unknown. Further, while there are data showing obesity is a polygenic disease with a greater proportion of the variance in BMI explained with greater coverage of the genome in a polygenic risk score, it is unknown if genome-wide polygenic risk scores perform better in phenome-wide association studies. Therefore, we identified phenotypes associated with class 3 obesity in a clinical cohort and replicated these findings in two separate genetic cohorts using genome-wide polygenic risk scores for BMI, elucidating the complex genomic and phenomic characteristics of this prevalent disease. Class 3 obesity and polygenic risk for obesity was associated with 199 distinct phenotypes. The burden of disease associated with obesity was significant with a predicted 17.1% of disease in obese individuals potentially preventable if individuals maintained a normal BMI.

Research Synthesis

Through these specific aims, we are able to advance the knowledge on approaches to phenotyping methods to elucidate patterns of disease. Through the use of different phenotyping methods across diverse disease processes, we were able to illustrate the strengths and weaknesses of extracting data for research that was curated for clinical medicine. These methods allowed for curation of the largest dataset of individuals with systemic loxoscelism and characterization of this rare disease. This thesis is also the first application of genome-wide risk scores in a phenome-wide approach, demonstrating that genome-wide polygenic risk scores have improved ability to define disease risk and associations. This body of work reduced phenome-wide phenotyping uncertainties by grouping of billing codes, large cohort sizes, and the requirement of multiple instances of the billing codes on separate days, providing results that were validated across cohorts and with both clinical and genomic predictors. These novel methods allowed us to demonstrate the full extent of the role obesity has on postoperative complications and the overall burden of disease driven by obesity in society. Translation of these findings could involve applying genome-wide risk profiling methods to identification of individuals who would benefit from environmental modifications or heightened medical awareness prior to the onset of obesity and its comorbid conditions.

References

1. Denny JC, Bastarache L, Roden DM. 2016. Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* 17:353–73

2. Hebbring SJ. 2014. The challenges, advantages and future of phenome-wide association studies. *Immunology*. 141(2):157–65

3. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. 2013. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 14(3):187–91

4. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* 26(9):1205–10

5. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31(12):1102–10

6. Futrell JM. 1992. Loxoscelism. *Am. J. Med. Sci.* 304(4):261–67

7. Murray LM, Seger DL. 1994. Hemolytic anemia following a presumptive brown recluse spider bite. *J. Toxicol. Clin. Toxicol.* 32(4):451–56

8. Rosen JL, Dumitru JK, Langley EW, Meade Olivier CA. 2012. Emergency department death from systemic loxoscelism. *Ann. Emerg. Med.* 60(4):439–41

9. Nance WE. 1961. Hemolytic anemia of necrotic arachnidism. *Am. J. Med.* 31:801–7

10. Rees RS, Altenbern DP, Lynch JB, King LE. 1985. Brown recluse spider bites. A comparison of early surgical excision versus dapsone and delayed surgical excision. *Ann. Surg.* 202(5):659–63

11. Global BMI Mortality Collaboration null, Di Angelantonio E, Bhupathiraju S, Wormser D, Gao P, et al. 2016. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet Lond. Engl.* 388(10046):776–86

12. Prospective Studies Collaboration, Whitlock G, Lewington S, Sherliker P, Clarke R, et al. 2009. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet Lond. Engl.* 373(9669):1083–96

13. Emerging Risk Factors Collaboration, Wormser D, Kaptoge S, Di Angelantonio E, Wood AM, et al. 2011. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *Lancet Lond. Engl.* 377(9771):1085–95

14. Ogden CL, Carroll MD, Kit BK, Flegal KM. 2014. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA*. 311(8):806–14

15. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 518(7538):197–206

CHAPTER II

Defining Phenotypes from Clinical Data to Drive Genomic Research

Jamie R. Robinson, M.D., M.S.[1, 2], Wei-Qi Wei, M.D., Ph.D.[1], Dan M. Roden, M.D.[1,3,4], Joshua C. Denny, M.D., M.S.[1, 3]

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

[2] Department of General Surgery, Vanderbilt University Medical Center, Nashville, TN

[3] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

[4] Department of Pharmacology, Vanderbilt University Medical Center

This manuscript was published in the journal Annual Reviews of Biomedical Data Science as follows: Robinson JR, Wei W, Roden DM, Denny JC. Defining phenotypes from clinical data to drive genomic research. *Annual Review of Biomedical Data Science*. 2018; 1:69-92.

Abstract

The rise in available longitudinal patient information in electronic health records (EHRs) and their coupling to DNA biobanks has resulted in a dramatic increase in genomic research using EHR data for phenotypic information. EHRs have the benefit of providing a deep and broad data source of health-related phenotypes, including drug response traits, expanding the phenome available to researchers for discovery. The earliest efforts at repurposing EHR data for research involved manual chart review of limited numbers of patients but now typically involve applications of rule-based and machine learning algorithms operating on sometimes huge corpora for both genome-wide and phenome-wide approaches. We highlight here the current methods, impact, challenges, and opportunities for repurposing clinical data to define patient phenotypes for genomics discovery. Use of EHR data has proven a powerful method for elucidation of genomic influences on diseases, traits, and drug-response phenotypes and will continue to have increasing applications in large cohort studies.

Introduction

The widespread adoption of electronic health records (EHRs) has raised the possibility of using these data in clinical research. Abundant evidence now supports the idea that the EHR repurposed for research represents a rich data set of a patient's health trajectory, including diseases, laboratory and radiology tests, and medications and their response, much of which can be hard to acquire in a research setting. Simultaneously, high-throughput methods for genotyping have considerably decreased the cost of genetic discovery while also improving in accuracy and availability. A key component of these are genome-wide association studies (GWAS) and whole genome and exome sequencing technologies, which systematically analyze variation across the genome. Since 2005, over 3000 GWAS have identified almost 40,000 unique SNP-trait associations (1).

The vast majority of early genomic research studies before 2010 were performed using observational cohorts or randomized controlled trial data. Perhaps in part as a result, some of the largest GWAS published have been traits that are common to many studies, such as height and body mass index. Nearly coinciding with the growth of genetic investigation has been the national adoption of EHRs across the United States. The national adoption rate among non-federal acute care hospitals was only 9.4% in 2008 but reached 94% by 2013 (2). The formation of EHR-linked DNA biobanks that repurpose EHR data from their clinical data stores to research-oriented databases, combined with advances in informatics tools and terminologies, led to the beginning of successful EHR utilization for genetic studies in 2010 (3–7) The first EHR-based studies recapitulated non-EHR GWAS by focusing on well-defined clinical phenotypes to test against variation in the human genome but later gave rise to other innovative, reverse-genetics approaches such as phenome-wide association studies (PheWAS), which provide a systematic approach to the analysis of many phenotypes potentially associated with a specific genotype (8).

Critical to discovery through genomic and phenomic investigation is the accumulation of sufficiently large sample sizes with well-defined phenotypes. These phenotypes include specific diseases or observable traits and are used to decipher the genetic determinants of human diseases, physiologic attributes, and medication response. In this review, we highlight the growth of and approaches to phenotyping using clinical data within EHRs as it pertains to genomics research, limitations of these techniques, and future opportunities for integration of genotypic and phenotypic clinical data.

Use of Clinical Data in Research

Cohorts Available for Genetic Study

Traditional genetic studies have used population- or clinical trial-based cohorts with prospective participant enrollment and questionnaires to gather data on specific aspects. While this approach can result in high-quality phenotypes, there are significant challenges due to the time and monetary expense required for participant accrual, retention, questionnaire completion, and validation (9, 10). Patient accrual can take months to years. Further, long-term follow-up and patient retention can be tedious or unfeasible, with exclusion of populations of patients important to include, such as those too ill to participate or at the extremes of age. Prospective cohorts are also generally guided by a particular clinical research question, and thus phenotype data may be limited to those conditions, hindering reuse of these data for future studies examining different phenotypes.

Accordingly, EHRs have emerged as an efficient method for obtaining dense patient information for research over the last two decades. Historically, clinical data and documentation has been collected primarily to support patient care and administrative functions, such as billing. Thus, cross-sectional aggregation and querying of EHR data was not a priority. In 2003, the National Academy of Medicine (then the Institute of Medicine) released a report on the key capabilities of an EHR, noting that facilitating research is an important secondary use of EHRs (11). Early studies utilizing aggregated EHR data often consisted of epidemiological studies performed at early EHR adoption sites with well-maintained databases such as the Veterans Health Administration or UK general practice research database (12–14). Recent efforts bringing together diverse international healthcare data have been able to study treatment protocols in as many as 250 million individuals (15). Some examples of network initiatives utilizing EHR data for genomic research are detailed below.

Current Biobanking Efforts

Biobanking of genetic data linked to the longitudinal patient data available within the EHR aggregates otherwise disparate information, potentially making it available to both clinicians and researchers. Some of the earliest biobanks linked to EHR data derived their samples from left over blood collected as part of clinical care, and would have otherwise been discarded (9). These include the Harvard Crimson, which pursued as-needed sample collection for phenotypes of interest, and Vanderbilt University Medical Center's BioVU, which started as a prospective collection of all individuals who did not opt-out of DNA collection as part of their routine consent to treatment (18–25)(16, 17). Since 2015, BioVU has converted to an opt-in consent model (electronically requested at the point of care) due to changes in the National Institutes of Health (NIH) Genomic Data Sharing policy requiring subject consent for data sharing (18).

Several other initiatives have been launched across the world for the development of very large EHR-linked biobanks which have started to deliver biomedical data sets comprising extensive phenotype and genotype information on hundreds of thousands of subjects. In the United States, the Electronic Medical Records and Genomics Network (eMERGE) and the Million Veteran Program (MVP) represent examples (19, 20). The MVP has recruited more than 580,000 participants and is establishing a longitudinal study of Veterans for future genomic and clinical research that combines data from survey instruments, the EHR, and biospecimens (20).

One of the more robust EHR phenotyping efforts has been performed by the eMERGE consortium, a national network organized and funded by the National Human Genome Research Institute with the goal of combining DNA biorepositories with EHRs for high-throughput and generalizable genomic discovery. An integral aim of eMERGE is to support the creation, validation, and dissemination of phenotype algorithms by providing tools that guide the user through the stages of development, public sharing, and reuse (21).

In Europe, the UK Biobank is a large prospective study of more than 500,000 individuals to investigate the role of genetic factors, environmental exposures, and lifestyle in the causes of major diseases (22). Participants aged 40-69 years were enrolled over 4 years in 22 recruitment centers, each completing questionnaires and donating biospecimens. The emphasis is now on further phenotyping of participants and ascertaining their health outcomes through follow-up and linkages to healthcare-based datasets (23). In Asia, the China Kadoorie Biobank also has more than 500,000 individuals and has prospectively linked genomic information to both EHR data and participant surveys.

The *All of Us* Research Program[1] is a NIH-funded initiative to build a United States research cohort of more than one million individuals, including prospective participant provided information, molecular data (including genomics), and linkage to health information in EHRs (24). Participants will be recruited from diverse healthcare centers located across the country and as “direct volunteers”, individuals who may not have a direct connection with a recruiting healthcare system.

Repurposing of EHR Data

Rationale for use of Clinical Data for Genomics Research

EHRs offer longitudinal patient information in a form that is relatively unbiased to particular diseases or research agendas, allowing for study of diverse genomic risk, diseases, and outcomes. The rich phenotypic clinical documentation coupled with laboratory data, medication receipt, family history,

[1] Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins With You are service marks of the U.S. Department of Health and Human Services.

and environmental exposures, makes the EHR a practical data source for reuse in genomic studies. The key advantage of repurposing EHRs for research is that they are already created and maintained for healthcare delivery and prospectively accrue clinical observations and costly tests at regular intervals driven by an individual's health trajectory.

To quantify the density of information in the EHR, we explored the data available for BioVU participants (Figure 1). We found that individuals had an average of 7.8 years of clinical data, including many years before enrollment and up to 33 years of longitudinal recorded information. Much of these data are costly and may be infeasible to collect for a research trial. For example, consider the cost to obtain and clinically interpret the 3.4 million radiology tests (~14 radiology tests per patient) presented in Figure 1; these included over 628,000 computed tomography (CT) scans and nearly 192,000 magnetic resonance imaging (MRI) scans, all with clinical interpretations. Assuming a conservative estimate of $500 per test, these CT and MRI scans alone would cost $410 million. Thus, there is a significant reduction in research time and expense for accrual of large sample sizes with a breadth of clinical data (9, 25).

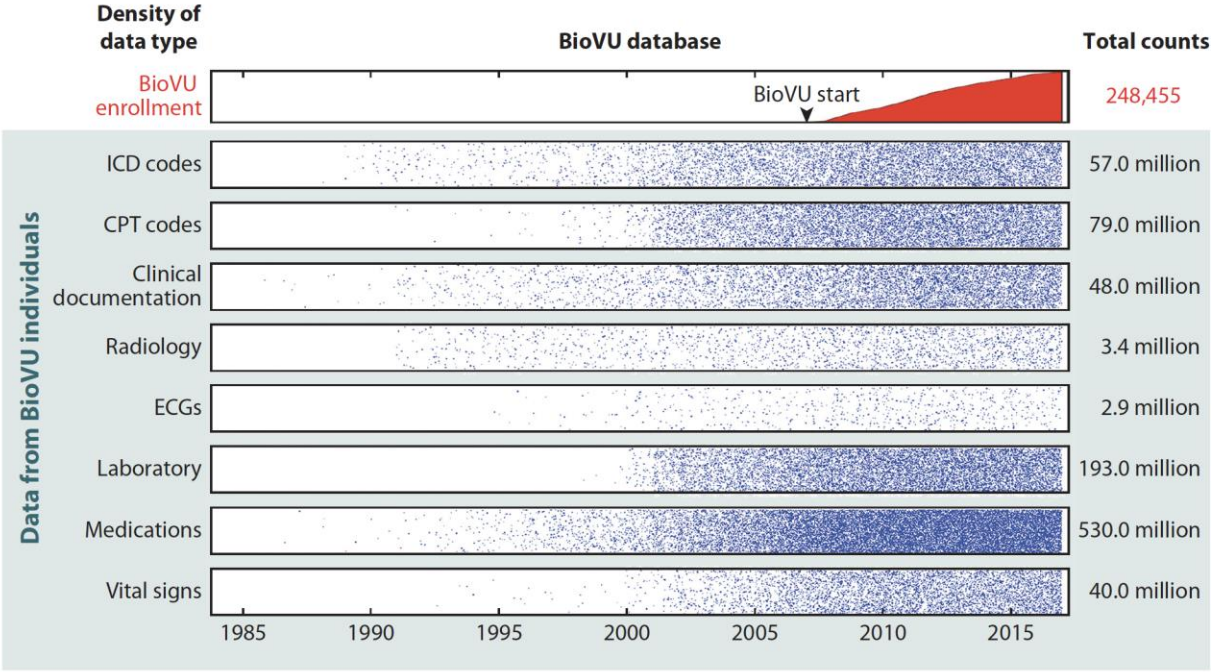


Figure 1. Density of data in the Vanderbilt EHR-linked biobank, BioVU. While enrollment in BioVU and accrual of samples for DNA analysis started in 2007, clinical data within the electronic health record on the individuals enrolled dated as far back as 1984. Data points were transformed by taking the square root and dividing by 20.

Another potential benefit of linking DNA repositories to EHRs is the inclusivity of EHRs in comparison to traditional population-based cohorts, which often will exclude certain diseases, children, minority or poor populations, and the elderly. This is critical for both identification of a range of cases as well as controls. Because EHR-based cohorts provide significant variability in phenotypic traits, a single cohort can be reused many times for many phenotypes or genetic variants examined (9, 26). Once the genetic data have been collected, the majority of cost and effort is thus expended at defining, refining, and validating phenotypes of interest.

Large sample sizes for modern genetic research methods, such as GWAS and PheWAS, are critical to the discovery of novel findings and afforded by the EHR (27). For both GWAS and PheWAS, there is a need to correct for multiple comparisons, increasing the threshold for statistically-significant results. For GWAS, the threshold has been established at 5x10^{-8} (28). While the threshold for statistical significance is less well-established for PheWAS, a Bonferroni correction is often applied in these analyses, resulting in a conservative significance level that assumes independence across all phenotypes, which is unlikely given that many phenotypes of human diseases and traits are closely related. Since statistical power is a function of number of tests performed, effect size (often low for many variants), and minor allele frequency, use of this stringent threshold has necessitated use of larger and larger cohorts to enable identification of significant associations, especially those that are rare or low-frequency variants of moderate-to-large effect.

Classes of Data Available in EHRs for Phenotype Curation

EHR phenotyping is the process of identifying individuals with an explicit observable trait from large quantities of imperfect clinical patient data (Figure 2) (29). The earliest approach to phenotyping is via manual chart review, typically performed by thorough searching of clinical documents, laboratory, and medication information by individuals with medical domain knowledge. For automation of phenotyping, EHR data, which is stored in both structured and unstructured formats, is extracted and utilized for analysis (Table 1). Structured data is typically easier for computerized extraction, with examples being billing codes, laboratory results, vital signs, and often medications. A large proportion of the EHR is relatively unstructured, including almost all clinical documentation, radiology reports, and some test or laboratory results. For many phenotypes, accurate case selection is best achieved through use of a combination of structured and unstructured data (10).

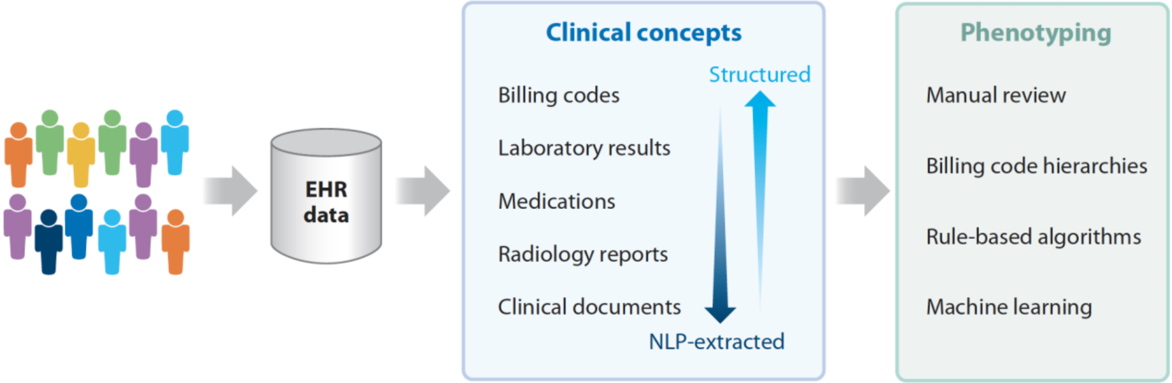


Figure 2. Methods of electronic health record (EHR)-based phenotyping. Clinical concepts are in both structured and unstructured (require machine-extraction) in the EHR. These concepts can be used for various forms of phenotyping.

Table 1. Common data classes within electronic health records (EHRs).

	Demographics	**Diagnoses**	**Procedures**	**Medication records**	**Laboratory results**	**Imaging**	**Clinical text documentation**
Data format	Structured	Structured	Structured	Partially structured	Mostly structured	Partially structured	Mostly unstructured
Data standard	None	ICD9/10	CPT	RxNorm	LOINC	DICOM for images	SNOMED-CT
Query method	Simple	Simple	Simple	Simple, text searching, NLP	Simple, text searching	Simple, text searching, NLP	NLP
Recall	Moderate	Moderate	Variable	Moderate	Moderate	Moderate	Moderate
Precision	Moderate	Low	High	Moderate	High	High	High
Affected by healthcare fragmentation	No	Low/moderate (chronic) to high (acute)	High	Moderate	High	High	Moderate
Strengths	Easy to query	Easy to query	Easy to query	High validity for inpatient setting	High validity	High validity	Most dense clinical information; can capture out-of-hospital history
Weaknesses	Variable recall and precision based on demographic	Susceptible to inaccuracies	Susceptible to missing data	Susceptible to missing data	Variable recall and precision based on test; susceptible to missing data	Susceptible to missing data; difficult to process raw images	Most difficult to process and interpret at scale
Used in Phenotyping	Most common	Most common	Common	Common	Somewhat common	Somewhat common	Less common

*CPT, Current Procedural Terminology; DICOM, Digital Imaging and Communications in Medicine; ICD-9/ICD-10, International Classification of Diseases, Ninth Revision/Tenth Revision; LOINC, Logical Observation Identifiers Names and Codes; SNOMED-CT; NLP, Natural Language Processing; Systematized Nomenclature of Medicine-Clinical Terms

Billing data are the most commonly used resource for identifying phenotypes in both clinical and genomic research (21, 30). This structured data typically consists of International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) codes. The ICD coding system classifies diseases, symptomatology, and procedures based on a hierarchical terminology structure maintained by the World Health Organization (WHO). CPT codes were created by the American Medical Association, and are used to bill for clinical services, such as an imaging study or surgical procedure. Both classes of billing data are ubiquitous and easily queried within EHRs, making them highly utilized as at least a portion of most phenotyping algorithms. A query of the data types used in phenotyping algorithms in the Phenotype Knowledgebase (www.pheKB.org) shows that 122 of 154 (73%) algorithms used ICD codes (Table 2), all of which used ICD-9 codes and 26 that used both ICD-9 and -10 codes. This demonstrates one challenge of using billing codes within longitudinal patient information as coding systems change over time, such as the migration from ICD-9 to ICD-10 coding in 2015 in the US, resulting in the need for mapping strategies to combine codes from different systems. Sensitivity and specificity of billing codes alone are variable across phenotypes, with one study showing a range of positive predictive values (PPVs) for ICD codes from 0.12 to 0.56 across ten diseases (29). ICD codes generally have low specificity but are highly sensitive for diseases, as a clinician may bill an ICD code for a diagnosis based upon clinical suspicion rather than confirmation of disease (31). CPT codes tend to have higher specificity, as procedural coding is quite accurate, but lower sensitivity in comparison to ICD codes due to procedures being performed at other institutions, demonstrating fragmentation of EHR data (32).

Other types of structured data within the EHR, such as laboratory results and medications, are also often used to identify phenotypes (Table 2). In particular, 53% of algorithms available on PheKB utilize medication data. Record of medication receipt can be in various forms in the EHR; however, inpatient computerized provider order entry systems and outpatient drug prescribing systems have increased the availability of drug exposures as structured data. Medication data in the absence of corroborating evidence has not been found to be especially useful, with area under the receiver operator characteristic curve (AUC) of 0.54 for 10 diseases (29). Importantly, the capture of medications is essential to provide exposure data for pharmacogenomic studies (31, 33). Challenges to analyzing laboratory data in particular include the repeated measures, naming conventions, and various specimen sources, resulting in difficult interpretation. While laboratory and medication data must be placed into appropriate clinical context with careful selection, they can improve phenotyping accuracy for many conditions.

Table 2. Data modalities used in phenotyping algorithms available on PheKB. Data as of 10/15/2017. Non-public algorithms include algorithms in development and those whose performance has not yet been validated.

	Public (n = 44)	**Non-public (n = 110)**	**Percent of total**
ICD-9 or -10 codes	39	73	73%
Medications	31	51	53%
CPT codes	23	44	44%
Natural language processing	28	36	42%
Laboratory test results	21	37	38%
Vital signs	5	14	12%

* ICD-9/ICD-10, International Classification of Diseases, Ninth Revision/Tenth Revision; NLP, Natural Language Processing; CPT, Current Procedural Terminology

The main source of unstructured data within the EHR is clinical documentation, consisting of the most accurate record of the providers' thoughts and richest information for phenotype algorithms. To be useful for electronic-based phenotyping, clinical documentation must be in a format that is computable. The majority of clinical notes consist of narrative text, lacking uniform format or structure, thus they must be processed with either basic keyword searching or modern tools such as natural language processing (NLP), discussed further in the next section.

Processing of EHR Text Data

Narrative clinical documentation includes a wealth of information about diagnoses, signs and symptoms, risk factors, treatments, family history, exposures, and clinical decision making, many of which are not well captured by structured information in most EHRs. Indeed, the gold standard in validating whether a patient has a diagnosis or a given trait generally involves a review of the clinical notes (30). NLP is a tool for producing computable representations from this narrative unstructured text (34). There are many approaches to NLP, ranging from rule, grammar, and machine learning (ML)-based approaches for producing comprehensive “understandings” of the text (so-called “general-purpose” NLP systems (35–40)) to focused applications applied to particular tasks (e.g., identifying medications and their features (41), or left ventricular ejection fraction (42)). General-purpose NLP systems often seek to parse unstructured text documents into phrases that can be mapped to concepts within controlled terminologies such as the Unified Medical Language System (UMLS), the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), or RxNorm (for medications) (37, 41, 43–46). Mapping the textual elements within clinical free-text documents to a semantic terminology provides a standardized method to represent the data for downstream computation. Other tasks that often improve the performance of NLP systems include identification of negation and qualifiers (47), semantic role labeling (48), word sense disambiguation (49), and temporal analysis (50).

Some important clinical features often have historically only been found in narrative data, such as family history and smoking status. Smoking status has been the focus of many dedicated rule-based, ML, and hybrid NLP systems with F-measures ranging from 84-97% (38, 51–53). NLP also has the capacity to identify family history information with variable accuracy in prior studies (54–57). Identifying these important features within clinical documents allows the features to be reused and add meaningful information to diverse research studies. The introduction of Meaningful Use Stages 2 and 3 with the requirement for structured data entry for smoking status and family history will make these data more available over time with NLP tools (58, 59). Further, functionality such as Duke University’s MeTree, which allows a patient to complete his or her own family history, are providing new avenues for increased entry of structured information in the EHR (60, 61).

Performance for concept extraction varies with recall and precision ranging from 70-90%, depending on the specific application of the NLP (36–38, 44, 62). For this reason, EHR-based phenotyping algorithms have typically combined NLP features with other data components. Specific phenotypes such as adverse events and diseases can be successfully extracted using NLP (63–68), with several studies reporting higher predictive capability for case identification through NLP (either alone or in combination with other methods) in comparison to use of billing codes alone (19, 32, 34, 64, 69, 70).

Researchers have used both general-purpose NLP systems and focused applications, but the latter have become more common, since most phenotype algorithms do not need to identify all concepts but instead require a high-precision text mining approach to identify a set of terms applicable to specific concepts. For example, nearly all Vanderbilt phenotypes involving medications have employed a general-purpose NLP tool for medication extraction (MedEx) (41), but we have used a mix of regular expressions, general-purpose NLP tools (66, 71), or ML-based approaches for specific phenotypes (72). While general-purpose NLP systems can be utilized across a range of phenotypes, they also generally have inferior recall and precision to purpose-built approaches for defined phenotypes. Further, while NLP methods have significantly improved in the ability to identify negation and sentence structure, these algorithms remain imperfect and secular shifts in documentation, such as the transition from use of dictation to typed templates to “point-and-click” note writers, can result in instability of algorithm performance.

Challenges of Repurposing EHR Data for Research

While EHRs contain a wealth of extractable information for phenotype classification, their interface and the data generated within them are used primarily for clinical care and reimbursement, typically with little consideration towards research impact. The secondary use of EHRs for clinical, genomic, and pharmacogenomics discovery can be challenged by variable accuracy, lack of standardization, irregular follow-up, incompleteness of patient records, and significant amounts of unstructured information. Inaccuracy within EHRs can result from clinical uncertainty, omissions, or billing errors. Omission can occur due to provider workload and perceptions on what is deemed to be clinically relevant to report at the time of the encounter. Due to the lack of EHR centralization, the length and depth of a patient's record can vary greatly due to when a patient inhabits a region, what insurance the patient carries and the hospital accepts, and where a patient receives his or her care, with patients often seeing multiple disconnected providers within a region. One study to evaluate the effect of potential data fragmentation on the accuracy of a phenotyping algorithm for type 2 diabetes found that almost one-third of cases were missed if EHR data from only a single site was used (73). While completeness of the EHR is difficult to define, it is important for researchers to understand the likely limitations of the data and how it may affect study findings (74).

Approaches to Identifying Phenotypes in the EHR

Large-scale, efficient phenotyping methods that utilize data within the EHR provide benefits that are not limited to genomic studies. Phenotyping is crucial to the identification of a population of patients that satisfy a set of criteria, which is important for clinical trial recruitment, retrospective cohort or outcomes studies, and cost analyses, among others. However, identification of patients that belong in a particular cohort is time-consuming and challenging. Different approaches to automated phenotyping have been pursued; we describe the most frequent and promising (Figure 3, Table 3).

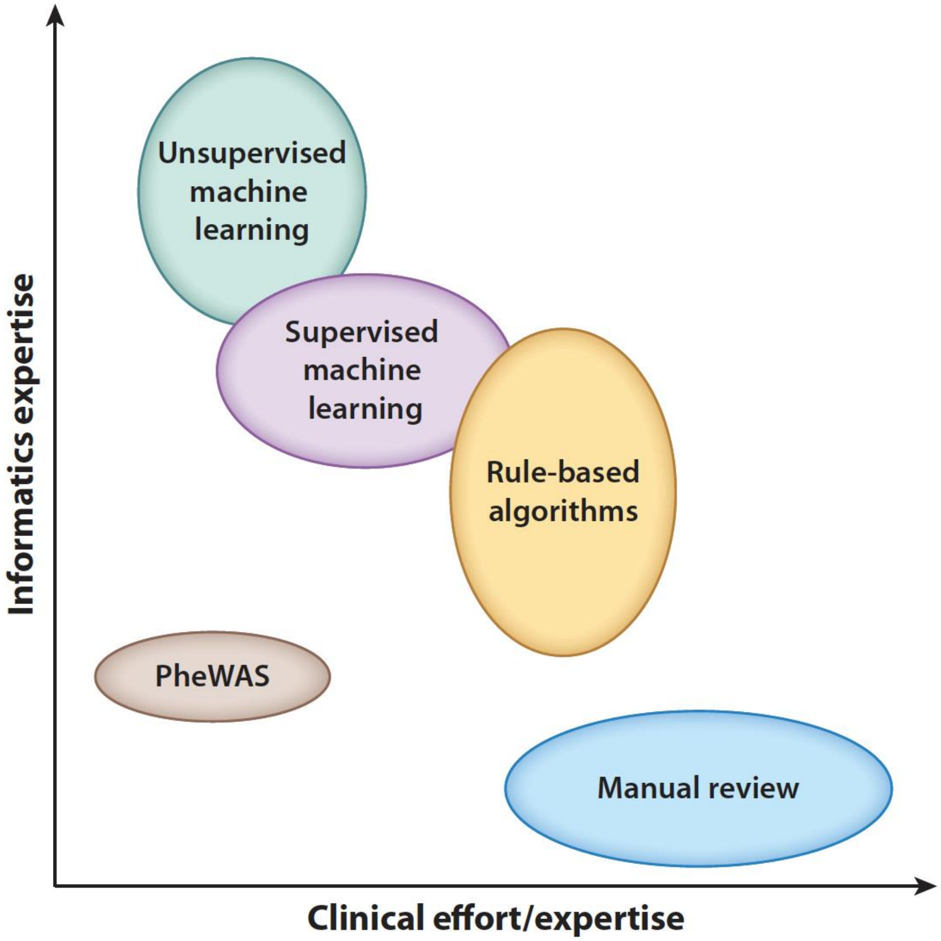

Unsupervised machine learning
Supervised machine learning
Rule-based algorithms
PheWAS
Manual review
Informatics expertise
Clinical effort/expertise

Figure 3. Phenotyping methodological approaches differ in their clinical and informatics resources needed. The width/height of each oval represents the range of that type of resource needed.

Table 3. Strengths and Weaknesses of Current Phenotyping Approaches using Electronic Health Records (EHRs).

	Manual chart review	**Rule-based computerized phenotyping**	**Supervised machine learning**	**Unsupervised machine learning**	**Phenome-wide approaches (PheWAS)**
Description	Clinician review of medical records	Manually-created Boolean logic	Classification based on a set of features	Phenotypes learned through clustering of computer-identified features	Systematic analysis of many phenotypes, usually using very simple rule-based approaches
Data variables	Can be used to narrow records for manual review	Structured data, NLP	Structured data, NLP	Any	Most commonly billing codes but can be lab or NLP

Recall	High	Variable	Variable	Variable	High
Precision	High	High	High	Variable	Variable
Costs	High	Moderate; major cost is in algorithm development	Moderate; major cost is in creating training set	Moderate; usually needs very large data collections and computing resources	Low
Scalability of cohort size	Poor; costs directly proportional to N	High	High	High (large cohorts required)	High
Scalability of phenotype number	Poor	Poor	Poor	High	High
Transportability	High	Variable but often high	Variable; often less than rule-based	Variable; often less than rule-based	High
Expertise Required	Clinical domain knowledge	Clinical domain knowledge and informatics support	Clinical domain knowledge and informatics support	Little to no clinical domain knowledge, significant informatics support needed	Little to no clinical domain knowledge, minimal informatics support needed
Strengths	High validity, considered gold-standard; little informatics support needed	High validity; Ease of interpretation	Typically less costly than manual review	Can identify unspecified phenotypes; Unbiased; No domain expertise or manual review	Ascertain broad range of phenotypes; can be used for very large populations
Weaknesses	Significant knowledge and time required	Requires iterative algorithm development and manual review; Domain and informatics expertise needed	Require manually classified training sets; may require expert feature selection	Potentially poor interpretability or irrelevant findings if not coupled with supervised approaches	Often phenotype detail and accuracy limited by billing codes
Current use	Idea for small cohorts and validation for computational methods	Often used in GWAS for single phenotypes	Often used in GWAS for single phenotypes	Rare, but increasing for identifying phenotypes from unstructured data	Hypothesis generation; drug repurposing; interpretation of novel genomic loci

*PheWAS, Phenome-wide association studies; NLP=natural language processing

Logical Constraint-Based Approaches

The earliest automated approach applied to recognize patients with a particular phenotype of interest is through the use of Boolean logic. The simplest example of this is utilizing hierarchical billing code structures to determine cases and controls, as is performed in a standard PheWAS analysis. However, more commonly these phenotype algorithms are more complex, manually-curated, and based upon rules applied in a step-wise fashion (75). Similar to the manner in which a content expert would determine case status, the logical constraint-based algorithms incorporate information from various sources of the EHR, including billing codes, clinical documents, laboratory data, and medication exposures.(19) Structured data, either data that is extracted from the EHR in structured way or is processed into a standardized format using methods such as NLP, are necessary as input for the algorithm.

Construction and validation of Boolean algorithms are typically an iterative process with collaborations between clinical domain experts, bioinformaticians, clinical informaticians, NLP experts, and genomics researchers. Clinical experts are typically required for creation of the algorithm itself, and manual effort is required for review of at least a subset of the case and control cohorts classified by each algorithm to ensure the algorithm's accuracy. The time and effort required is extremely phenotype dependent as the algorithms can vary from fairly simple to very complex. For example, the number of rules in an algorithm can differ dramatically: a 2012 review of 9 phenotypes found that algorithms contained between 8 and 174 rules for case identification (76).

Logical constraint-based approaches to phenotyping have several advantages. The most profound benefit is that they are the simplest to interpret and implement, allowing them to be more readily replicable and transferable across different EHRs or clinical enterprises (21, 26, 66). In 2011, a GWAS was performed to identify associations with primary hypothyroidism using a phenotyping algorithm developed at a single site and implemented to be transportable across EHRs and institutions (26). This algorithm incorporated billing codes, laboratory values, text queries, and medication records within five separate EHRs. Overall, the algorithms' PPVs were 92.4% and 98.5% for cases and controls, respectively. For the controls, PPVs at all sites were above 95%, while PPV for cases as varying sites ranged from 82-98%, with the lowest PPV mainly due to misclassification of individuals who had undergone thyroidectomy elsewhere or in the distant past. Experience has shown that algorithms developed for identification of rare phenotypes or those including only billing codes for case determination typically have the poorest performance.

These algorithms work well for circumstances in which high validity is needed for a single phenotype for a disease or simple trait that may need to be applied across multiple institutions. The ease of interpretation makes logical constraint-based algorithms attractive to clinicians, and thus research results potentially more translatable to clinical practice, and they can easily be applied to large data sets. The principal limitation is that a new algorithm needs to be created for each new phenotype pursued.

Machine Learning (ML) Approaches to Defining Phenotypes

ML-based algorithms have been proposed as a method to achieve the improved accuracy and breadth needed to scale phenotype annotation. ML approaches can automate the identification of complex patterns to classify individuals into different groups, such as a case or control for a given phenotype. The traditional approach to ML is that of supervised learning, in which an expert creates a "gold standard" of classified individuals and a feature set used for determination and then the trained algorithm can be used to make predictions on unlabeled examples (77). When individuals are not labeled into groups, an unsupervised learning approach can be applied, which attempts to find natural clusters or patterns of data and individuals. An unsupervised approach can also be used for feature extraction and then tested against an annotated outcome of interest for creation of the classifier (which can be viewed as a type of supervision). Fully unsupervised ML requires no need for domain expert annotation or feature selection; thus this approach is high-throughput and scalable but also can lead to greater difficulty in interpretation.

Supervised Machine Learning

Supervised ML requires a set of training examples belonging to either a phenotype case or control and then can build a model that can be used on other examples for classification. Several different methods can be used for supervised ML, such as support vector machines (SVM) (66, 72, 78, 79), logistic regression (66, 80), random forests (70), or neural networks (81, 82).

While supervised ML has been shown to be extremely effective for individual tasks, the requirement for manual annotation, a time-consuming and costly process, and feature selection results in limited scalability (83). The conventional approach to building a set of annotated samples is to select a random pool of individuals to manually classify. Active learning approaches to ML-based phenotyping can potentially overcome the need for large annotated datasets. Chen et al. demonstrated the use of an uncertainty sampling algorithm to find and annotate only the most informative samples, those samples with the most uncertain features (83). This approach achieved similar classification results with annotation of a fraction of samples compared to the use of a randomly annotated set. ICD-9 codes have been used as a surrogate for defining training cases and controls to limit manual input; however, this confines phenotypes to those defined by billing codes (79, 84).

Feature selection is the technique of selecting a subset of potential terms or features to use in the ML model. ML-based classification algorithms can incorporate a range of features, including billing data, clinical documentation extracted with NLP, semantic terminologies, medication exposures, and laboratory data. Features can either be chosen from a set of all potential clinical concepts or from a refined set through application of algorithms (e.g., using univariate statistical associations, ML approaches, or penalized regression models) or by domain expert curation. Potential features within the EHR are vast, and for many phenotypes, it is impractical to find an optimal subset of features using manual approaches. Prior studies have suggested the use of unrefined feature sets including all clinical concepts to reduce the domain knowledge required for phenotyping (72). While this provides effective performance for some phenotypes, algorithms using unrefined feature sets typically have lower accuracy than those using expertly curated features, unless very large training sets are used. Bejan et al. proposed statistical feature selection by ranking features based upon their association with a case or control to improve algorithm performance for phenotype identification of pneumonia (85). Yu et al. created the Automated Feature Extraction for Phenotyping (AFEP) approach for disease phenotyping by using NLP applied to knowledge sources (such as Medscape and Wikipedia) to develop medical concepts relevant to a disease followed by concept screening, in which concepts that were either too rare, too common, or not relevant enough were excluded from the feature sets in automated ways (86). AFEP algorithms achieved accuracy comparable to the use of expert-curated feature sets for rheumatoid arthritis and coronary artery disease.

Another approach to limit the need for manually engineered features is to use deep learning. Deep learning allows for the construction of a hierarchy of progressively complex feature layers, with transformation into more abstract representations at higher levels by training a neural network. The key aspect of deep learning is that these layers of features are not designed by domain experts, and rather they are learned from the data (87). Gulshan et al. applied supervised deep learning to more than 128,000 images of patients with diabetic retinopathy along with controls, with the development of algorithms that identified diabetic retinopathy with similar performance to ophthalmologists (AUC of 0.99) (81). Similarly, Esteva et al. used deep neural networks for automated classification of skin lesions, demonstrating the capability of identifying skin cancer with performance comparable to dermatologists (82).

In an early demonstration of deep learning neural networks applied to longitudinal health data, Lasko et al. applied an unsupervised feature learning approach to produce phenotypic features from uric-acid measurements for inputs in a supervised logistic regression classifier, and the model was capable of accurately distinguishing (with an AUC of 0.97) between gout and acute leukemia (88). A unique aspect of this study was managing the variable representation of time intervals, since labs are taken at irregular intervals for patients. Lab values for unobserved times were estimated using Gaussian probability distributions to allow drawing of continuous curves that formed the learned features from the deep learning model.

An important limitation of supervised ML is that the methods require training sets labeled with the phenotypes that they will "learn" to find. Thus supervised learning can be successful in finding patterns that explain phenotypes we have the knowledge to label, but not when we don't have the knowledge or ability to label the phenotypes and rather aim to discover phenotypes from the data (88).

Unsupervised Machine Learning

Unsupervised ML has demonstrated success using an unbiased approach towards phenotype discovery. This method could conceivably identify all phenotypes in a data set, including disease subtypes, medication response or adverse events, and previously unrecognized disease patterns. There are several types of unsupervised ML methods, including clustering, dimensionality reduction techniques, and tensor factorization. Prior studies have used unsupervised ML and hierarchical clustering to learn subphenotypes (or sets of comorbidities that group together) of autism (89, 90). Others have used clustering methods to identify distinct phenotypes of bicuspid aortic valves (91) and bronchiectasis (92). Nonnegative tensor factorization has also been shown to have the ability to capture diverse multi-attribute phenotypes consisting of combinations of diagnoses and medications with over 80% found to be clinically meaningful (93).

Unsupervised ML-based approaches may have the potential to scale phenotype classification to thousands of phenotypes in a computable and accurate way. The limitation of unsupervised ML models is their difficulty with interpretability, necessary for wide adoption by clinical providers (94). Further, ML-based algorithms are only as useful as the features used for classification, thus ontologies and advancement of NLP techniques will continue to improve their predictions as well.

Phenotyping for Genomic Research

Phenotype Creation, Validation, and Implementation for Genomic Research

Phenotype creation and validation is often an iterative process and critical step in determining the success and scalability of the phenotype as shown in Figure 4 (30). Phenotype accuracy has been shown in several studies to improve with combinations of data within phenotype algorithms (29, 34, 70). Even simply adding the requirement of 2 or more ICD codes can significantly increase PPV and sensitivity of algorithms; however, inclusion of EHR components, such as medication records and clinical notes, often results in further phenotyping performance improvement (29). Following initial creation and execution of the algorithm at the development site, it is important to validate the results at the primary site before executing at subsequent sites. Validation is typically performed by manual review of a subset of clinical records, and can be achieved by either an “expert review” by a seasoned clinician who reviews the clinical record to determine holistically if the patient meets the case definition or not (e.g., does the patient clinically appear to be diagnosed by appropriate clinicians with the disease of interest) or using formal case review algorithms, which specify finding a number of elements in the chart to verify the case is truly a case. The latter may be particularly important for more precise phenotypes, such as defining resistant hypertension or a subtype of eye disease, or when a clinical expert is not available for chart review.(30) The scope of review can vary based on the phenotype, for example while some algorithms may only require review of 1 year of a clinical record, others may require a full review of patient’s health record. After validation and tuning at one site, then the algorithm is executed and validated at secondary sites, preferably occurring across multiple EHRs and institutions, ensuring its reproducibility and transportability.

Curation and validation of a single or few phenotyping algorithms using the above described method to ensure high PPV is feasible and remains a highly-utilized method for GWAS.

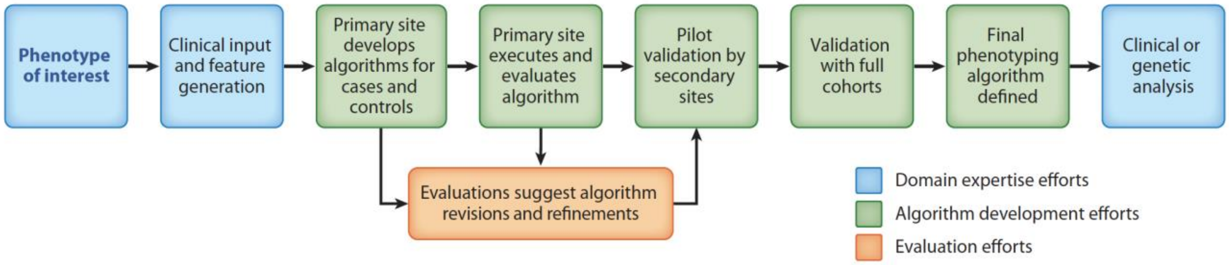

Phenotype of interest
Clinical input and feature generation
Primary site develops algorithms for cases and controls
Primary site executes and evaluates algorithm
Pilot validation by secondary sites
Validation with full cohorts
Final phenotyping algorithm defined
Clinical or genetic analysis
Evaluations suggest algorithm revisions and refinements
Domain expertise efforts
Algorithm development efforts
Evaluation efforts

Figure 4. Phenotype development and validation. A primary site first develops and executes the phenotype, and then secondary sites execute the phenotype. At each step feedback to primary and secondary sites may lead to revisions in the methods (arrows).

Scalability and Portability of EHR-derived Phenotyping Algorithms

For multi-site collaboration and accrual of large datasets, it is important to consider the portability of phenotypes across different institutions with varying EHRs, infrastructures, clinical domain knowledge, and informatics support. Researchers within eMERGE and other networks have demonstrated the portability of well-defined logical constraint-based phenotyping algorithms and have fostered sharing through public availability on PheKB. Kirby et al. described the results of 43 phenotyping algorithms including multisite validation data in PheKB, with a median of 3 (range 1–8) external validations per algorithm (21). Performances on case and control algorithms for development-site evaluations were similar to performance by external-site evaluations, with median case and control PPV over 95% for both. PheKB currently contains 154 phenotypes, 44 of which are publicly available (Table 2). As there is presently no uniformly-adopted method for data representation and extraction across varying institutions and EHRs, the algorithms are typically represented as ‘pseudocode’ to improve transportability and guide other sites at implementation, containing all necessary variables and the rules to combine them (30). Pathak et al. suggested that standardization of phenotype data dictionaries in these phenotyping algorithms using common data elements and biomedical ontologies could also help facilitate multi-site and cross-study collaborations (95); these ideas were later expanded into a desiderata for phenotype algorithms which includes logic formalism, use of ontologies, and the need for a common phenotype language that removes the need for human reimplementation from pseudocode (96).

In 2012, Carroll et al. utilized a previously published logistic regression phenotyping algorithm for identification of cases and controls for rheumatoid arthritis as a model to demonstrate portability of a ML algorithm across institutions (66). Features obtained from structured data were comprised of billing codes, laboratory results, medication orders, and use of general-purpose NLP systems. Applying the previously published regression phenotyping model to 2 external institutions with different EHRs and NLP systems showed an AUC to be similar to that obtained at the development site (92% and 95% at external sites compared to 97% at development site).

Logical constraint-based algorithms are largely rule-based, relying heavily on domain experts for curation. Similarly, supervised ML algorithms require annotated sets and often defined features for a specific phenotype. While these algorithms achieve high accuracy and portability across institutions, their scalability to a phenome-wide approach remains limited. In addition, many traits are not dichotomous but present along a continuum in a population. A recent study by Wells et al. demonstrated the use of left ventricular function determined by systolic ejection fraction, a continuous measure, to analyze associations with drug side effects in a GWAS analysis (97). Other opportunities for defining phenotypes include incorporating ontologies such as SNOMED-CT or the human phenotype ontology (HPO), which was originally designed to capture phenotypes related to Mendelian disease but has grown to encompass an increasing representation of common diseases (98).

As genotyping costs continue to decline and efforts become more widespread, the limiting factor to identification of genotype-phenotype associations will be accurate labeling of phenotype cases and controls. Research questions that require identification of many cohorts or phenotypes can become exponentially more challenging, with the resultant need for high-throughput phenotyping methods. Unsupervised ML algorithms to broadly define phenomes may have potential to scale phenotype discovery; however, their current use in genetic association analyses is limited.

Phenome-wide Association Studies

Analogous to a GWAS, PheWAS leverages the breadth of phenotypes in the EHR to perform systematic interrogation for associations with an independent variable, typically a genotype. The first PheWAS was performed in 2010, in which four known SNP-disease associations were replicated and several new proposed (99). Since then, dozens of studies have used PheWAS to explore both genetic and phenotypic associations to specific traits.

The PheWAS method requires a broad set of phenotypes collected in an unbiased approach to create a complete phenome of diseases and traits. PheWAS can use thousands of phenotypes; thus manual curation and validation is not practical (99–101). Thus to define a complete phenome across a large cohort, many PheWAS have used phenotypes derived from custom groupings of ICD-9 codes, also referred to as phecodes (99, 102). Typically, 2 or more ICD-9 codes are required for mapping to a single phecode. While the phecode groupings have been shown to better align with clinical diseases in practice, other methods of phenotype classification also are effective for PheWAS studies (100, 101, 103–105). Several studies have reported success with a PheWAS method using raw ICD-9 codes and parent ICD-9 three-digit groups as phenotypes (100, 103, 104). Leader et al. compared five gold standard phenotypes to ICD-9-Clinical Modification (CM) 5-digit and 3-digit diseases and phecodes (103). They found that phecodes may not be granular enough for some phenotypes. Others have also used Agency for Healthcare Research and Quality (ARHQ) Clinical Classification Software for ICD-9-CM (CCS), which reorganizes disparate ICD-9-CM codes into a smaller number of clinically meaningful categories (105, 106). Similar to three-digit ICD-9-CM codes or phecodes, CCS provides a hierarchical grouping of ICD-9-CM or ICD-10-CM codes but at a more aggregate level than other approaches. Wei et al. recently compared different diseases studied in prior GWAS to determine which phenotyping method better aligned with clinical practice and prior genomic association results and found that phecodes replicated more known SNP-phenotype associations (153 SNP-phenotype pairs) than use of ICD-9-CM (143) or CCS (139) (105).

Finding specific phenotypes that are not captured by billing codes or uncommon disease associations is crucial and will require increased granularity and accuracy of phenotyping techniques. An example of a trait rarely captured by billing codes alone is that of drug response, important for pharmacogenomic studies (107). Hebbring et al. has shown that PheWAS can be performed by defining the phenome solely on textual data within clinical documentation (101). Using clinical text extracted using NLP along with billing codes and other data available in the EHR could help automate and refine phenotypes at the phenome-wide scale.

PheWAS has shown the feasibility of analyzing genomic associations with thousands of phenotypes across a cohort of individuals and finding novel associations (99, 100, 102, 108). This approach can also be applied using various predictive attributes in the PheWAS analysis, including genetic risk scores, a set of SNPs aggregated into a single continuous score, or gene expression data. Krapohl et al. demonstrated how a genetic risk score could be applied to a phenome consisting of behavioral or psychiatric traits to explain some of the phenotypic variation seen in a population (109). Gamazon et al. have described PrediXcan, a method that predicts tissue-specific gene expression and could be used in PheWAS to measure trait associations as well (110). Mosley et al. applied generalized linear mixed models in a phenome-wide approach to estimate the additive genetic variance underlying phenotypes to prioritize those diseases and traits more likely to have genetic drivers, identifying a few conditions for which novel genetic signals were discovered via subsequent GWAS (111).

The flexibility and ease of application of the PheWAS approach makes the methods highly generalizable. Expansion of PheWAS-like methods with other high-throughput phenotyping techniques that capture the breadth of phenotypes in clinical practice has the potential to further advance discovery.

Challenges and Opportunities of EHR Phenotypes for Genomics Research

Challenges Learned from Current Phenotyping Efforts

Several challenges remain to optimizing use of EHR data. In addition to the inherent limitations of reusing EHR data for research, such as data fragmentation, variable accuracy, and extent of unstructured data, there are significant challenges to phenotype algorithm creation, stability, and scalability. Evolving billing standards, such as the transition to ICD-10, changing documentation practices, and advancing EHR systems all weaken the stability of phenotyping algorithms. EHR-linked biobanks introduce further challenges. In particular, there can be significant loss to follow-up as individuals change providers or institutions. Further, many EHR-linked biobanks prevent the ability to recontact patients due to de-identification of the EHR prior to the conduct of genotyping and research.

One major challenge both within EHR-linked biobanks as well as with combining genomic data from different institutions in multi-site collaborations is the usage of selected patient cohorts genotyped on varying platforms. Often genotyping of all consented individuals at an institution is not financially or logistically feasible. Thus, priorities are made for genotyping of individuals with known phenotypes of specific research interest or external funding for genotyping expenses. This has resulted in relatively large sample sizes of specific phenotypes at any one site, which can also influence what other phenotypes are observed in the genotyped population. In addition to population selection for genotyping, a variety of different genotyping platforms may be used. These constraints result in specific biases within certain biobanks that must be considered by researchers.

Another challenge is variable success of some algorithms. For example, type 2 diabetes has shown replicability with PPV greater than 98% for cases across institutions while the PPV for accurate identification of dementia ranged from 73-90% at different sites (30). For dementia in particular, cases were compared to a research-quality dementia diagnosis at a specific site obtained through a longitudinal cohort study, showing that billing codes and medications alone were not sufficient for accurate prediction of a dementia diagnosis. Unfortunately, there is currently no reliable method of determining preemptively which algorithms will perform well across different sites; thus validation, an iterative and time-consuming process, remains a crucial step.

Future Directions

Collaborations between researchers and clinicians could allow for integration of data obtained in prospective research initiatives into EHRs. In addition, linkage of EHR data to medical claims data and nationally public databases, such as prescription data and national death indices could also improve accuracy of phenotyping algorithms (30). In contrast to EHRs, which are used primarily by healthcare providers, participant-facing collection of health data has grown dramatically. While not a focus of this review, phenotypes for genomic research obtained through consumer-facing organizations such as 23andMe and the upcoming *All of Us* Research Program also have the potential to accrue a wealth of phenotypic information for genomic research (24, 112–114). Further, as personal health records and consumer health tracking applications on smartphones or other consumer devices increase in functionality and integration with other aspects of patient health information, these have the potential for the development of more refined phenotype definitions (18, 107–109).

Although use of ML and novel statistical methods for phenotype classification are growing, rule-based approaches continue to dominate. Use of standard terminologies within phenotype algorithms improves portability. Globally, there are challenges with collaboration due to differences in billing standards, such as the use of ICD-9, ICD-10, or ICD-10 with Clinical Modifications in the US. For example, while SNOMED-CT is the most common emerging terminology used in the United States, the United Kingdom has historically used Read codes in general practice clinics.

With no current unified programming language or phenotype implementation approach, the actual code for phenotyping algorithms remains largely non-portable today (118). The Phenotype Execution Modeling Architecture (PhEMA) project has the goal of developing reusable and machine-executable phenotype algorithms across sites and EHR systems (119). A key challenge, however, is not the programming language itself but the semantic representation of the data within the EHR. The Shared Health Research Information Network (SHRINE) has allowed for shared phenotype queries across systems implementing the Integrating Biology and the Bedside (i2b2) platform (120). Another promising approach to harmonizing data from heterogeneous EHRs is the use of a common data model (CDM) to create a common format for the data elements. Several different CDMs are in widespread use, including the Observational Medical Outcomes Partnership (OMOP) CDM (121) and the Patient-centered Clinical Research Network (PCORnet) CDM (122). When transformed into a CDM, ideally the same code can be executed across disparate sites and EHR systems.

The creation of precision medicine initiatives through biobanks allows for discovery into phenotypes across populations of patients with diverse genetic and environmental backgrounds. Genomic discovery has historically lacked in ancestral diversity (123, 124). Research collaboration using these large cohorts will help to better understand the role of genomics in phenotype variation internationally.

The overall goal of integration of genomic and phenotypic information with the EHR is not only to foster advances in research, but also to drive clinical decision making to support precision medicine (125). For example, implementation of algorithms to predict who may be at risk for being placed on particular medications could guide who is genotyped prior to medication therapy (126). Phenotyping algorithms could improve identifying populations for public health measures and decision support.

The National Academy of Medicine has recognized that healthcare systems are falling short it the ability integrate the wealth of knowledge and innovation into improvement in quality, outcomes, and cost (127). Much of the genetic information from genome-wide sequencing or genotyping is either not accessible to clinicians, not actionable at this time, or not in a computable form available for clinical decision support and future research (128). Providing accessibility of genomic information to clinicians through the EHR comes with many challenges, including data storage, structuring, and visualization, and can be overwhelming to a clinician (129). A vital step in achieving a learning healthcare system that incorporates genomic knowledge is the development of genomic clinical decision support for clinicians.

The ability to integrate genomic information into the EHR in a way that is accessible, logical, interpretable, and reusable will leverage the translation of genomic discovery to the bedside.

Conclusions

While integration of genomic information into EHRs has yet to fully reach its potential, significant work in phenotyping and biobanking efforts are providing an improved understanding of potential challenges to guide the future. The capability of data collection and extraction in the EHR has made it feasible for genomic and phenomic studies to have significant impacts on modern genomic research. Efforts towards the centralization of information and design of systems that meet the needs of collaborative research along with clinical and billing requirements will greatly facilitate advancement in phenotyping studies. Current successful initiatives worldwide provide the framework for phenotyping and genomic progress and innovation. Continued advances and standardization in EHR data abstraction and visualization, along with structured methods for integrating and representing genomic information will afford generalizable approaches to promote precision medicine.

Acknowledgments

JR Robinson support by the T15 LM007450 training grant from the NIH National Library of Medicine. This work is also supported by NIH funding through R01 LM010685, U01 HG008672, R01 GM 103859, and P50 GM115305.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106(23):9362–67

2. The Office of the National Coordinator for Health Information Technology. *Adoption of electronic health record systems among u.s. non-federal acute care hospitals: 2008-2015*. /evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php

3. Ginsburg GS, Burke TW, Febbo P. 2008. Centralized biorepositories for genetic and genomic research. *JAMA*. 299(11):1359–61

4. Denny JC. 2014. Surveying recent themes in translational bioinformatics: big data in EHRs, omics for drugs, and personal genomics. *Yearb. Med. Inform.* 9:199–205

5. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. 2010. A genome-wide association study of red blood cell traits using the electronic medical record. *PloS One*. 5(9):

6. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86(4):560–72

7. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, et al. 2010. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation*. 122(20):2016–21

8. Denny JC, Bastarache L, Roden DM. 2016. Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* 17:353–73

9. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, et al. 2014. Biobanks and electronic medical records: enabling cost-effective research. *Sci. Transl. Med.* 6(234):234

10. Wei W-Q, Denny JC. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* 7(1):41

11. Institute of Medicine. 2003. Key capabilities of an electronic health record system: Letter report

12. Kaye JA, del Mar Melero-Montes M, Jick H. 2001. Mumps, measles, and rubella vaccine and the incidence of autism recorded by general practitioners: a time trend analysis. *BMJ*. 322(7284):460–63

13. Asch SM, McGlynn EA, Hogan MM, Hayward RA, Shekelle P, et al. 2004. Comparison of quality of care for patients in the Veterans Health Administration and patients in a national sample. *Ann. Intern. Med.* 141(12):938–45

14. Croen LA, Yoshida CK, Odouli R, Newman TB. 2005. Neonatal hyperbilirubinemia and risk of autism spectrum disorders. *Pediatrics*. 115(2):e135-138

15. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, et al. 2016. Characterizing treatment pathways at scale using the OHDSI network. *Proc. Natl. Acad. Sci. U. S. A.* 113(27):7329–36

16. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. 2016. Building the Partners Healthcare Biobank at Partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *J. Pers. Med.* 6(1):

17. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. 2010. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin. Transl. Sci.* 3(1):42–48

18. *NOT-OD-14-124: NIH Genomic Data Sharing Policy*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

19. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, et al. 2011. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* 3(79):79re1

20. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, et al. 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70:214–23

21. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, et al. 2016. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc. JAMIA*. 23(6):1046–52

22. Elliott P, Peakman TC, UK Biobank. 2008. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* 37(2):234–44

23. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12(3):e1001779

24. National Institutes of Health (NIH) - All Of Us | National Institutes of Health (NIH). https://allofus-nih-gov.proxy.library.vanderbilt.edu/

25. Kohane IS. 2011. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* 12(6):417–28

26. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, et al. 2011. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* 89(4):529–42

27. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, et al. 2009. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int. J. Epidemiol.* 38(1):263–73

28. Sham PC, Purcell SM. 2014. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* 15(5):335–46

29. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. 2016. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc. JAMIA.* 23(e1):e20-27

30. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, et al. 2013. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J. Am. Med. Inform. Assoc. JAMIA.* 20(e1):e147-154

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

32. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, et al. 2010. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J. Am. Med. Inform. Assoc. JAMIA*. 17(4):383–88

33. Robinson JR, Denny JC, Roden DM, Van Driest SL. 2017. Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records. *Clin. Transl. Sci.*

34. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, et al. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 350:h1885

35. CLAMP Development Team. *CLAMP | Natural Language Processing (NLP) Software*. http://clamp.uth.edu/

36. Denny JC, Smithers JD, Miller RA, Spickard A. 2003. "Understanding" medical school curriculum content using KnowledgeMap. *J. Am. Med. Inform. Assoc. JAMIA*. 10(4):351–62

37. Friedman C, Shagina L, Lussier Y, Hripcsak G. 2004. Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc. JAMIA*. 11(5):392–402

38. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, et al. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc. JAMIA*. 17(5):507–13

39. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, et al. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J. Am. Med. Inform. Assoc. JAMIA*. 18(5):601–6

40. Salmasian H, Freedberg DE, Friedman C. 2013. Deriving comorbidities from medical records using natural language processing. *J. Am. Med. Inform. Assoc. JAMIA*. 20(e2):e239-242

41. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. 2010. MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc. JAMIA*. 17(1):19–24

42. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, et al. 2012. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *J. Am. Med. Inform. Assoc. JAMIA*. 19(5):859–66

43. Friedman C, Hripcsak G, Shagina L, Liu H. 1999. Representing information in patient reports using natural language processing and the extensible markup language. *J. Am. Med. Inform. Assoc. JAMIA*. 6(1):76–87

44. Denny JC, Spickard A, Miller RA, Schildcrout J, Darbar D, et al. 2005. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. *AMIA Annu. Symp. Proc. AMIA Symp.*, pp. 196–200

45. Elkin PL, Ruggieri AP, Brown SH, Buntrock J, Bauer BA, et al. 2001. A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. *Proc. AMIA Symp.*, pp. 159–63

46. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, et al. 2008. NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA Annu. Symp. Proc. AMIA Symp.*, pp. 172–76

47. Chapman WW, Chu D, Dowling JN. 2007. ConText: An algorithm for identifying contextual features from clinical text

48. Zhang Y, Tang B, Jiang M, Wang J, Xu H. 2015. Domain adaptation for semantic role labeling of clinical text. *J. Am. Med. Inform. Assoc. JAMIA*. 22(5):967–79

49. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, et al. 2015. A preliminary study of clinical abbreviation disambiguation in real time. *Appl. Clin. Inform.* 6(2):364–74

50. Sun W, Rumshisky A, Uzuner O. 2015. Normalization of relative and incomplete temporal expressions in clinical narratives. *J. Am. Med. Inform. Assoc. JAMIA*. 22(5):1001–8

51. Sohn S, Savova GK. 2009. Mayo clinic smoking status classification system: extensions and improvements. *AMIA Annu. Symp. Proc. AMIA Symp.* 2009:619–23

52. Uzuner O, Goldstein I, Luo Y, Kohane I. 2008. Identifying patient smoking status from medical discharge records. *J. Am. Med. Inform. Assoc. JAMIA*. 15(1):14–24

53. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, et al. 2012. A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu. Symp. Proc. AMIA Symp.* 2012:577–86

54. Friedlin J, McDonald CJ. 2006. Using a natural language processing system to extract and code family history data from admission reports. *AMIA Annu. Symp. Proc. AMIA Symp.*, p. 925

55. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *J. Am. Med. Inform. Assoc. JAMIA*. 16(6):806–15

56. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. 2014. Automated extraction of family history information from clinical notes. *AMIA Annu. Symp. Proc. AMIA Symp.* 2014:1709–17

57. Mehrabi S, Krishnan A, Roch AM, Schmidt H, Li D, et al. 2015. Identification of patients with family history of pancreatic cancer--investigation of an NLP system portability. *Stud. Health Technol. Inform.* 216:604–8

58. Centers for Medicare and Medicaid Services. 2012. Stage 2 Eligible Professional Meaningful Use Core Measures Measure 5 of 17

59. Centers for Medicare and Medicaid Services. 2012. Stage 2 Eligible Professional Meaningful Use Menu Set Measures Measure 4 of 6

60. Orlando LA, Buchanan AH, Hahn SE, Christianson CA, Powell KP, et al. 2013. Development and validation of a primary care-based family health history and decision support program (MeTree). *N. C. Med. J.* 74(4):287–96

61. Wu RR, Himmel TL, Buchanan AH, Powell KP, Hauser ER, et al. 2014. Quality of family history collection with use of a patient facing family history assessment tool. *BMC Fam. Pract.* 15:31

62. Garla V, Re VL, Dorey-Stein Z, Kidwai F, Scotch M, et al. 2011. The Yale cTAKES extensions for document classification: architecture and application. *J. Am. Med. Inform. Assoc. JAMIA.* 18(5):614–20

63. Melton GB, Hripcsak G. 2005. Automated detection of adverse events using natural language processing of discharge summaries. *J. Am. Med. Inform. Assoc. JAMIA.* 12(4):448–57

64. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, et al. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA.* 306(8):848–55

65. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. 2012. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin. Pharmacol. Ther.* 92(2):228–34

66. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, et al. 2012. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J. Am. Med. Inform. Assoc. JAMIA.* 19(e1):e162-169

67. Delaney JT, Ramirez AH, Bowton E, Pulley JM, Basford MA, et al. 2012. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin. Pharmacol. Ther.* 91(2):257–63

68. Mosley JD, Shaffer CM, Van Driest SL, Weeke PE, Wells QS, et al. 2016. A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. *Pharmacogenomics J.* 16(3):231–37

69. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* 6:30

70. Teixeira PL, Wei W-Q, Cronin RM, Mo H, VanHouten JP, et al. 2017. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J. Am. Med. Inform. Assoc. JAMIA*. 24(1):162–71

71. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF. 2009. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. *Int. J. Med. Inf.* 78 Suppl 1:S34-42

72. Carroll RJ, Eyler AE, Denny JC. 2011. Naïve electronic health record phenotype identification for rheumatoid arthritis. *AMIA Annu. Symp. Proc. AMIA Symp.* 2011:189–96

73. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, et al. 2012. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J. Am. Med. Inform. Assoc. JAMIA*. 19(2):219–24

74. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. 2013. Defining and measuring completeness of electronic health records for secondary use. *J. Biomed. Inform.* 46(5):830–36

75. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, et al. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc. JAMIA*. 21(2):221–30

76. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, et al. 2012. An evaluation of the nqf quality data model for representing electronic health record driven phenotyping algorithms. *AMIA. Annu. Symp. Proc.* 2012:911–20

77. Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16(6):321–32

78. Wei W-Q, Tao C, Jiang G, Chute CG. 2010. A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes. *AMIA Annu. Symp. Proc. AMIA Symp.* 2010:857–61

79. Peissig PL, Santos Costa V, Caldwell MD, Rottscheit C, Berg RL, et al. 2014. Relational machine learning for electronic health record-driven phenotyping. *J. Biomed. Inform.* 52:260–70

80. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, et al. 2010. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res.* 62(8):1120–27

81. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 316(22):2402–10

82. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 542(7639):115–18

83. Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, et al. 2013. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J. Am. Med. Inform. Assoc. JAMIA*. 20(e2):e253-259

84. Chiu P-H, Hripcsak G. 2017. EHR-based phenotyping: Bulk learning and evaluation. *J. Biomed. Inform.* 70:35–51

85. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. 2012. Pneumonia identification using statistical feature selection. *J. Am. Med. Inform. Assoc. JAMIA*. 19(5):817–23

86. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, et al. 2015. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J. Am. Med. Inform. Assoc. JAMIA*. 22(5):993–1000

87. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature*. 521(7553):436–44

88. Lasko TA, Denny JC, Levy MA. 2013. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS One*. 8(6):e66341

89. Doshi-Velez F, Ge Y, Kohane I. 2014. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*. 133(1):e54-63

90. Lingren T, Chen P, Bochenek J, Doshi-Velez F, Manning-Courtney P, et al. 2016. Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. *PloS One*. 11(7):e0159621

91. Wojnarski CM, Roselli EE, Idrees JJ, Zhu Y, Carnes TA, et al. 2017. Machine-learning phenotypic classification of bicuspid aortopathy. *J. Thorac. Cardiovasc. Surg.*

92. Guan W-J, Jiang M, Gao Y-H, Li H-M, Xu G, et al. 2016. Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics. *Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis.* 20(3):402–10

93. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, et al. 2014. Limestone: high-throughput candidate phenotype generation via tensor factorization. *J. Biomed. Inform.* 52:199–211

94. Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R. 2015. Causal phenotype discovery via deep networks. *AMIA Annu. Symp. Proc. AMIA Symp.* 2015:677–86

95. Pathak J, Wang J, Kashyap S, Basford M, Li R, et al. 2011. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J. Am. Med. Inform. Assoc. JAMIA*. 18(4):376–86

96. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, et al. 2015. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J. Am. Med. Inform. Assoc. JAMIA*. 22(6):1220–30

97. Wells QS, Veatch OJ, Fessel JP, Joon AY, Levinson RT, et al. 2017. Genome-wide association and pathway analysis of left ventricular function after anthracycline exposure in adults. *Pharmacogenet. Genomics*. 27(7):247–54

98. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, et al. 2017. The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* 45(D1):D865–76

99. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* 26(9):1205–10

100. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. 2013. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 14(3):187–91

101. Hebbring SJ, Rastegar-Mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. 2015. Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinforma. Oxf. Engl.* 31(12):1981–87

102. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31(12):1102–10

103. Leader JB, Pendergrass SA, Verma A, Carey DJ, Hartzel DN, et al. 2015. Contrasting association results between existing PheWAS phenotype definition methods and five validated electronic phenotypes. *AMIA Annu. Symp. Proc. AMIA Symp.* 2015:824–32

104. Verma A, Verma SS, Pendergrass SA, Crawford DC, Crosslin DR, et al. 2016. eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. *BMC Med. Genomics*. 9 Suppl 1:32

105. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS One*. 12(7):e0175508

106. Pathak J, Kiefer RC, Bielinski SJ, Chute CG. 2012. Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. *J. Biomed. Semant.* 3(1):10

107. Van Driest SL, Wells QS, Stallings S, Bush WS, Gordon A, et al. 2016. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA*. 315(1):47–57

108. Hebbring SJ. 2014. The challenges, advantages and future of phenome-wide association studies. *Immunology*. 141(2):157–65

109. Krapohl E, Euesden J, Zabaneh D, Pingault J-B, Rimfeld K, et al. 2016. Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry*. 21(9):1188–93

110. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47(9):1091–98

111. Mosley JD, Witte JS, Larkin EK, Bastarache L, Shaffer CM, et al. 2016. Identifying genetically driven clinical phenotypes using linear mixed models. *Nat. Commun.* 7:11433

112. 23andMe. *DNA Genetic Testing & Analysis - 23andMe*. https://www.23andme.com/

113. Annas GJ, Elias S. 2014. 23andMe and the FDA. *N. Engl. J. Med.* 370(11):985–88

114. Precision Medicine Initiative (PMI) Working Group. 2015. *The Precision Medicine Initiative Cohort Program—building a research foundation for 21st century medicine*. https://www-nih-gov.proxy.library.vanderbilt.edu/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf

115. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. 2006. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *J. Am. Med. Inform. Assoc. JAMIA*. 13(2):121–26

116. Roehrs A, da Costa CA, Righi R da R, de Oliveira KSF. 2017. Personal health records: a systematic literature review. *J. Med. Internet Res.* 19(1):e13

117. Gay V, Leijdekkers P. 2015. Bringing health and fitness data together for connected health care: mobile apps as enablers of interoperability. *J. Med. Internet Res.* 17(11):

118. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, et al. 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 15(10):761–71

119. Jiang G, Kiefer RC, Rasmussen LV, Solbrig HR, Mo H, et al. 2016. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. *J. Biomed. Inform.* 62:232–42

120. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, et al. 2009. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J. Am. Med. Inform. Assoc. JAMIA*. 16(5):624–30

121. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. 2012. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc. JAMIA*. 19(1):54–60

122. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. 2016. Data interchange using i2b2. *J. Am. Med. Inform. Assoc. JAMIA*. 23(5):909–15

123. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, et al. 2016. Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* 375(7):655–65

124. Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature*. 538(7624):161–64

125. Collins FS, Varmus H. 2015. A new initiative on precision medicine. *N. Engl. J. Med.* 372(9):793–95

126. Pulley JM, Denny JC, Peterson JF, Bernard GR, Vnencak-Jones CL, et al. 2012. Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. *Clin. Pharmacol. Ther.* 92(1):87–95

127. Institute of Medicine. 2013. Best care at lower cost: the path to continuously learning health care in america

128. Starren J, Williams MS, Bottinger EP. 2013. Crossing the omic chasm: a time for omic ancillary systems. *JAMA*. 309(12):1237–38

129. Kho AN, Rasmussen LV, Connolly JJ, Peissig PL, Starren J, et al. 2013. Practical challenges in integrating genomic data into the electronic health record. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 15(10):772–78

CHAPTER III

Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records

Jamie R. Robinson, M.D., M.S.[1,2], Joshua C. Denny, M.D., M.S. [1,3],
Dan M. Roden, M.D.[1,3,4], Sara L. Van Driest, M.D., Ph.D.[3,5]

Affiliations:

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center

[2] Department of Surgery, Vanderbilt University Medical Center

[3] Department of Medicine, Vanderbilt University Medical Center

[4] Department of Pharmacology, Vanderbilt University Medical Center

[5] Department of Pediatrics, Vanderbilt University Medical Center

This manuscript was published in the journal Clinical and Translational Science as follows:

Robinson JR, Denny JC, Roden DM, Van Driest SL. Genome-wide and phenome-wide approaches to understand variable drug actions in electronic health records. *Clinical and Translational Science*. 2018; 11(2): 112-122. PMID: 29148204.

Introduction

Genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS) have provided powerful methods for investigating the impact of genetic variation on individual drug response and have added extensive knowledge to the understanding of drug targets and effects. We highlight here recent advances in drug development, repurposing, and personalization accelerated by applying GWAS and PheWAS to longitudinal health data information, along with limitations of these methods.

Importance of understanding drug targets and effects

The challenges facing modern clinical pharmacology can be grouped into two major categories: the efficient development of novel therapeutics and understanding individual variability in response. The development of novel therapeutics is hampered by the problem that despite major advances in the knowledge of disease mechanisms, drug targets, and biomarkers, as well as a continual rise in investment into pharmaceutical research, the number of new drugs approved each year has remained steady.(1) Most drugs fail in Phase II clinical trials, with 50% of the failures due to lack of efficacy.(2, 3) Thus, there is concern that preclinical disease models do not reliably predict efficacy in patients. Human genetics has been proposed as a mechanism to prioritize molecular targets early in the stages of drug discovery towards potentially more efficacious models.(4, 5)

The problem of variable drug efficacy and susceptibility to side effects has been recognized since the advent of therapeutics. There is now evidence that medication exposure data, outcome data, and genetic information linked together via longitudinal electronic health records (EHRs) can provide a more thorough understanding of drug effects, including response patterns and individuals at risk for potentially rare side effects.(6) Knowledge of the genetic mechanisms that drive drug response variation and adverse events can help guide the tailoring of medication therapy. GWAS and PheWAS are modern genetic tools for the exploration of datasets in order to efficiently identify potential targets for novel therapeutics and to provide evidence-based individualized therapy.

Overview of GWAS and PheWAS approaches

GWAS is a hypothesis generating method to systematically analyze variants across the entire genome (i.e. "genome-wide") for association to a phenotype of interest (Figure 1a). Over the past 10 years, the genotyping assays have evolved from early versions assessing hundreds of thousands of single nucleotide polymorphisms (SNPs) to current panels including millions of SNPs.(7) At the same time, pipelines for quality control, imputation of genotypes, and statistical analysis for dichotomous, categorical and continuous traits have matured and become standardized across high-quality laboratories. The threshold for statistically-significant results, given the need to correct for multiple comparisons, has been established at $5x10^{-8}$. Despite this stringent threshold, use of larger and larger cohorts (including some recent cohorts including more than 700,000 individuals) has enabled identification of many SNPs with statistically significant p-values.(8) GWAS focuses on detection of associations with relatively common variants (e.g., minor allele frequencies 1-5%) so the odds ratios are often small (OR < 1.5). Thus, the major outcome of many GWAS is a better understanding of the genetic architecture of complex traits, but not a set of high effect size variants that would be clinically actionable.(7) As described below, examining drug response with GWAS has provided an interesting counter-example to these generalizations since small numbers can yield signals with large enough effect sizes to be considered for implementation. This may reflect the idea that drug response represents an example of a controlled environmental intervention (drug) interacting with a genome, rather than a multifactorial complex disease state with many potential environmental inputs. Since 2005, over 3000 GWAS have identified almost 40,000 unique SNP-trait associations within the GWAS catalog provided by the National Human Genome Research Institute (NHGRI) and European Bioinformatics Institute (EBI).(7) The rapid increase in knowledge of common genetic variants in complex diseases has provided significant opportunities for analyzing the association of genetic variation with disease phenotypes and response to therapies.

The integration of this wealth of genetic information with phenotypic data by linkage of DNA biobanks with EHRs has led to the development of a reverse genetics approach with EHR-based genomic studies, termed PheWAS (Figure 1b).(6) The first PheWAS study was performed in 2010, in which Denny et al. successfully replicated four known SNP-disease associations.(9) Since then, the use of PheWAS has continued to rise in popularity with 58 current PubMed indexed publications. PheWAS provides a systematic approach to analyze the many phenotypes potentially associated with a specific genotype, with the ability to identify pleiotropy, or the finding of multiple independent associations with a single genetic association.(10) The threshold for statistical significance is less well-established for PheWAS; therefore, often a Bonferroni correction is applied in the analyses. However, this is highly stringent as it assumes independence across all phenotypes, unlikely given that many phenotypes are closely related. Despite this, use of large cohorts have allowed for PheWAS to not only replicate known findings, but also identify novel associations.

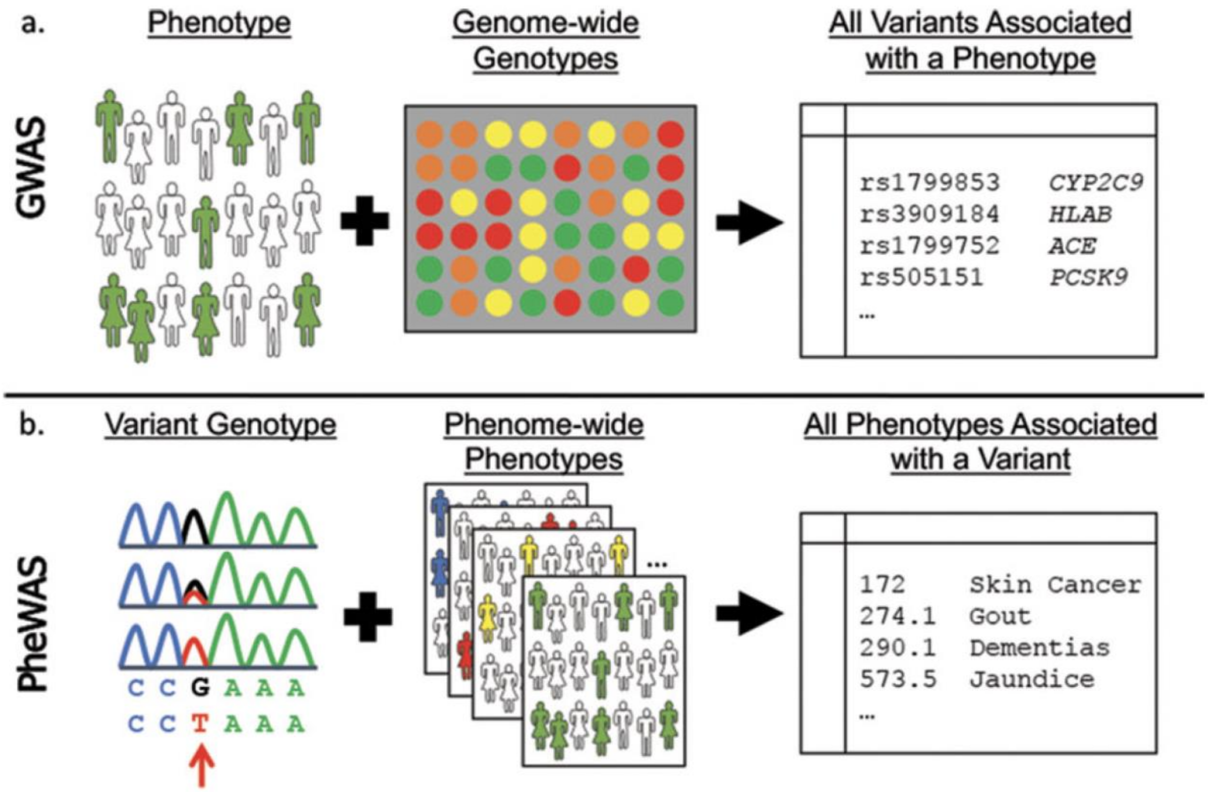


Figure 1.

GWAS and PheWAS approaches are complementary, with the ability to replicate and validate the other's findings. Representing the capacity for PheWAS to replicate GWAS, a comprehensive comparison of known GWAS associations within the NHGRI GWAS catalog against the PheWAS method was performed in 2013, showing that 210 of 751 (28%) known SNP-disease associations were replicated with PheWAS, including 66% of those associations that were adequately powered to detect the association.(11) This method also identified 63 potentially novel SNP-disease associations, again demonstrating pleiotropic effects of the variants. In a pediatric cohort, PheWAS replicated many prior known GWAS associations including SNPs associated with juvenile rheumatoid arthritis, asthma, autism and pervasive developmental disorder, and type 1 diabetes.(12) Several new SNP-disease associations were identified within the pediatric population as well, including a cluster of association near the *NDFIP1* gene (associated with mental retardation), *PLCL1* (developmental delay), and the IL5-IL13 region (eosinophilic esophagitis).(12)

Leveraging EHRs for drug-based genomic and phenomic research

The EHR provides a longitudinal collection of phenotypic information coupled with medication exposures, thus making it an important platform for study of drug effects.(6) A broad set of phenotypes collected in an unbiased approach is essential to the PheWAS method.(13) To accomplish this, many PheWAS have used phenotypes derived from custom groupings of billing codes, also referred to as phecodes.(9) The billing codes used for phecode groupings currently are International Classification of Diseases, Ninth Revision (ICD9) codes. ICD codes are published and maintained by the World Health Organization for classification of diseases and services for reimbursement of medical services. While the phecode groupings have been shown to better align with clinical diseases in practice, other methods of phenotype classification also are effective for PheWAS studies.(14) Hebbring et al. reported a PheWAS method using individual ICD9 codes and parent ICD9 three-digit groups as phenotypes.(15) They not only replicated a known association of an HLA class II allele, *HLA-DRB1*, with multiple sclerosis, but also replicated associations of *HLA-DRB1* with erythematous conditions and benign neoplasms of the respiratory and intrathoracic organs, found to be significant in a prior study by Denny and colleagues.(9, 15) These and other studies highlight the importance of PheWAS techniques for identifying pleiotropic effects.

Billing codes are not the only source of phenotypes from the EHR. Hebbring et al. have shown that PheWAS can be performed by defining the phenome solely on textual data within clinical documentation.(16) For drug effects, a phenome based upon billing codes or clinical text alone may not accurately capture drug efficacy or adverse events, nor do they provide the necessary information about drug exposure, including dosing data. One potential method of obtaining this drug exposure and outcome data is to use prospective cohort-based studies. It has been shown that PheWAS can be used with data obtained through clinical trials, representing a more biased, but targeted, approach at defining a phenome.(17, 18)

The benefit of leveraging EHRs for both GWAS and PheWAS as opposed to prospective cohort-based studies is the ability to obtain large sample sizes with relatively less time or expense. While the EHR phenome may be incomplete, it includes conditions that are medically relevant, as opposed to clinical trial cohorts in which phenotypes may not represent conditions that necessitate medical attention. Biobanking of genetic data linked to the longitudinal patient data available within the EHR provides an efficient method for aggregating otherwise disparate information. EHR-based biobanks have the potential to integrate genomic data with medication receipt, laboratory results, or textual data, thus refining both exposures and phenotypes, essential for research on drug effects.

Genomic investigation aids in understanding drug mechanisms

Several features of GWAS suggest its potential for elucidating drug mechanism and identifying relevant novel drug targets. An estimated 21% of published genes within the GWAS catalog are amenable to pharmacological modulation by small molecules.(19) Further, the GWAS gene set is enriched with drug targets in comparison to the entire human genome, many of which align with the disease-gene pair identified by GWAS analysis.(19)

Prior studies support the role of using GWAS to identify alleles that contribute to disease risk and druggable targets. Early GWAS efforts retrospectively identified the genetic basis for drugs already in use for a particular indication. Statins have been used to inhibit 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase and treat hyperlipidemia for decades,(20) and GWAS studies in 2008 showed that low-density lipoprotein (LDL) levels are associated with variation in *HMGCR*, the gene which encodes HMG-CoA reductase.(21, 22) Further, pharmacogenetics studies have shown that genetic variation in *HMGCR* is associated with statin efficacy.(23, 24) Since then, variants in other genes involved in lipid metabolism but not direct targets for statin action (*APOE, LPA, SORT1/CELSR2/PSRC1* and *SLCO1B1*) have been found in GWAS to be associated with the LDL-lowering effect of statin therapy.(25, 26) A recent GWAS also found that several variants with the *LPA* locus appear to be associated with coronary heart disease events during statin therapy, independent of the extent of LDL-cholesterol lowering.(27)

Other examples of drugs with mechanism replicated by GWAS are ustekinumab, a monoclonal antibody against interleukin (IL)-12 used for treatment of psoriasis and inflammatory bowel disease(28, 29), and denosumab, a monoclonal antibody to the receptor activator of nuclear factor-kappaB ligand (RANKL) for treatment of osteoporosis.(30) Metformin has long been used to lower blood glucose levels in individuals with diabetes; however, a 2011 GWAS of 3920 type 2 diabetes patients clarified the genetic basis for its mechanism with polymorphisms in the *ATM* gene found to be associated with glycemic control.(31) Okada et al. evaluated the role of GWAS in validating the current therapeutic drug targets for rheumatoid arthritis (RA).(32) Through a comprehensive genetic study with nearly 100,000 subjects, they found that 18 of 27 currently approved drugs for RA target genes identified as RA risk loci, and also suggest several potential novel therapeutics, some of which had supporting animal studies.(32) These early successes fuel enthusiasm for using GWAS to elucidate disease mechanisms and drug targets.(19)

Early evidence for drug discovery using genomic approaches

In the context of drug development, GWAS advances are relatively recent and are only now being applied to have a potential impact on target discovery. Nevertheless, prior linkage and candidate gene studies have shown that genetics can drive development of novel therapeutics. The development of proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors represents this realization. In 2003, it was found that autosomal dominant hypercholesterolemia and an increased incidence of coronary heart disease were associated with gain-of-function mutations in the *PCSK9* gene.(33) Subsequent candidate gene association studies in 2005 and 2006 revealed that *PCSK9* loss-of-function mutations correlate with reduced levels of LDL cholesterol and a lower incidence of coronary heart disease.(34, 35) In 2012, almost 10 years after the first genetic discovery, randomized controlled trials demonstrated that PCSK9-

specific monoclonal antibodies significantly reduce LDL cholesterol levels.(36, 37) There are now two United States Food and Drug Administration (FDA) approved PCSK9 inhibitors. Similar to how candidate gene studies led to the target of PCSK9 and subsequent development of a novel therapy for familial hypercholesterolemia, GWAS and PheWAS hold promise as means to identify novel drug targets. However, the timeline from target to an approved drug is often over 10 years. As findings from GWAS have exponentially increased over the last decade and PheWAS is gaining similar recognition, we anticipate the next decade will show progress toward utilization of that knowledge and drug development.

GWAS for understanding impact of genetic variation on drug efficacy

A considerable amount of variability can exist in patient's response to drug therapy, including differing efficacy, adverse side effects, and toxicity. A better understanding of the genetic determinants of drug response and mechanism is thought to have potential to individualize drug treatment toward improved efficacy and side effect profiles.(38) While candidate gene studies have shown success in identifying genetic variants that contribute to drug response and effects, for many drugs, the biological mechanism, metabolic pathways, and potential genetic associations impacting individual response is unknown, limiting the potential for focused gene analysis. In contrast, GWAS are a hypothesis-free method that can be utilized to determine associations of genetic variation with effects of drug treatment (Figure 2a).

Drug efficacy in particular is often considered to be along a continuum in a patient population. The known genetic variants contributing to statin efficacy discussed above are an indication of the significant clinical and genetic variability that can be seen in a population.(25, 26) A high-yield area of pharmacogenetic investigation utilizing GWAS has been the study of drugs with a narrow therapeutic window and variable efficacious dosing regimens, such as warfarin. While candidate gene studies were used to initially describe the associations of *CYP2C9* and *VKORC1* with the ability for warfarin to achieve anticoagulation,(39-42) subsequent GWAS have confirmed these findings, showing these to be the strongest genetic predictors of warfarin dose required in individuals of European descent.(43, 44) Subsequent GWAS in individuals of African ancestry has also found that in addition to the well-known *CYP2C9*2, CYP2C9*3,* and *VKORC1* polymorphisms, the CYP2C locus exerts influence by a variant outside of those well-established, and this new variant could improve dose prediction in this population.(45) Differences in the variants associated with warfarin effect across populations may be due to the differences in mean allele frequencies, ancestry-specific gene-gene interactions, or population specific gene-environment interactions.

Response to clopidogrel therapy is also known to be highly variable. Clopidogrel is a prodrug, and the bioactivation pathway is largely CYP2C19-dependent.(46) Candidate gene studies of cardiovascular events on clopidogrel indicated that *CYP2C19* loss of function variants increased risk.(47-49) Subsequent GWAS in Amish individuals found the most common loss of function allele, *CYP2C19*2*, had the strongest genetic association with the effect of clopidogrel on platelet aggregation; however, this single variant only accounts for approximately 12% of the variability in response seen in this population.(50) More recently, Zhong et al. identified two novel variants in a Chinese population that were associated with the antiplatelet effect of clopidogrel, as measured by P2Y12-mediated platelet aggregation, as well as formation of H4, an active metabolite of clopidogrel.(51) They estimate that the identified variants, in association with *CYP2C19*2*, *CYP2C19*3* (a variant common in Asian populations), and clinical factors, can improve the predictability of clopidogrel effect to 37.7%.

Another example of GWAS elucidating the genetic underpinnings of variability in drug efficacy is in the use of interferon-alpha for treatment of hepatitis C infection. A polymorphism adjacent to *IL28B* has shown to predict treatment response and viral clearance in individuals on interferon-alpha for hepatitis C in several GWAS.(52-54) Because the genotype associated with improved response is more common in individuals of Asian and European ancestry than African ancestry populations, this genetic polymorphism may explain the difference in response rates between patients of African and European ancestry.(52)

These GWAS of statin, clopidogrel, and interferon-alpha effects all emphasize that variability in drug response is ancestry dependent due to the vast difference in distribution of genetic variants, such as *CYP2C9*, *CYPC19*, and *IL28B*, across populations. These findings, along with others, have increased focus towards a personalized approach to disease treatment and encouragement of research efforts from individuals of diverse backgrounds. Studies across ancestries are needed to fully capture the genetic architecture of human traits, including drug response, and ultimately appropriately implement such variants in clinical practice.

GWAS for understanding impact of genetic variation on drug toxicity

GWAS has been used to determine potential associations of drug toxicities and adverse drug reactions (Figure 2a). Human leukocyte antigen (HLA) variation, in particular, has been associated by GWAS with susceptibility to adverse drug reactions. Drug-induced liver injury (DILI) is a rare but serious adverse effect secondary to many drug therapies, with increased susceptibility in HLA regions implicated in several studies.(55-57) The first study was in 2008 and focused on ximelagatran, an oral direct thrombin inhibitor that was removed from the market in 2006 due to the development of transaminitis in some patients. In this study, Kindmark et al. performed a GWAS which suggested an association between DILI during use of ximelagatran with HLA class II alleles, which was confirmed with candidate gene studies.(58) In 2009, Daly et al. found strong association of *HLA-B*5701* with DILI following treatment with flucloxacillin.(55) Singer et al., in 2010, identified an association of hepatotoxicity after use of lumiracoxib, a selective cyclooxygenase-2 inhibitor, with common HLA class II haplotypes.(56) In 2011, Lucena et al. performed GWAS of 201 cases of DILI after treatment with amoxicillin-clavulanate compared to 532 controls, finding HLA class I and II SNPs may confer susceptibility to liver injury after this antibiotic treatment.(57) Due to the rarity of DILI, the finding of genetic predispositions in GWAS may be limited. Nicoletti et al. recently attempted to overcome this limitation by grouping DILI caused by any drug other than the common causes (flucloxacillin and amoxicillin-clavulanate) to determine predisposition for DILI.(59) They found a strong association of *HLA-A*33:01* with DILI, appearing to be heavily influenced by the effects of terbinafine.(59) This demonstrates how novel methods in studies will allow researchers in some cases to overcome the power limitations of GWAS and find rare variants with large effect size.

The HLA locus has also been implicated in other adverse drug reactions, including skin hypersensitivity. In 2004, Chung et al. reported a strong association in a Han Chinese population between *HLA-B*1502* and Stevens-Johnson syndrome induced by carbamazepine.(60) Candidate gene studies were also used to ascertain an association between variation in the HLA region, *HLA-B*5701* (OR = 117), with abacavir skin hypersensitivity, which has since been elucidated both at a structural and mechanistic level.(61, 62) Several subsequent GWAS studies across ancestral populations have shown that skin hypersensitivity reactions, ranging from skin rash to severe reactions such as Stevens-Johnson syndrome/Toxic Epidermal Necrolysis, can occur secondary to a wide range of drug therapies. In 2011, Ozeki et al. identified the *HLA-A*3101* allele as a genetic risk factor with a modestly large effect size (OR = 10.8) for carbamazepine-induced hypersensitivity in a cohort of 53 cases and 882 controls from Japan.(63) McCormack et al. shortly after reported the same genetic association with carbamazepine-induced hypersensitivity reaction in individuals of European descent, finding a large effect size as well (OR = 12.4).(64)

One important early GWAS example is the study by Link et al., which discovered a single strong association of statin-induced myopathy in Europeans with a SNP located within *SLCO1B1*, known to encode an organic anion-transporting polypeptide that regulates the hepatic uptake of statins (OATP1B1).(65) While the variant allele frequency of this significant polymorphism is 0.13 in European populations, carriage of the variant allele resulted in an 18% incidence of myopathy over 5 years, with 60% of cases attributable to the variant allele. Thus, further studies are needed to define the mechanism(s) underlying this “variable penetrance”. A recent study by Mosley et al. evaluated the association of genetic variation with angiotensin-converting enzyme inhibitor (ACEi)-induced cough.(66) Cough is the most common side effect of ACEi therapy with epidemiologic variation that suggests a potential genetic predisposition. In GWAS consisting of 1695 cases of ACEi-induced cough compared to 5485 controls, SNPs in *KCNIP4* were associated with increased risk for developing cough with ACEi. In recent GWAS, genetic variation has also been implicated as increasing susceptibility to anthracycline-induced cardiotoxicity and reduced left ventricular function.(67, 68) Vancomycin, a commonly used antibiotic, is known to be nephrotoxic, with a GWAS suggesting variation at the chromosome 6q22.31locus could modulate that risk as well.(69)

PheWAS for understanding drug response variability

PheWAS also has the potential to uncover associations with drug effects, including therapeutic response and side effect profiles (Figure 2b). Neuraz et al. described in 2013 the use of a study population with thiopurine exposure to determine associations with clinical traits after drug exposure to identify adverse events.(70) They grouped 442 individuals with thiopurine exposure into three categories based upon thiopurine S-methyltransferase (TPMT) activity, a quantitative trait available from the EHR for patients with clinical TPMT testing. They found that very high TPMT activity was associated with diabetes mellitus and iron-deficiency anemia. Similarly, they analyzed associations with laboratory data, finding that very high TPMT activity was associated with increased incidence of hyperglycemia and anemia by test results. This study shows the ability for PheWAS to identify adverse events potentially associated with drug use, as well as the feasibility of cross-validation of conventional PheWAS analyses with biological test results.(70)

Others have noted the potential ability to leverage the identification of pleiotropic effects through PheWAS methods to predict potential adverse events. Diogo et al. analyzed associations of RA-protecting variants for additional indications and potential adverse events.(71) They did not identify any associations with adverse events meeting statistical significance in their study, suggesting that inhibition of tyrosine kinase 2 (TYK2) may not result in serious adverse events in the treatment of RA. However, the ranking of associations did potentially prioritize adverse events for study in a trial, and represents an analytical framework that could show success in the future. This study also highlights the very large populations needed for this study design; among over 20,000 individuals in the PheWAS of one cohort, a total of 2612 had pneumonia, the potential adverse event most trending toward statistical significance. Using PheWAS to suggest deleterious effects of evolving therapeutics early in the drug development stages could allow resources to target therapeutics with greater potential or can identify patient populations for whom the drug may be contraindicated.

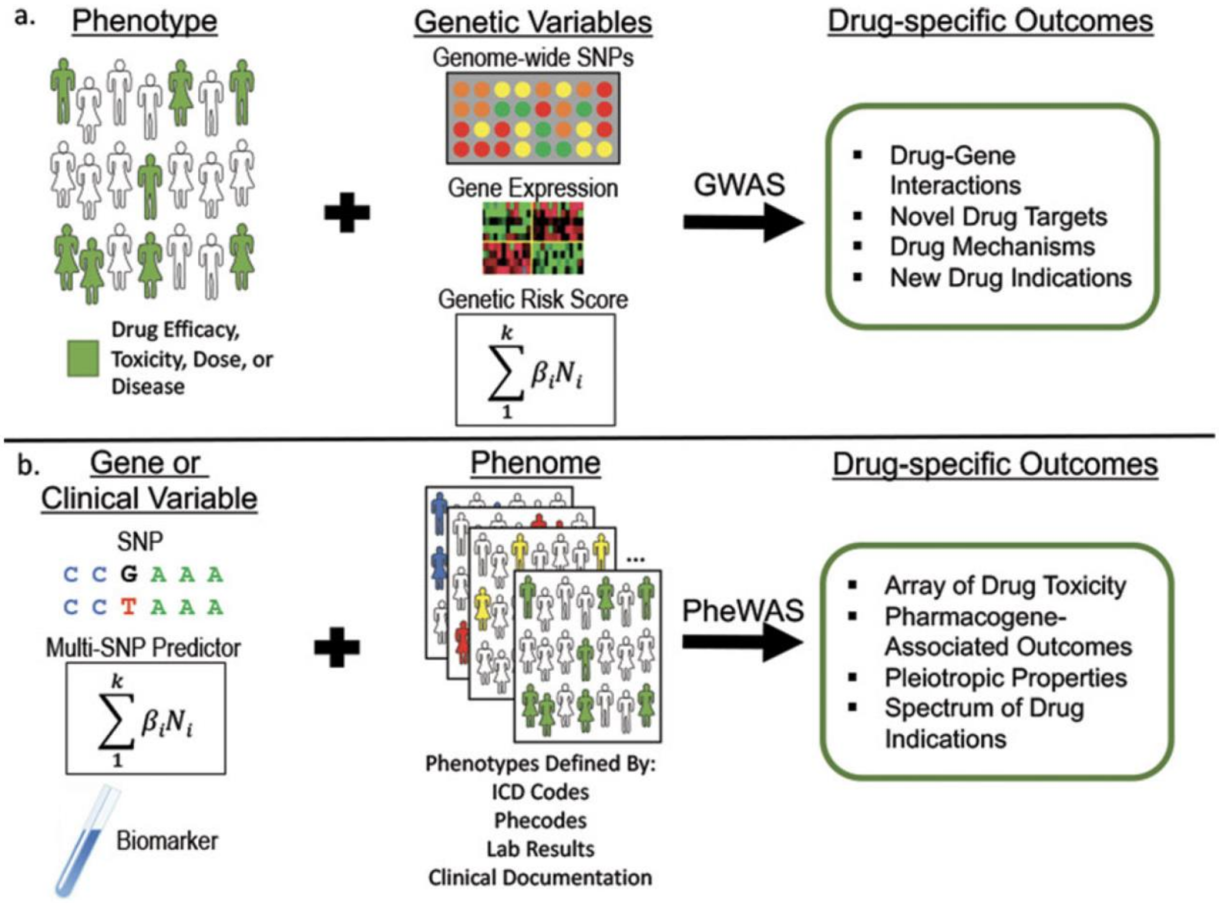


Figure 2. Drug-specific outcomes identified through genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). a. Use of drug-specific phenotypes of interest with genomic predictor variables such as genome-wide single nucleotide polymorphisms (SNPs), measured or predicted gene expression, or genetic risk scores, can be used in GWAS analysis to gain information for drug mechanisms and discovery. **b.** In PheWAS, genetic or clinical variables can be used to search for

associations in phenomes curated from different sources of information in the EHR for analysis of drug-specific associations.

Use of GWAS and PheWAS to identify opportunities for drug repurposing

In addition to elucidating drug mechanism and response variability, GWAS and PheWAS can be used to identify novel treatment methods through drug repurposing (Figure 3).(72, 73) Drug repurposing, also termed drug repositioning, is the application of an existing therapeutic drug for new indications. Drug repurposing could significantly speed up the typical >10 years lag time for FDA approval and drug marketing, as preclinical and phase I clinical trials are already complete. While the GWAS gene set is enriched with targets already pursued by drugs that align with the disease-gene pair identified by GWAS analysis, there are also mismatches in which the indication for the drug is not congruent with the associated disease by GWAS, and examples of pleiotropy, where multiple diagnoses are associated with the same genetic signal.(19) By comparing known GWAS-disease associations to the indications of drugs with known gene targets, Sanseau et al. identified 92 individual genes that are targets of drug projects that mapped to a GWAS trait different than their drug indication.(19) These instances represent potential drug repurposing opportunities.

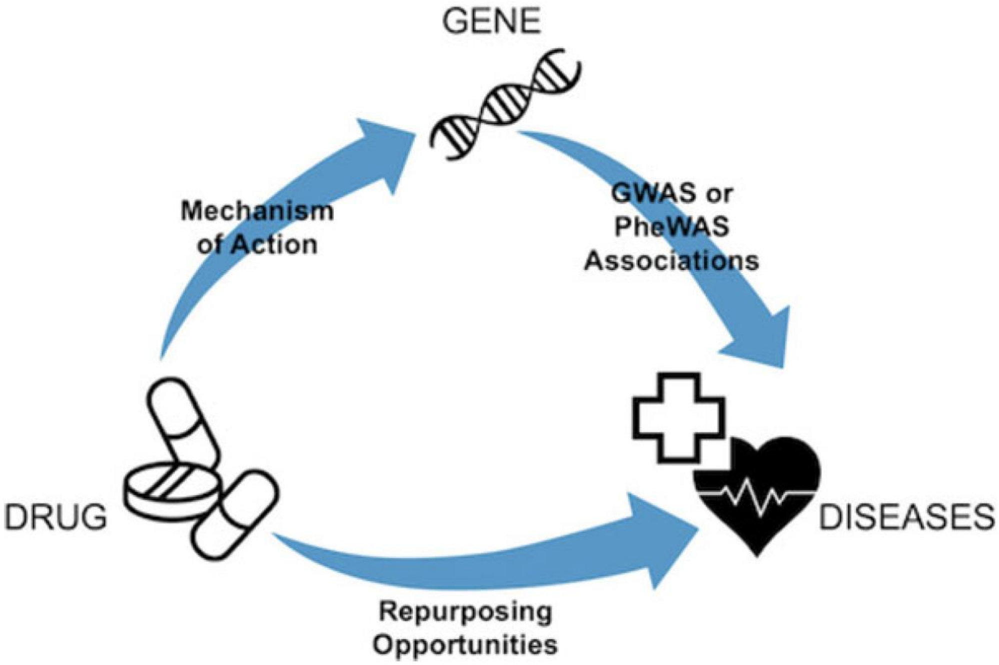


Figure 3. Opportunities for drug repurposing using results of genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). Given a known mechanism of action or genetic target of a currently approved drug, GWAS and PheWAS reveal drug repurposing opportunities through identification of diseases with common genetic associations with the known drug genetic target.

Prior studies have demonstrated success of this approach, for example the use of complement inhibitors for the treatment of age-related macular degeneration (AMD). One of the first GWAS in 2005 found the complement factor H gene to be strongly associated with risk of AMD.(74) At that time, complement inhibitors had been developed for the treatment of sepsis and paroxysmal nocturnal hemoglobinuria.(75-77) This has led to the targeting of factors in the alternative complement pathway in clinical trials with promising findings for reducing the severity of AMD.(78) In addition to validating currently approved RA drug therapies, Okada et al. identified several drugs used for other diseases that target biological genes containing RA risk SNPs and thus proposed these as drug repurposing opportunities.(32) They found that *CDK6* and *CDK4*, targets of three approved drugs for cancer (palbociclin, capridine, and flavopiridol), include RA risk SNPs, suggesting they should be investigated for efficacy in RA as well.

Analogous to the use of GWAS to identify novel drug uses, the ability of PheWAS to identify pleiotropic effects creates opportunities for drug repurposing (Figure 3).(13, 73) As PheWAS can identify diseases that share a common etiology, one can theorize that drugs used to treat one disease may also have efficacy to treat another.(16) A hypothesis-generating study by Rastegar-Mojarad et al. evaluated the potential for drug repurposing by linking current drug-targeted genes in DrugBank to the gene-phenotype associations in the PheWAS catalog.(73) They validated the disease indications for drugs in 127 cases, but also identified 2583 strongly supported potential novel drug-disease associations, available within a cataloged database to the public.(79) There are several factors that can influence the ability for a drug-disease identified in PheWAS and poised for drug repurposing to come to fruition. In particular, methods must be developed to narrow the results to candidate drug-disease pairs that are supported in the literature or by mechanistic knowledge. Rastegar-Mojarad et al. started this approach by cross-referencing all pairs with the clinical trial registry, noting that incorporation of other biomedical databases could also significantly improve prioritization.(73) Recently, Pulley et al. specifically described six genes with pleiotropic effects identified in PheWAS, three of which are currently underway to study repurposing opportunities of drugs with respect to the relevant molecular target.(80)

Millwood et al. in 2016 used PheWAS methods applied to ICD10 codes in the China Kadoorie Biobank to evaluate the potential efficacy of lipoprotein-associated phospholipase A_2 (Lp-PLA_2) inhibitors for the treatment of atherosclerotic disease.(81) Loss-of-function variants in the *PLA2G7* gene is associated with reduced Lp-PLA_2 activity and is relativity common among East Asian populations. Through their PheWAS analysis, they determined there was no association of a loss-of-function variant in *PLA2G7* with improvement in vascular diseases, such as stroke and coronary events, or non-vascular diseases in a Chinese population. They note that these findings correlate with the lack of efficacy in a 2014 randomized controlled trial with the Lp-PLA_2 inhibitor darapladib.(82) Use of PheWAS results such as these in the design of clinical trials could thus help guide study design, saving time and resources.

Challenges of GWAS and PheWAS in drug discovery, drug repurposing, and pharmacogenomics

Statistical power in GWAS and PheWAS is determined by the size of the study cohort, the frequency of the variant, and the effect size of the variant. Both methods are limited by reduced ability to achieve statistical significance given the large number of hypotheses tested.(13) While the number of phenotypes tested in PheWAS is relatively small compared to the number of genotypes tested in GWAS, testing of multiple genotypes against a large set of phenotypes exponentially increases the number of statistical tests, requiring smaller and smaller p-values for statistical significance with Bonferroni correction.

GWAS and PheWAS for evaluation of drug effects is challenged by small sample sizes with a subsequent lack of power to detect small or moderately sized effects.(83) For example, rare but serious adverse events or drug non-responders may be associated with rare variants with clinically relevant effect size, but could be potentially missed in traditional GWAS. Due to the rarity of these events in a population, sample size is often much smaller than typical in GWAS performed for evaluation of disease risk. While the sample size for GWAS in pharmacogenomics studies is typically less than one thousand individuals, GWAS for common diseases often use thousands of subjects with meta-analyses containing even tens of thousands, realized by the pairing of genomic information with EHRs.(84) Non-EHR cohorts often have focused clinical information, lacking drug response trait information. While the EHR can be leveraged to identify drug response traits, the rarity of events necessitates collaboration between biobanks to reach adequate statistical power. Several efforts to encourage data sharing have evolved over the last decade, such as the Electronic Medical Records and Genomics (eMERGE) network, UK Biobank, China Kadoorie Biobank, and Million Veterans Project.

Another challenge facing GWAS and PheWAS is due to the complex architecture of phenotypes with non-Mendelian inheritance patterns. Disease-associated alleles, and thus druggable genes, often have a very small effect on the overall risk of the disease, thus variability in drug effects can also only be partially accounted for by an identified genetic variant.(4) GWAS and PheWAS are designed and powered to detect associations with common genetic variants in a population, with the majority of these variants having small genetic effects. Thus, while an association may be present between a drug effect and genetic variant, many other environmental and genetic factors are also simultaneously contributing to that variation, resulting in a significant proportion of "missing heritability". Further, GWAS and PheWAS results are population-specific, with the majority of large studies being performed in populations of European descent.(85, 86) The extent to which these findings can be translated to other populations is unknown as there are significant differences in linkage disequilibrium and allele frequencies between ancestries.

While GWAS may identify many alleles contributing to disease risk, not all of those alleles or potential gene targets will be disease-causing or able to be modulated for disease treatment.(4) Those genes which harbor causal alleles must be differentiated from the rest in order to narrow the search for potential drug targets. Once a causal allele is identified, it can be difficult to understand the mechanism by which the gene variation contributes to the disease, thus functional studies are required to fully understand the disease risk attributed to the gene and the potential mechanism of a modifying drug. Another factor that has limited the success of GWAS findings from being translated to marketed therapeutics is the long duration, often over 10 years, before a gene target is translated into an approved marketed drug.(19) As previously discussed, drug repurposing is a method to potentially decrease this development time.

The use of GWAS and PheWAS to investigate drug efficacy and adverse events relies upon accurate description of the drug response or adverse event. Although EHRs have greatly eased the ability for researchers to identify phenotypes in a population, accurate drug response and side effect phenotypes remain a challenge to assemble in large cohorts.(83) EHR-based GWAS and PheWAS rely on the ability to readily extract structured data from the medical record. For PheWAS, this is often in the form of billing codes which are unreliably accurate and rarely used for describing drug effects, drug efficacy, and adverse event phenotypes. Thus, manually-curated and validated phenotyping algorithms from the EHR must be developed and implemented. While EHRs have allowed for accrual of and access to clinical information, algorithms are necessary to extract this information from the various parts of the EHR, including clinical documents, laboratory data, nursing records, etc. Development and validation of these algorithms can be time-consuming and require both clinical and technical expertise.

While curation of a single (or few) phenotypes for a GWAS is manageable, this is much more difficult for the thousands of potential phenotypes used in a PheWAS. Phecodes have shown efficacy for PheWAS analyses; however, they do not align precisely with clinical diseases and may not have adequate granularity or specificity for some phenotypes.(14, 87) Phecodes currently use ICD-9-CM codes as their sole source of information. Efforts are underway to map the codes to ICD-10, but more importantly, billing codes alone do not capture all medically-relevant phenotypes. For drug effects, while integration of billing codes with other portions of information from the EHR can refine phenotypes and exposures, a significant limitation in obtaining these well-specified phenotypes from various sources is first the clinical expertise to define the phenotype, followed by the informatics support to extract the information from the EHR.(88; 89) Curated phenotypes have been developed for individual diseases, but there is currently no high-throughput mechanism to produce cases and controls from thousands of detailed phenotype algorithms. New methods are needed to study drug exposures with events at scale while appropriately assessing the timing of both. Currently, drug response phenotypes are best pursued one-phenotype-at-a-time.

As we have previously discussed, drug efficacy can vary significantly in a population. However, accurate ascertainment of drug response as a continuous outcome is difficult. For some phenotypes, such as blood pressure reduction or blood glucose control, multiple measurements may allow for more accurate determination of response; however, for the majority of therapeutics there is not a defined scale for response or adverse effect, nor are these measurements made routinely part of clinical care to enable large GWAS studies in EHR cohorts. The recent study by Wells et al. shows how a phenotype along a continuum, left ventricular function by systolic ejection fraction, can be used as the outcome in a GWAS analysis to determine drug side effects, rather than a dichotomous variable, such as the presence or absence of heart failure.(68) When feasible, GWAS to measure drug-phenotype associations should use phenotypes defined along a continuum to allow improved accuracy of prediction.

Other limitations of EHR-based genetics research are secondary to the current confines of EHRs. Due to the decentralization of EHRs, data within the record itself it may be incomplete due to the various providers and institutions a patient may visit. Also, EHRs are designed for exchange of clinical information and billing purposes, not specifically for research. Thus, inaccuracies can be introduced by clinical uncertainty or billing errors, and the amount of information available can vary greatly. Further efforts to improve EHR data, centralize information, and allow for phenotype curation from EHR data more efficiently and accurately will greatly facilitate advancement in phenotyping studies.

Emerging GWAS and PheWAS-related techniques for pharmacogenomics

While GWAS methods have provided insight into thousands of variants associated with complex traits, the biological mechanisms underlying the associations remain poorly understood. Gene expression is an intermediate between genetic information and phenotypes and can play an important role in drug response. One proposed method to gain information on biological mechanisms and gene expression is through PrediXcan, a technique that estimates the component of gene expression determined by an individual’s genetic profile through use of reference transcriptome data sets and correlates that gene expression with the phenotype of interest.(90) PrediXcan can be likened to a limited PheWAS using imputed gene expression as the PheWAS predictor variable. A major benefit of PrediXcan is its ability to increase power by aggregating the effects of SNPs associated with gene expression. PrediXcan also provides direction of the effect of the genetic variant, for example increased or decreased gene expression. This is significant for drug discovery and repurposing, as the development of therapeutics that downregulate a gene, and thus gene expression, is often easier to attain than development of drugs that upregulate a gene.(90)

In addition to potential opportunities with drug development, PrediXcan can provide insights into drug effects. One recent example in which this has been employed is in evaluation of genetically-determined expression levels association with chemotherapy-induced peripheral neuropathy. Dolan et al. analyzed associations with cisplatin-induced peripheral neuropathy using GWAS and PrediXcan.(91) While no SNPs met genome-wide significance in GWAS, lower expression of *RPRD1B*, which is predicted by twenty SNPs on chromosome 20 and codes for a protein that regulates transcription of genes involved in the cell cycle, was associated with decreased risk for cisplatin-induced peripheral neuropathy in PrediXcan ($p = 0.0089$).(91) These recent findings suggest a promising role for PrediXcan methods in the future.

Techniques such as PrediXcan, which aim to increase the power of GWAS methods, may overcome some of the limitations for GWAS to identify associations with rare variants or small effects. Further, although analysis of GWAS data often uses stringent thresholds for statistical significance, there is likely information that can be gleamed from associations with p-values that fail to meet the $5x10^{-8}$ threshold. Some have proposed analyzing GWAS data using a multiple-locus-based approach, drawing on protein pathway- or domain-based data to develop a candidate gene data set, which can then be integrated with known drug-gene target sets to identify potential drug repurposing opportunities.(72, 92) This has been suggested for a wide range of complex diseases including type 1 and 2 diabetes, bipolar disorder, Crohn’s disease, hypertension, coronary artery disease, and RA.(72, 93)

Although the initial applications of the PheWAS methodology have focused on identification of phenotypes that are associated with single SNPs, recent approaches have involved a search for associations with aggregated genetic information or other phenotypic data.(94) These advances also aim to overcome the power and effect size limitations of traditional PheWAS studies. Use of a set of SNPs as the input for a PheWAS can be one way to increase effect size in PheWAS. The set of SNPs can be used to generate a genetic risk score derived from GWAS data and weighted based upon individual SNP effect size. Krapohl et al. used genetic risk scores of thousands of SNPs derived from GWAS of psychiatric traits to determine associations with phenotypes.(95) They also demonstrate the use of a limited phenome, consisting only of behavioral phenotypes only, which can be used to yield greater power.

Similarly, methods for the joint testing of multiple correlated traits can be performed to increase the power in a PheWAS analysis.(96) As many phenotypes are known to be correlated, the Bonferroni correction often applied to PheWAS is likely overly conservative, resulting in significant associations being missed. Performing the analysis on an *a priori* grouping of correlated traits could increase likelihood of finding associations. Any significant association could then be more closely analyzed individually, decreasing the number of tests performed in a single PheWAS compared to analysis using the entire phenome.(96)

It is the curation of the EHR phenome that enables PheWAS, and the technique is not limited to the study of genetic effects. PheWAS methods can also be used to investigate the association of other factors, such as laboratory parameters or comorbidities, with human traits, an analysis that can be termed a phenotype-only PheWAS. Using this approach, Warner et al. demonstrated that elevated white blood counts (WBC) in an intensive care unit are associated with diagnoses of *Clostridium difficile* infection and bacterial sepsis.(97) This study also takes advantage of the non-binary features of many clinical traits, such as continuous laboratory measurements, to show the varying WBC across the phenome. Limiting dichotomization of these features, which could lead to loss of significant information and ability to find associations, will be important in future PheWAS.

Phenotype-only PheWAS can also be used to describe features associated with a disease process as shown recently in the description of features associated with systemic loxoscelism.(98) In another study, Liao et al. used the predictor in a PheWAS as the presence of autoantibodies among a cohort of patients with RA, and determined a significant association between several different epitopes and comorbidities.(99) A similar approach was used by Doss et al. to define subgroups of RA patients based upon serology for rheumatoid factor, finding that seronegative RA was associated with fibromyalgia and seropositive RA was associated with chronic airway obstruction.(100) In addition to demonstrating the use of non-genetic information for PheWAS analyses, these studies show the ability for PheWAS to identify subtypes within diseases, for example associations with other diseases, severity of disease, variable phenotypic manifestations of disease, or differing response to therapeutics. Outside of clinical phenotypes as predictors in PheWAS, another opportunity for the future is to apply PheWAS to PrediXcan, in which predictors of gene expression can be used to identify traits associated with predicted increased or decreased expression of a gene. Each of these developing techniques have the potential to add insight on subgroups of diseases that respond to medication therapy differently, including patient populations with the development of adverse effects or lack of efficacy.

While the potential for evolution of PheWAS techniques are vast, the goal will remain the same – to improve the ability for PheWAS to identify novel associations by increasing power and improving predictive capacity.

Conclusions

GWAS and PheWAS do not only provide insight into biology of diseases, but also provide opportunities for drug targeting, development, and identification of populations at risk for drug-related adverse events. Further investigations using current and future methods will provide the linkages between disease-gene associations, cellular mechanisms, and therapeutic approaches. GWAS and PheWAS pharmacogenomic studies with larger sample sizes, facilitated by multi-institutional collaboration and consistent phenotyping through utilization of EHRs, can allow future studies to achieve greater power to identify small to moderate genetic effects on drug response. Techniques such as genetic risk scores to analyze all risk genes, including those with small and large effect size in a population, will further facilitate greater accuracy in prediction of response to drug therapy.

Acknowledgments

JR Robinson receives salary and tuition support by the T15 LM007450 training grant from the NIH National Library of Medicine. SL Van Driest is supported by an Innovation in Regulatory Science Award from the Burroughs Wellcome Foundation (1015006) and a Clinical Scientist Development Award from the Doris Duke Foundation, and has been supported by the NIH National Center for Advancing Translational Sciences KL2 TR000446. This work is also supported by NIH funding through R01 LM0106085 and P50 GM115305.

References

1. Scannell JW, Blanckley A, Boldon H, Warrington B. 2012. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews. Drug discovery* 11:191-200

2. Kola I, Landis J. 2004. Can the pharmaceutical industry reduce attrition rates? *Nature reviews. Drug discovery* 3:711-5

3. Arrowsmith J, Miller P. 2013. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature reviews. Drug discovery* 12:569

4. Plenge RM, Scolnick EM, Altshuler D. 2013. Validating therapeutic targets through human genetics. *Nature reviews. Drug discovery* 12:581-94

5. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, et al. 2015. The support of human genetic evidence for approved drug indications. *Nature genetics* 47:856-60

6. Denny JC. 2014. Surveying Recent Themes in Translational Bioinformatics: Big Data in EHRs, Omics for Drugs, and Personal Genomics. *Yearbook of medical informatics* 9:199-205

7. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106:9362-7

8. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, et al. 2017. Rare and low-frequency coding variants alter human adult height. *Nature* 542:186-90

9. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)* 26:1205-10

10. Hebbring SJ. 2014. The challenges, advantages and future of phenome-wide association studies. *Immunology* 141:157-65

11. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology* 31:1102-10

12. Namjou B, Marsolo K, Caroll RJ, Denny JC, Ritchie MD, et al. 2014. Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Frontiers in genetics* 5:401

13. Denny JC, Bastarache L, Roden DM. 2016. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annual review of genomics and human genetics* 17:353-73

14. Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS one* 12:e0175508

15. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. 2013. A PheWAS approach in studying HLA-DRB1*1501. *Genes and immunity* 14:187-91

16. Hebbring SJ, Rastegar-Mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. 2015. Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics (Oxford, England)* 31:1981-7

17. Moore CB, Verma A, Pendergrass S, Verma SS, Johnson DH, et al. 2015. Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. *Open forum infectious diseases* 2:ofu113

18. Hall MA, Verma A, Brown-Gentry KD, Goodloe R, Boston J, et al. 2014. Detection of pleiotropy through a Phenome-wide association study (PheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLoS genetics* 10:e1004678

19. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, et al. 2012. Use of genome-wide association studies for drug repositioning. *Nature biotechnology* 30:317-20

20. Mabuchi H, Haba T, Tatami R, Miyamoto S, Sakai Y, et al. 1981. Effect of an inhibitor of 3-hydroxy-3-methyglutaryl coenzyme A reductase on serum lipoproteins and ubiquinone-10-levels in patients with familial hypercholesterolemia. *N Engl J Med* 305:478-82

21. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. 2008. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics* 40:189-97

22. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R, et al. 2008. Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arteriosclerosis, thrombosis, and vascular biology* 28:2078-84

23. Chasman DI, Posada D, Subrahmanyan L, Cook NR, Stanton VP, Jr., Ridker PM. 2004. Pharmacogenetic study of statin therapy and cholesterol reduction. *Jama* 291:2821-7

24. Krauss RM, Mangravite LM, Smith JD, Medina MW, Wang D, et al. 2008. Variation in the 3-hydroxyl-3-methylglutaryl coenzyme a reductase gene is associated with racial differences in low-density lipoprotein cholesterol response to simvastatin treatment. *Circulation* 117:1537-44

25. Chasman DI, Giulianini F, MacFadyen J, Barratt BJ, Nyberg F, Ridker PM. 2012. Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction: the Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) trial. *Circulation. Cardiovascular genetics* 5:257-64

26. Postmus I, Trompet S, Deshmukh HA, Barnes MR, Li X, et al. 2014. Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nature communications* 5:5068

27. Wei W, Li X, Feng Q, Kubo M, Kullo IJ, et al. LPA variants are associated with residual cardiovascular risk in patients receiving statins. *Proc. American Heart Association Scientific sessions, Anaheim, CA, 2017*

28. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491:119-24

29. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, et al. 2012. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics* 44:1341-8

30. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, et al. 2012. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature genetics* 44:491-501

31. Zhou K, Bellenguez C, Spencer CC, Bennett AJ, Coleman RL, et al. 2011. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nature genetics* 43:117-20

32. Okada Y, Wu D, Trynka G, Raj T, Terao C, et al. 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506:376-81

33. Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, et al. 2003. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature genetics* 34:154-6

34. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. 2005. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nature genetics* 37:161-5

35. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH. 2006. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* 354:1264-72

36. Stein EA, Gipe D, Bergeron J, Gaudet D, Weiss R, et al. 2012. Effect of a monoclonal antibody to PCSK9, REGN727/SAR236553, to reduce low-density lipoprotein cholesterol in patients with heterozygous familial hypercholesterolaemia on stable statin dose with or without ezetimibe therapy: a phase 2 randomised controlled trial. *Lancet* 380:29-36

37. Stein EA, Mellis S, Yancopoulos GD, Stahl N, Logan D, et al. 2012. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* 366:1108-18

38. Roden DM, Wilke RA, Kroemer HK, Stein CM. 2011. Pharmacogenomics: the genetics of variable drug responses. *Circulation* 123:1661-70

39. Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Hortnagel K, et al. 2004. Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* 427:537-41

40. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, et al. 2005. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med* 352:2285-93

41. Lindh JD, Lundgren S, Holm L, Alfredsson L, Rane A. 2005. Several-fold increase in risk of overanticoagulation by CYP2C9 mutations. *Clinical pharmacology and therapeutics* 78:540-50

42. Wadelius M, Chen LY, Downes K, Ghori J, Hunt S, et al. 2005. Common VKORC1 and GGCX polymorphisms associated with warfarin dose. *The pharmacogenomics journal* 5:262-70

43. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, et al. 2008. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 112:1022-7

44. Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, et al. 2009. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS genetics* 5:e1000433

45. Perera MA, Cavallari LH, Limdi NA, Gamazon ER, Konkashbaev A, et al. 2013. Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. *Lancet* 382:790-6

46. Hulot JS, Bura A, Villard E, Azizi M, Remones V, et al. 2006. Cytochrome P450 2C19 loss-of-function polymorphism is a major determinant of clopidogrel responsiveness in healthy subjects. *Blood* 108:2244-7

47. Kim KA, Park PW, Hong SJ, Park JY. 2008. The effect of CYP2C19 polymorphism on the pharmacokinetics and pharmacodynamics of clopidogrel: a possible mechanism for clopidogrel resistance. *Clinical pharmacology and therapeutics* 84:236-42

48. Mega JL, Close SL, Wiviott SD, Shen L, Hockett RD, et al. 2009. Cytochrome p-450 polymorphisms and response to clopidogrel. *N Engl J Med* 360:354-62

49. Simon T, Verstuyft C, Mary-Krause M, Quteineh L, Drouet E, et al. 2009. Genetic determinants of response to clopidogrel and cardiovascular events. *N Engl J Med* 360:363-75

50. Shuldiner AR, O'Connell JR, Bliden KP, Gandhi A, Ryan K, et al. 2009. Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *Jama* 302:849-57

51. Zhong WP, Wu H, Chen JY, Li XX, Lin HM, et al. 2017. Genomewide Association Study Identifies Novel Genetic Loci That Modify Antiplatelet Effects and Pharmacokinetics of Clopidogrel. *Clinical pharmacology and therapeutics* 101:791-802

52. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, et al. 2009. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* 461:399-401

53. Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, et al. 2009. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nature genetics* 41:1100-4

54. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, et al. 2009. Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nature genetics* 41:1105-9

55. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, et al. 2009. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nature genetics* 41:816-9

56. Singer JB, Lewitzky S, Leroy E, Yang F, Zhao X, et al. 2010. A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury. *Nature genetics* 42:711-4

57. Lucena MI, Molokhia M, Shen Y, Urban TJ, Aithal GP, et al. 2011. Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. *Gastroenterology* 141:338-47

58. Kindmark A, Jawaid A, Harbron CG, Barratt BJ, Bengtsson OF, et al. 2008. Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis. *The pharmacogenomics journal* 8:186-95

59. Nicoletti P, Aithal GP, Bjornsson ES, Andrade RJ, Sawle A, et al. 2017. Association of Liver Injury From Specific Drugs, or Groups of Drugs, With Polymorphisms in HLA and Other Genes in a Genome-Wide Association Study. *Gastroenterology* 152:1078-89

60. Chung WH, Hung SI, Hong HS, Hsih MS, Yang LC, et al. 2004. Medical genetics: a marker for Stevens-Johnson syndrome. *Nature* 428:486

61. Mallal S, Nolan D, Witt C, Masel G, Martin AM, et al. 2002. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* 359:727-32

62. Illing PT, Vivian JP, Dudek NL, Kostenko L, Chen Z, et al. 2012. Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature* 486:554-8

63. Ozeki T, Mushiroda T, Yowang A, Takahashi A, Kubo M, et al. 2011. Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. *Human molecular genetics* 20:1034-41

64. McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperaviciute D, et al. 2011. HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *N Engl J Med* 364:1134-43

65. Link E, Parish S, Armitage J, Bowman L, Heath S, et al. 2008. SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N Engl J Med* 359:789-99

66. Mosley JD, Shaffer CM, Van Driest SL, Weeke PE, Wells QS, et al. 2016. A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. *The pharmacogenomics journal* 16:231-7

67. Aminkeng F, Bhavsar AP, Visscher H, Rassekh SR, Li Y, et al. 2015. A coding variant in RARG confers susceptibility to anthracycline-induced cardiotoxicity in childhood cancer. *Nature genetics* 47:1079-84

68. Wells QS, Veatch OJ, Fessel JP, Joon AY, Levinson RT, et al. 2017. Genome-wide association and pathway analysis of left ventricular function after anthracycline exposure in adults. *Pharmacogenet Genomics* 27:247-54

69. Van Driest SL, McGregor TL, Velez Edwards DR, Saville BR, Kitchner TE, et al. 2015. Genome-Wide Association Study of Serum Creatinine Levels during Vancomycin Therapy. *PloS one* 10:e0127791

70. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, et al. 2013. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS computational biology* 9:e1003405

71. Diogo D, Bastarache L, Liao KP, Graham RR, Fulton RS, et al. 2015. TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PloS one* 10:e0122271

72. Grover MP, Ballouz S, Mohanasundaram KA, George RA, Sherman CD, et al. 2014. Identification of novel therapeutics for complex diseases from genome-wide association data. *BMC medical genomics* 7 Suppl 1:S8

73. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. 2015. Opportunities for drug repositioning from phenome-wide association studies. *Nature biotechnology* 33:342-5

74. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385-9

75. Nurnberger W, Gobel U, Stannigel H, Eisele B, Janssen A, Delvos U. 1992. C1-inhibitor concentrate for sepsis-related capillary leak syndrome. *Lancet* 339:990

76. Hack CE, Voerman HJ, Eisele B, Keinecke HO, Nuijens JH, et al. 1992. C1-esterase inhibitor substitution in sepsis. *Lancet* 339:378

77. Hillmen P, Hall C, Marsh JC, Elebute M, Bombara MP, et al. 2004. Effect of eculizumab on hemolysis and transfusion requirements in patients with paroxysmal nocturnal hemoglobinuria. *N Engl J Med* 350:552-9

78. Yaspan BL, Williams DF, Holz FG, Regillo CD, Li Z, et al. 2017. Targeting factor D of the alternative complement pathway reduces geographic atrophy progression secondary to age-related macular degeneration. *Science translational medicine* 9

79. Moosavinasab S, Patterson J, Strouse R, Rastegar-Mojarad M, Regan K, et al. 2016. 'RE:fine drugs': an interactive dashboard to access drug repurposing opportunities. *Database : the journal of biological databases and curation* 2016

80. Pulley JM, Shirey-Rice JK, Lavieri RR, Jerome RN, Zaleski NM, et al. 2017. Accelerating Precision Drug Development and Drug Repurposing by Leveraging Human Genetics. *Assay and drug development technologies* 15:113-9

81. Millwood IY, Bennett DA, Walters RG, Clarke R, Waterworth D, et al. 2016. A phenome-wide association study of a lipoprotein-associated phospholipase A2 loss-of-function variant in 90 000 Chinese adults. *International journal of epidemiology* 45:1588-99

82. White HD, Held C, Stewart R, Tarka E, Brown R, et al. 2014. Darapladib for preventing ischemic events in stable coronary heart disease. *N Engl J Med* 370:1702-11

83. Daly AK. 2010. Genome-wide association studies in pharmacogenomics. *Nature reviews. Genetics* 11:241-6

84. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, et al. 2014. Biobanks and electronic medical records: enabling cost-effective research. *Science translational medicine* 6:234cm3

85. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. 2010. Genome-wide association studies in diverse populations. *Nature reviews. Genetics* 11:356-66

86. Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* 538:161-4

87. Leader JB, Pendergrass SA, Verma A, Carey DJ, Hartzel DN, et al. 2015. Contrasting Association Results between Existing PheWAS Phenotype Definition Methods and Five Validated Electronic Phenotypes. *AMIA Annu Symp Proc* 2015:824-32

88. Wei WQ, Denny JC. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome medicine* 7:41

89. Pathak J, Kho AN, Denny JC. 2013. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 20:e206-11

90. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* 47:1091-8

91. Dolan ME, El Charif O, Wheeler HE, Gamazon ER, Ardeshir-Rouhani-Fard S, et al. 2017. Clinical and Genome-Wide Analysis of Cisplatin-Induced Peripheral Neuropathy in Survivors of Adult-Onset Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*

92. Ballouz S, Liu JY, Oti M, Gaeta B, Fatkin D, et al. 2014. Candidate disease gene prediction using Gentrepid: application to a genome-wide association study on coronary artery disease. *Molecular genetics & genomic medicine* 2:44-57

93. Grover MP, Ballouz S, Mohanasundaram KA, George RA, Goscinski A, et al. 2015. Novel therapeutics for coronary artery disease from genome-wide association study data. *BMC medical genomics* 8 Suppl 2:S1

94. Roden DM. 2017. Phenome-wide association studies: a new method for functional genomics in humans. *The Journal of physiology*

95. Krapohl E, Euesden J, Zabaneh D, Pingault JB, Rimfeld K, et al. 2016. Phenome-wide analysis of genome-wide polygenic scores. *Molecular psychiatry* 21:1188-93

96. Bush WS, Oetjens MT, Crawford DC. 2016. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature reviews. Genetics* 17:129-45

97. Warner JL, Alterovitz G. 2012. Phenome based analysis as a means for discovering context dependent clinical reference ranges. *AMIA Annu Symp Proc* 2012:1441-9

98. Robinson JR, Kennedy VE, Doss Y, Bastarache L, Denny J, Warner JL. 2017. Defining the complex phenotype of severe systemic loxoscelism using a large electronic health record cohort. *PloS one* 12:e0174941

99. Liao KP, Sparks JA, Hejblum BP, Kuo IH, Cui J, et al. 2017. Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. *Arthritis & rheumatology (Hoboken, N.J.)* 69:742-9

100. Doss J, Mo H, Carroll RJ, Crofford LJ, Denny JC. 2017. Phenome-Wide Association Study of Rheumatoid Arthritis Subgroups Identifies Association Between Seronegative Disease and Fibromyalgia. *Arthritis & rheumatology (Hoboken, N.J.)* 69:291-300

CHAPTER IV

Defining the Complex Phenotype of Severe Systemic Loxoscelism Using a Large Electronic Health Record Cohort

Jamie R. Robinson M.D.[1, 2], Vanessa E. Kennedy, M.D.[3], Youssef Doss[4], Lisa Bastarache, B.S.[1], Joshua Denny M.D., M.S.[1, 5], Jeremy L. Warner M.D., M.S.[1, 5]

Affiliations:

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

[2] Department of General Surgery, Vanderbilt University Medical Center, Nashville, TN

[3] Vanderbilt University School of Medicine, Nashville, TN

[4] University School of Nashville, Nashville, TN

[5] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

This manuscript was published in the journal Plos One as follows:

Robinson JR, Kennedy VE, Doss Y, Bastarache L, Denny JC, Warner, JL. Defining the complex phenotype of severe systemic loxoscelism using a large electronic health record cohort. Plos One. 2017; 12(4). PMID: 28422977.

Abstract

Objective: Systemic loxoscelism is a rare illness resulting from the bite of the recluse spider and, in its most severe form, can lead to widespread hemolysis, coagulopathy, and death. We aim to describe the clinical features and outcomes of the largest known cohort of individuals with moderate to severe loxoscelism.

Methods: We performed a retrospective, cross sectional study from January 1, 1995 to December 31, 2015 at a tertiary-care academic medical center, to determine individuals with clinical records consistent with moderate to severe loxoscelism. Age-, sex-, and race-matched controls were compared. Demographics, clinical characteristics, laboratory measures, and outcomes of individuals with loxoscelism are described. Case and control groups were compared with descriptive statistics and phenome-wide association study (PheWAS).

Results: During the time period, 57 individuals were identified as having moderate to severe loxoscelism. Of these, only 33% had an antecedent spider bite documented. Median age of individuals diagnosed with moderate to severe loxoscelism was 14 years old (IQR 9.0-24.0 years). PheWAS confirmed associations of systemic loxoscelism with 29 other phenotypes, e.g., rash, hemolytic anemia, and sepsis. Hemoglobin level dropped an average of 3.1 g/dL over an average of 2.0 days (IQR 2.0-6.0). Lactate dehydrogenase and total bilirubin levels were on average over two times their upper limit of normal values. Eighteen individuals of 32 tested had a positive direct antiglobulin (Coombs') test. Mortality was 3.5% (2/57 individuals).

Conclusion: Systemic loxoscelism is a rare but devastating process with only a minority of patients recalling the toxic exposure; hemolysis reaches a peak at 2 days after admission, with some cases taking more than a week before recovery. In endemic areas, suspicion for systemic loxoscelism should be high in individuals, especially children and younger adults, presenting with a cutaneous ulcer and hemolysis or coagulopathy, even in the absence of a bite exposure history.

Introduction

Systemic loxoscelism is a constitutional illness resulting from the bite of spiders of the genus *Loxosceles*, which is distributed worldwide. In parts of the United States, the *Loxosceles reclusa*, commonly referred to as the brown recluse spider, is endemic. Such bites commonly cause local necrosis, referred to as necrotic arachnoidism.(1) Systemic toxicity may also occur and in its mild form consists of nausea, vomiting, fever, chills, or arthralgia. In its more severe form, brown recluse bites may cause massive hemolysis, hemoglobinuria, acute renal failure, disseminated intravascular coagulation, and rarely death.(2-5) The most significant morbidity in systemic loxoscelism results from hemolysis and coagulopathy.(6, 7) Because hemolysis resulting from loxoscelism is uncommon, there is little known about its clinical manifestations, diagnosis, or outcomes.(8) Use of dapsone, once considered a treatment for systemic loxoscelism, has declined due to the suggestion of increased risk of hemolysis.(9) The underlying pathogenesis of systemic loxoscelism remains incompletely understood, but sphingomyelinase D, a component of the venom toxin, has been shown to have a central role in the process.(10-12) Recent literature suggests it causes both direct toxin-mediated hemolysis and complement-mediated erythrocyte destruction.(8, 11, 13) There is also an indication that hemolysis may be partly immune-mediated, given that a certain proportion of individuals reported in the literature have shown positive direct antiglobulin testing (DAT; Coombs') for surface immunoglobulin G (IgG).(7)

The majority of brown recluse spider bite victims lack systemic symptoms, and severe systemic symptoms are even more rare.(14) In 2014, only 1,330 brown recluse spider bites were reported in the United States; of these, 481 individuals required treatment in a health care facility.(15) The brown recluse spider is endemic to the southeastern and Midwestern United States, and likelihood of envenomation outside of these areas is extremely low.(16, 17) Due to the limited geographic nature of the brown recluse and infrequent occurrence of systemic loxoscelism, there is little published on the clinical features and outcomes of these individuals.

In this study, we describe clinical characteristics and outcomes of the largest known cohort of individuals with systemic loxoscelism to date, leveraging our large de-identified electronic clinical data warehouse. We then performed a phenome-wide association study (PheWAS) of these individuals matched to a control population to identify key differences in ~1800 phenotypes between individuals who develop systemic loxoscelism and those who do not. PheWAS has previously been successfully applied to genomic and laboratory results with high validity to replicate known associations.(18-21) Our goal was to highlight clinical characteristics of this rare and potentially lethal illness, and to potentially uncover previously undocumented phenotypic associations.

Materials and Methods

A retrospective, cross-sectional study was performed to analyze suspected cases of loxoscelism at Vanderbilt University Medical Center (VUMC), a tertiary-care academic medical center, over a 20-year time span. VUMC consists of an adult and children's hospital in the epicenter of the brown recluse geographical range.(17) Data collection was performed using the VUMC Synthetic Derivative (SD), a de-identified version of over 2.4 million patient electronic health records.(22) Dates are shifted at random +/- 365 days for each individual with relative time preservation. We identified all records with shifted dates between January 1st, 1995 and December 31st, 2015 containing any mention of "loxoscelism" in a clinical note, problem list, discharge summary, clinical communication, or letter. A two-person manual review of all flagged records was performed, and discrepancies were adjudicated by a third reviewer. This study was approved and designated as non-human subject research by the Institutional Review Board of Vanderbilt University Medical Center; therefore, consent was not necessary.

Individuals were manually excluded from the study if they lacked evidence of systemic loxoscelism, i.e., absence of fever, chills, abdominal pain, hemolysis or abnormal liver function tests. Individuals with moderate to severe loxoscelism were determined by the presence of a documented diagnosis of hemolysis or disseminated intravascular coagulation, need for blood transfusion, or hemodynamic instability.

Control individuals were extracted to be age-, sex-, and race-matched to the cases with a 50:1 control to case ratio.

Variable data extracted included demographics and clinical parameters, including length of hospital stay, intensive care unit (ICU) admission, need for hemodialysis, and mortality. Race was self-reported and extracted for population comparisons and matching of controls. We determined individuals who received dapsone prior to or during admission to VUMC. The presence of a toxicology consultation and/or operative intervention for the cutaneous lesion, if present, were also documented. Laboratory values obtained included hemoglobin (HGB), lactate dehydrogenase (LDH), total bilirubin, haptoglobin, creatinine, urinalysis, and DAT. We selected for laboratory data 1 week prior to and up 3 weeks after the first instance of either "loxoscelism" in a clinical document or ICD-9-CM billing code of 989.5 (toxic effect of venom). Descriptive statistics were performed, including mean or median with interquartile ranges for continuous variables or frequencies and percentages for categorical variables. Demographic differences between case and control groups were assessed using Chi-square or Fisher exact test, as appropriate. Wilcoxon rank sum test was used to compare ages of individuals with moderate-severe systemic loxoscelism to those with only cutaneous or mild systemic symptoms. PheWAS of cases versus controls was applied. PheWAS codes are aggregations of ICD-9-CM codes, as previously described.(19) All pairwise PheWAS comparisons were conducted using logistic regression with adjustment for age and sex. The minimum number of records to perform a test was 20 individuals in the combined case and control groups, with each individual having at least 2 instances of the PheWAS code. All statistical analyses were performed with R statistical software(23) using the PheWAS package and PheWAS code map version 1.2.(21)

Results

The initial search found 373 possible cases with “loxoscelism” documented within the SD. After manual review and exclusion of subjects due to the lack of systemic loxoscelism, 57 individuals with moderate to severe loxoscelism were included in the final analysis. Of the excluded individuals, 90 were found to have cutaneous-only symptoms consistent with brown recluse spider bite, and 58 individuals had mild symptoms of systemic loxoscelism; the remainder had negation terms for loxoscelism (e.g., “this presentation is *not consistent* with loxoscelism.”). The control cohort consisted of 2,850 individuals.

Demographic Characteristics

Of the individuals identified, 54% were female and the majority Caucasian (37/57 individuals, 65%). A significantly larger portion of those with loxoscelism was African American (26%) compared to the SD population of ever-admitted individuals (15% African American [p = 0.02]).

The ages of those with moderate to severe loxoscelism were highly skewed towards children and young adults with 82.5% (47/57) of subjects under 30 years of age (Figure 1). The median age of included individuals was 14 years old (IQR 9.0 - 24.0 yrs), significantly younger ($p = 2.0 \text{ x}10^{-7}$) than the median age of those identified with only cutaneous or mild systemic symptoms (n = 148, median age 30 years old [IQR 19.0 - 46.0 yrs]). Admitted SD individuals were also on average older than those with loxoscelism (n = 667,990, median age 33 years old [IQR 12.0 - 58.0 yrs]).

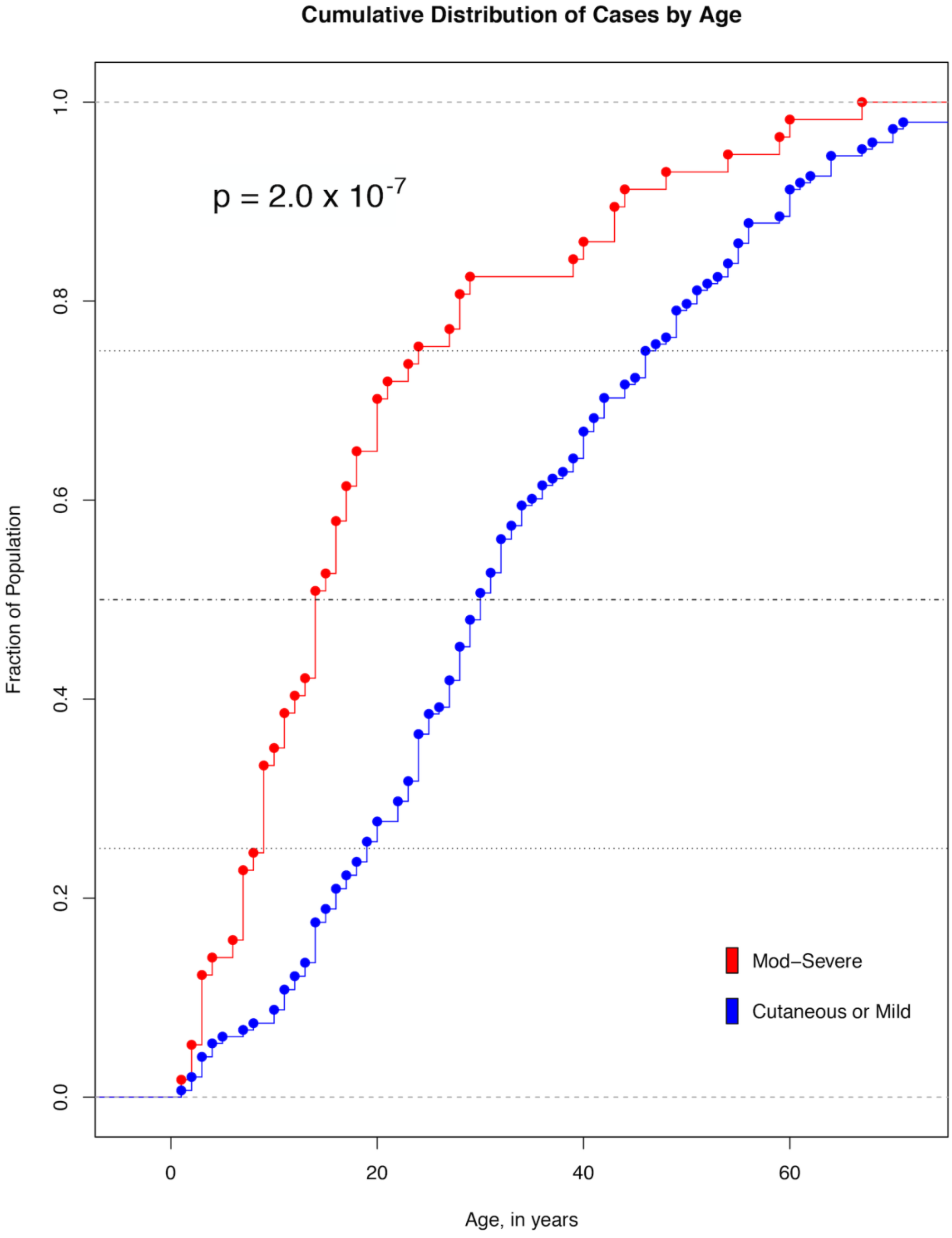


Figure 1. Cumulative Distribution of Cases by Age. Each point represents a case of either moderate-severe (mod-severe, red) loxoscelism or cutaneous/mild loxoscelism (cutaneous or mild, blue). The cumulative proportion of patients at or under a specific age are represented by each line. The median age of individuals with severe loxoscelism (14 years, IQR 9.0 - 24.0 yrs), was significantly lower than the median age (30 years old, IQR 19.0 - 46.0 yrs) of those identified with only cutaneous or mild systemic symptoms (p = 2.0 x 10^{-7}).

All individuals presented with a cutaneous ulcer consistent with a brown recluse spider bite, but many did not recall an antecedent spider bite. According to the documentation present in the SD, only 33% (19 of 57) witnessed the spider and confirmed site of a brown recluse at the time of the envenomation. Ulcer location was most commonly on the upper extremity (27 of 57 individuals, 47%) or lower extremity (10 of 57 individuals, 18%). Further demographics and clinical characteristics are in Table 1.

Table 1. Clinical characteristics and outcomes of systemic loxoscelism.

	Loxoscelism Cohort (n= 57)	**Reference**
Age, median (IQR)	14.0 (9.0-24.0)	-
Sex, No. (%) Male Female Unknown	 26 (46%) 31 (54%) 0 (0%)	-
Race, No. (%) White African American Asian Unknown/ Not Reported Other	 37 (65%) 15 (26%) 1 (2%) 4 (7%) 0 (0%)	-
Witnessed brown recluse spider bite), No. (%)	19 (33%)	-
Bite Location, No. (%) Upper extremity Lower extremity Chest Back Abdomen Head/neck Other	 27 (47%) 10 (18%) 6 (11%) 6 (11%) 4 (7%) 3 (6%) 1 (2 %)	-
Laboratory, median (IQR)		
HGB (g/dL)	10.2 (8.4-11.7)	11.8-16.0
Lowest HGB (g/dL)	8.7 (5.7-10.5)	11.8-16.0
Change in HGB (n=43, g/dL)	-3.1 (-1.8 to -5.6)	-
Average time to lowest HGB (n=43, days)	2.0 (2.0-6.0)	-
LDH (unit/L)	529.0 (265.5-833.5)	<226
Highest LDH per individual (n=37, unit/L)	739.0 (366.0-1344.0)	<226
Total Bilirubin (mg/dL)	2.9 (1.5-5.6)	0.2-1.2
Highest Total Bilirubin (n=47, mg/dL)	4.3 (1.9-7.4)	0.2-1.2
Haptoglobin (mg/dL)	3	16-200
Lowest Haptoglobin (n=23, mg/dL)	25.0 (2.5-129.5)	16-200
Creatinine (mg/dL)	0.8 (0.6-1.0)	0.70-1.50
Highest Creatinine (n=54, mg/dL)	0.9 (0.6-1.2)	0.70-1.50
Direct Antiglobulin positivity (n=32), No. (%)	18 (56%)	Negative
Dapsone treatment, No. (%)	2 (4%)	-
Operative Intervention, No. (%)	5 (9%)	-
Toxicology Consult, No. (%)	46 (80.7%)	-
ICU Admission, No. (%)	28 (49.1%)	-
Length of Hospital Stay, median days (IQR)	4.0 (2.0-5.0)	-
Dialysis, No. (%)	3 (5.3%)	-
Mortality, No. (%)	2 (3.5%)	-

IQR: Interquartile Range; HGB: Hemoglobin; LDH: Lactate Dehydrogenase; ICU: Intensive Care unit

Clinical Parameters

Laboratory parameters during the period with loxoscelism are in Table 1. Average HGB level at presentation was below normal (median 10.2 g/dL, reference 11.8-16.0). Furthermore, the average lowest HGB per case was significantly below normal at 8.7 g/dL (IQR 5.7-10.5). HGB decreased after admission in most individuals but gradually increased back to baseline, with or without supportive transfusion (Figure 2 Parts A-B). Relative decline in HGB was more severe for those admitted or transferred to the ICU during their hospitalization, compared to non-ICU areas. Not including the 12 individuals that arrived with their lowest recorded HGB, HGB level dropped an average of 3.1 g/dL from the first recorded level over an average of 2.0 days (IQR 2.0-6.0). There were 9 individuals with a decline in HGB of over 6 g/dL, of which 6 (67%) occurred 5-8 days after the first recorded value.

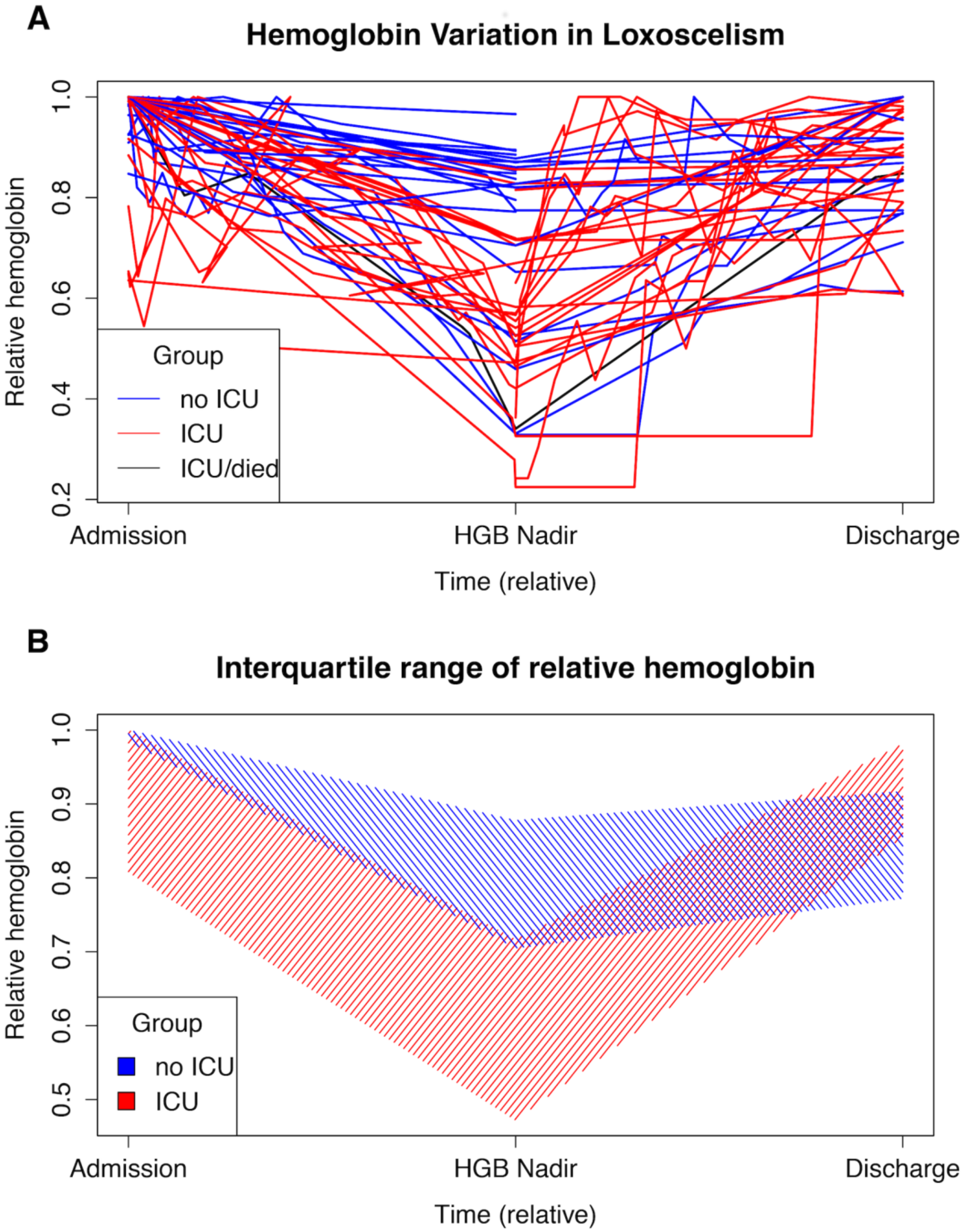


Figure 2. Hemoglobin Fluctuation During Loxoscelism. (A) Each line represents one individual with time and hemoglobin (HGB) graphed relative to the time point at the lowest HGB level (HGB Nadir) and the highest recorded HGB for each individual. Most HGB levels decline after admission and return to baseline. (B) ICU patients have lower interquartile ranges of HGB at presentation and at the HGB nadir, as compared to non-ICU patients. By the time of hospital discharge, the relative HGB level is similar between the two populations.

LDH (529.0 unit/L, reference < 226 unit/L) and total bilirubin (2.9 mg/dL, reference 0.2-1.2 mg/dL) were over two times their normal values. Median haptoglobin and creatinine levels were within normal range. Of the entire cohort, 32 individuals underwent DAT testing and 18 (56%) showed positivity. Of the 18 individuals with a positive DAT test, 9 (50%) were positive for C3 and IgG, 6 (33%) were positive for only IgG and 3 (17%) positive only for C3.

PheWAS for Loxoscelism Phenotype

A PheWAS for phenotypic associations with loxoscelism revealed many strong correlations including rash (p = 1.8 x 10^{-28}), toxic effect of venom (p = 1.5 x 10^{-28}), and hemolytic anemia (p = 2.0 x 10^{-27}), which were also the most frequent phenotypes associated with individuals with loxoscelism (Figure 3 Parts A-B). These were clinical parameters used to assist in confirming the loxoscelism phenotype. Phenotypes that align with intravascular hemolysis, including coagulation defects (p = 2.3 x 10^{-16}), hematuria (p = 3.1 x 10^{-14}), and thrombocytopenia (1.1 x 10^{-6}), were also strongly associated. The PheWAS analysis found strong associations between loxoscelism and superficial cellulitis/abscess (p = 2.1 x 10^{-23}), sepsis (p = 3.1 x 10^{-18}), and septicemia (p = 6.7 x 10^{-12}), all possible correlations with infection. All statistically significant associations are in Table 2.

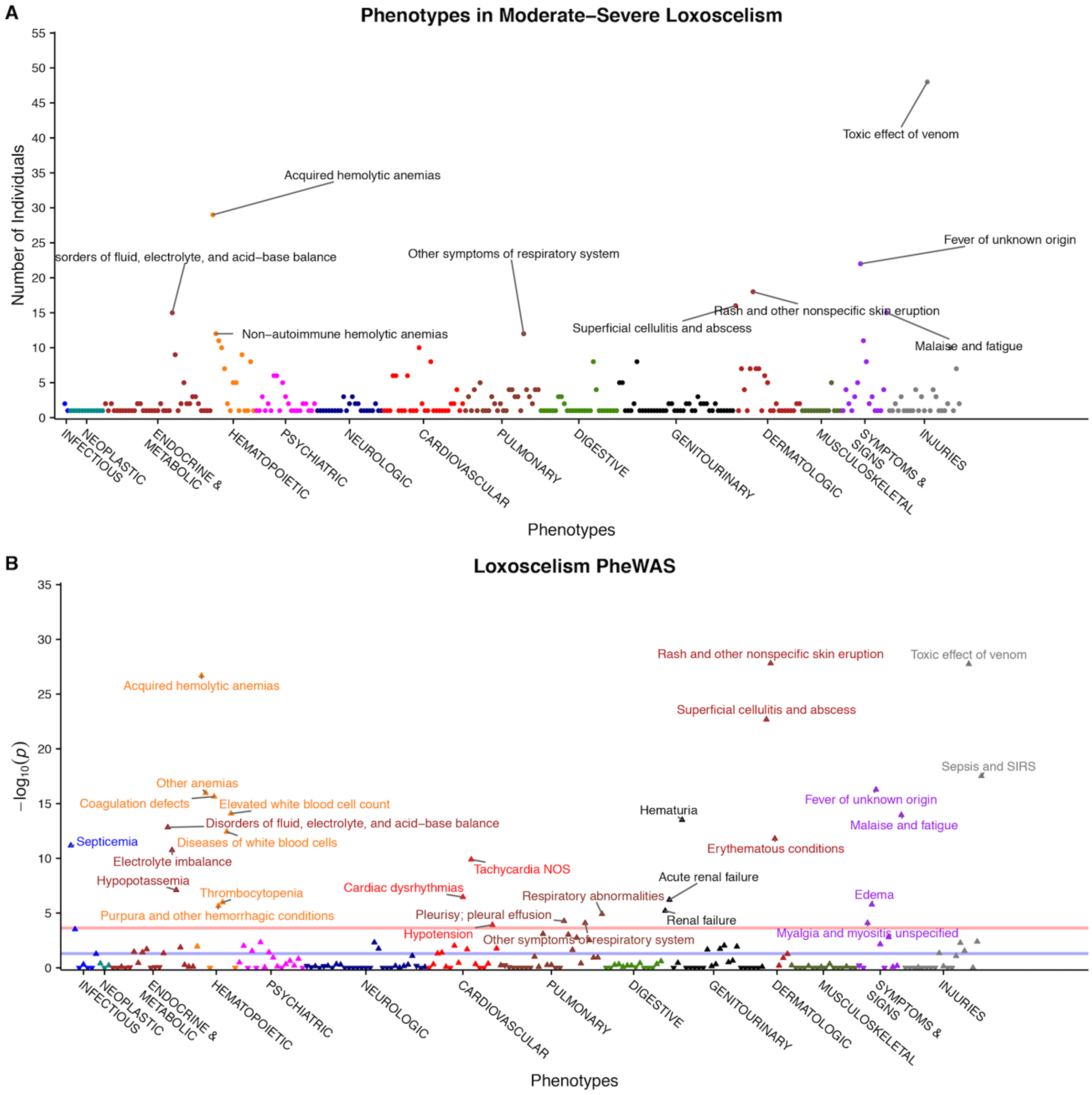


Figure 3. Phenotypes and PheWAS of Individuals with Moderate-Severe Loxoscelism. (A) Manhattan plot representing the number of individuals with moderate to severe loxoscelism with each phenotype. The most frequent phenotypes validated the loxoscelism definition and included the toxic effect of venom, acquired hemolytic anemia, fever of unknown origin, and rash/skin eruption. (B) PheWAS for moderate-severe loxoscelism. The blue line represents significance level without correction (p = 0.05). The red line is representative of the adjusted significance threshold using the Bonferroni correction for multiple comparisons ($p = 1.2 \times 10^{-4}$). 29 phenotypes showed a significant correlation ($p < 1.2 \times 10^{-4}$) with the loxoscelism phenotype when compared to controls.

Table 2. Significant findings for PheWAS of loxoscelism (adjusted significance level, *p* < 1.2 x 10^{-4}).

Clinical Phenotype	**Cases in Loxoscelism only cohort (n = 57), No. (%)**	**Cases in Entire Population, No.**	**Controls in Entire Population, No.**	**OR (95% CI)**	***p*-value**
Rash and other nonspecific skin eruption	18 (32%)	57	2700	54 (27-110)	$1.5x10^{-28}$
Toxic effect of venom	48 (84%)	51	2793	16745 (3714-130031)	$1.8\ x10^{-28}$
Acquired hemolytic anemias	29 (51%)	32	2674	2459 (700-12541)	$2.0\ x10^{-27}$
Superficial cellulitis and abscess	16 (28%)	65	2696	36 (18-75)	$2.1\ x10^{-23}$
Sepsis and SIRS	10 (18%)	22	2869	57 (23-144)	$3.0\ x10^{-18}$
Fever of unknown origin	22 (39%)	207	2484	17 (9-33)	$5.3\ x10^{-17}$
Other anemias	11 (19%)	93	2674	39 (16-94)	$1.1\ x10^{-16}$
Coagulation defects	10 (18%)	33	2790	33 (14-75)	$2.3\ x10^{-16}$
Elevated white blood count	8 (14%)	20	2736	64 (22-186)	$8.1\ x10^{-15}$
Malaise and fatigue	15 (26%)	113	2616	17 (8-35)	$1.1\ x10^{-14}$
Hematuria	8 (14%)	25	2630	36 (14-88)	$3.1\ x10^{-14}$
Disorders of fluid, electrolyte, and acid-base	15 (26%)	128	2632	12 (6-24)	$1.5\ x10^{-12}$
Diseases of white blood cells	9 (16%)	41	2736	23 (9-52)	$3.7\ x10^{-13}$
Erythematous conditions	7 (12%)	21	2765	44 (15-127)	$1.6\ x10^{-12}$
Septicemia	7 (12%)	26	2790	26 (10-65)	$6.7\ x10^{-12}$
Electrolyte imbalance	9 (16%)	57	2632	20 (8-47)	$1.8\ x10^{-11}$
Tachycardia	8 (14%)	44	2651	18 (7-40)	$1.3\ x10^{-10}$
Hypopotassemia	5 (9%)	28	2632	22 (7-65)	$7.9\ x10^{-8}$
Cardiac dysrhythmias	10 (18%)	118	2651	7 (3-15)	$3.4\ x10^{-7}$
Acute renal failure	5 (9%)	27	2812	14 (5-39)	$6.1\ x10^{-7}$
Thrombocytopenia	5 (9%)	35	2790	12 (4-32)	$1.1\ x10^{-6}$
Edema	5 (9%)	31	2825	13 (4-35)	$1.7\ x10^{-6}$
Purpura and other hemorrhagic conditions	5 (9%)	38	2790	11 (4-28)	$2.2\ x10^{-6}$
Renal failure	5 (9%)	35	2812	11 (3-29)	$6.1\ x10^{-6}$
Respiratory abnormalities	4 (7%)	23	2838	12 (3-34)	$1.2\ x10^{-5}$
Pleurisy, pleural effusion	4 (7%)	31	2735	11 (3-31)	$5.3\ x10^{-5}$
Myalgia and myositis unspecified	4 (7%)	32	2839	12 (3-37)	$8.1\ x10^{-5}$
Other symptoms of respiratory system	12 (21%)	229	2441	4 (2-8)	$8.2\ x10^{-5}$
Hypotension	4 (7%)	31	2834	9 (3-26)	$1.2\ x10^{-4}$

OR: Odds Ratio; SIRS: Systemic inflammatory response syndrome

Treatment and Outcomes

Clinical outcomes for individuals are reported in Table 1. All individuals were hospitalized except for one who died in the Emergency Department. The median length of hospital stay was 4.0 days (range, 1-28 days). Of the 57 subjects, 46 (80.7%) had a toxicology consult service assisting in diagnosis and management. Only 2 individuals (4%) received dapsone either prior to (per report) or during their hospital admission. Approximately half of individuals (49.1%) required initial admission or transfer to an ICU due to their critical condition and need for more intensive monitoring. Acute renal injury with a 2-fold increase in the creatinine level occurred in 6 individuals (10.5%). The increase in creatinine among these 6 individuals ranged from 0.7 - 6.9 mg/dL. Few individuals (3; 5.3%) required dialysis due to severe acute renal failure.

Only 2 individuals died during the 20-year study period from loxoscelism. One individual was a previously healthy 54-year-old man who developed severe hemolysis 5 days after a witnessed spider bite. He proceeded to multi-system organ failure with hemodynamic instability, renal and respiratory failure with ultimate cardiac arrest. The other individual was a previously healthy 3-year-old girl who developed signs of systemic loxoscelism within 6 hours of witnessed spider envenomation. She progressed to significant hematuria, anemia, thrombocytopenia, disseminated intravascular coagulopathy and shock within 19 hours of the bite resulting in death.(4)

Discussion

This represents, to our knowledge, the largest cohort analyzed with moderate to severe loxoscelism to date. In contrast to prior published data consisting only of individuals with life-threatening hemolysis from severe loxoscelism(8), our cohort includes a wider phenotypic range consisting of all individuals with evidence of hemolysis. The large majority (70%) of individuals with systemic loxoscelism in our cohort were under 20 years of age. According to the 2014 annual report of the American Association of Poison Control Centers' National Poison Data System, the majority (63%) of individuals who present in the United States with a brown recluse spider bite are 20 years or older. This suggests that although adults suffer from envenomation from brown recluse spiders more frequently (which was also seen in our data), children are subject to a much more severe reaction. Prior literature shows more frequent case reports of systemic loxoscelism occurring in pediatric individuals(4, 24-26) and this large retrospective review corroborates that children are at greater risk for systemic loxoscelism.

Hemolysis can occur in severe cases of loxoscelism(3), but may present differently than typical intravascular hemolysis. HGB is the most direct indicator of clinical severity in hemolytic disease(27) and its level can become extremely low (< 6 g/dL) in severe forms of loxoscelism, as seen in 15 (26%) of the individuals in our cohort. LDH is also known to be elevated during states of intravascular hemolysis(27) which was also demonstrated in this cohort with loxoscelism, whose average highest LDH was 3 times the upper limit of normal. Hyperbilirubinemia is also seen during hemolysis and can rise to > 4 mg/dL in severe acute hemolysis(27, 28) as was seen in our cohort of individuals. Haptoglobin is known to be decreased in periods of hemolysis(27); however, our cohort only had 11 (19%) individuals with a haptoglobin less than normal and an average haptoglobin low within the normal range. Sterile tissue injury or infection can initiate a local inflammatory response that mobilizes a systemic acute phase reaction, resulting in the induction of genes encoding the acute phase plasma proteins, including haptoglobin.(29) Therefore, a state of inflammation from the brown recluse spider bite in many of the individuals may result in difficulty in interpretation of the haptoglobin level.(30)

In our review, it is important to note that the disease rarely progressed to renal failure. Only 6 individuals had a significant increase in creatinine level, 5 of which were identified with the PheWAS to have renal failure as well. Only 3 of these patients required dialysis. Furthermore, renal failure did not correspond with mortality in our series as the individuals who progressed to death did so in such a quick and extreme fashion that dialysis was not undertaken. Treatment of systemic loxoscelism is mainly supportive.(9) Steroids have been used to prevent kidney failure and hemolysis, but their efficacy is subject to debate.(9, 31) Other treatments that may be considered are dapsone, hyperbaric oxygen, and surgical excision.(9) Adverse side effects have been attributed to dapsone including hemolysis, anemia, and hyperbilirubinemia.(9) Only two individuals in our cohort received treatment with dapsone, both for short time periods. Although not statistically significant, one individual who received dapsone was an adult who developed severe hemolysis and subsequently died, whereas the other was a teenage male who suffered severe hemolysis that resolved. Omitting these individuals who received dapsone, the remainder of the hemolysis within our cohort cannot be attributed to treatment patterns of dapsone use.

Literature on rates of DAT positivity in systemic loxoscelism is scant; several small case series have reported high rates of positive DAT.(7, 32) Our large case series confirms DAT is positive in more than 50% of severely affected individuals, although this does not necessarily correlate with severity of disease. These prior reports demonstrated both IgG and C3 on the red blood cell surface, as is seen in a portion of our cohort with loxoscelism. Our results support the hypothesis that the anemia in loxoscelism can result from direct toxin-mediated erythrocyte damage, complement-mediated immune destruction, or both. Autoimmune hemolytic anemia(33) as well as other immune-mediated illnesses(34) are known to be associated with Human Leukocyte Antigen (HLA) type. HLA-DQ6 has shown in a small case series to have a negative association with a positive DAT result in individuals with hemolysis(33); whether our cohort is enriched for this genotype is unknown.

As with most PheWAS analyses based on billing codes, all statistically significant phenotypes were of increased prevalence in the case population. We found that the PheWAS analysis recapitulated the described phenotype for systemic loxoscelism, and also suggested additional phenotypes of concern. Through PheWAS, some significant phenotypes were directly related to the defined definition of loxoscelism, whereas other phenotypes were likely due to secondary effects or potentially factors increasing risk for the disease process. Phenotypes of loxoscelism strongly replicated include toxic effect of venom (OR 16745, 95% CI 3714-130031) and hemolytic anemia (OR 2459, 95% CI 700-12541). The phenotype “toxic effect of venom” is mapped to a single ICD-9 code, 989.5, which applies to bites of venomous snakes, lizards, ticks, and spiders. We also found that the PheWAS analysis reconfirmed other known clinical findings in loxoscelism that were not within our loxoscelism case definition. These included rash or other nonspecific skin eruption (OR 54, 95% CI 27-110), hematuria (OR 36, 95% CI 14-88), coagulation defects (OR 33, 95% CI 14-75), malaise and fatigue (OR 17, 95% CI 8-35), and fever of unknown origin (OR 17, 95% CI 9-33).

Although hematuria was significant in PheWAS, hemoglobinuria was not. Hematuria and hemoglobinuria are both known findings in systemic loxoscelism, with hematuria occurring almost invariably in severe disease.(35-39) Urine dipstick analysis for blood is typically the initial screening test for hematuria or hemoglobinuria. In subsequent testing, the presence of a red blood cells on microscopic urinalysis are indicative of hematuria. However, if the blood is detectable on a dipstick with no or very few microscopically visible RBCs, hemoglobinuria or myoglobinuria is suggested.(40) It is possible that misclassification of these billing diagnosis codes occurred, resulting in a stronger phenotypic association with hematuria. We did note in our cohort that hematuria, sometimes gross hematuria, was a present in patients with very severe forms of loxoscelism.

Several phenotypes more likely to be secondary effects of loxoscelism were also found to be significant in the PheWAS analysis. In particular, there appears to be a strong signal for septicemia (OR 26, 95% CI 10-63), sepsis and systemic inflammatory response syndrome (OR 56, 95% CI 22-140) – suggesting that bacterial superinfection of the local wound site and/or iatrogenic infections are an important consideration in this population. There is also a strong signal for cardiac dysrhythmias (OR 7, 95% CI 3-14) and electrolyte imbalance (OR 20, 95% CI 8-46), suggesting that the systemic loxoscelism process or secondary effects (massive fluid resuscitation, renal failure, etc.) lead to significant issues requiring intensive medical management. With the well-defined phenotype captured by PheWAS analysis, it is possible to construct phenotype risk scores that could capture individuals that may not have been formally diagnosed; this remains the subject of future work.(41)

There are several limitations to our study. This study was undertaken at a single institution and may not be generalizable to others, especially institutions within other countries in North and South

America where antivenom is an option for medical treatment. Although the use and efficacy of antivenom is controversial, with studies indicating potentially limited capacity of the medication to neutralize the systemic effects of loxoscelism due to delayed presentation of illness, further research in this area is needed.(42) Our case identification relied on the use of the word "loxoscelism" in clinical notes; this word is uncommon in the medical jargon and could be hypothetically misspelled. However, a preliminary investigation using the keywords "brown recluse" resulted in a large false positive rate. Another limitation is that many of the cases were referred to VUMC for tertiary-level care and had very little preceding or subsequent history in the SD, limiting our ability to evaluate for long term sequelae in sufferers of systemic loxoscelism. Most importantly, our retrospective chart review was limited by the fact that systemic loxoscelism is a clinical diagnosis that is made upon the presence of systemic symptoms, a cutaneous lesion consistent with a brown recluse spider bite, and clinical presentation and history. There is no confirmatory test for its diagnosis. We believe our criteria to require laboratory markers of hemolysis and documented diagnosis of loxoscelism was the most accurate approach to determining the true prevalence of loxoscelism within our institution. Furthermore, review was performed by physicians to ensure accuracy. More stringent criteria, such as requirement that the spider responsible for envenomation is captured and confirmed to be a brown recluse, would likely lead to a profound underestimation of the actual occurrence of systemic loxoscelism in endemic areas. Lastly, as the severity of loxoscelism occurs along a spectrum without a clear definition of what determines severe disease, we did not distinguish between moderate and severe loxoscelism, including in our cohort all individuals with systemic signs of hemolysis or disseminated intravascular coagulation, need for blood transfusion, or hemodynamic instability. This allowed for greater inclusion of individuals affected by the disease; however, it also led to increased variability in our phenotype.

In conclusion, systemic loxoscelism is a rare occurrence, but within a region endemic to brown recluse spiders, multiple individuals present yearly with moderate to severe loxoscelism. A portion of these individuals develops moderate to severe hemolysis. Although children and possibly African

Americans appear to be at increased risk, it remains unclear what specific risk factors correlate with disease severity.

References

1. Sams HH, Dunnick CA, Smith ML, King LE, Jr. 2001. Necrotic arachnidism. *J Am Acad Dermatol* 44:561-73; quiz 73-6

2. Futrell JM. 1992. Loxoscelism. *Am J Med Sci* 304:261-7

3. Murray LM, Seger DL. 1994. Hemolytic anemia following a presumptive brown recluse spider bite. *J Toxicol Clin Toxicol* 32:451-6

4. Rosen JL, Dumitru JK, Langley EW, Meade Olivier CA. 2012. Emergency department death from systemic loxoscelism. *Ann Emerg Med* 60:439-41

5. Nance WE. 1961. Hemolytic anemia of necrotic arachnidism. *Am J Med* 31:801-7

6. Williams ST, Khare VK, Johnston GA, Blackall DP. 1995. Severe intravascular hemolysis associated with brown recluse spider envenomation. A report of two cases and review of the literature. *Am J Clin Pathol* 104:463-7

7. McDade J, Aygun B, Ware RE. 2010. Brown recluse spider (Loxosceles reclusa) envenomation leading to acute hemolytic anemia in six adolescents. *The Journal of pediatrics* 156:155-7

8. Gehrie EA, Nian H, Young PP. 2013. Brown Recluse spider bite mediated hemolysis: clinical features, a possible role for complement inhibitor therapy, and reduced RBC surface glycophorin A as a potential biomarker of venom exposure. *PloS one* 8:e76558

9. Andersen RJ, Campoli J, Johar SK, Schumacher KA, Allison EJ, Jr. 2011. Suspected brown recluse envenomation: a case report and review of different treatment modalities. *The Journal of emergency medicine* 41:e31-7

10. de Oliveira KC, Goncalves de Andrade RM, Piazza RM, Ferreira JM, Jr., van den Berg CW, Tambourgi DV. 2005. Variations in Loxosceles spider venom composition and toxicity contribute to the severity of envenomation. *Toxicon* 45:421-9

11. Tambourgi DV, Pedrosa MF, de Andrade RM, Billington SJ, Griffiths M, van den Berg CW. 2007. Sphingomyelinases D induce direct association of C1q to the erythrocyte membrane causing complement mediated autologous haemolysis. *Mol Immunol* 44:576-82

12. Gremski LH, Trevisan-Silva D, Ferrer VP, Matsubara FH, Meissner GO, et al. 2014. Recent advances in the understanding of brown spider venoms: From the biology of spiders to the molecular mechanisms of toxins. *Toxicon : official journal of the International Society on Toxinology* 83:91-120

13. Tambourgi DV, Goncalves-de-Andrade RM, van den Berg CW. 2010. Loxoscelism: From basic research to the proposal of new therapies. *Toxicon : official journal of the International Society on Toxinology* 56:1113-9

14. Wright SW, Wrenn KD, Murray L, Seger D. 1997. Clinical presentation and outcome of brown recluse spider bite. *Ann Emerg Med* 30:28-32

15. Mowry JB, Spyker DA, Brooks DE, McMillan N, Schauben JL. 2015. 2014 Annual Report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 32nd Annual Report. *Clinical toxicology (Philadelphia, Pa.)* 53:962-1147

16. Swanson DL, Vetter RS. 2005. Bites of brown recluse spiders and suspected necrotic arachnidism. *N Engl J Med* 352:700-7

17. Vetter RS. 2009. The distribution of brown recluse spiders in the southeastern quadrant of the United States in relation to loxoscelism diagnoses. *Southern medical journal* 102:518-22

18. Bush WS, Oetjens MT, Crawford DC. 2016. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature reviews. Genetics* 17:129-45

19. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26:1205-10

20. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology* 31:1102-10

21. Carroll RJ, Bastarache L, Denny JC. 2014. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 30:2375-6

22. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, et al. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical pharmacology and therapeutics* 84:362-9

23. Team RC. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

24. Said A, Hmiel P, Goldsmith M, Dietzen D, Hartman ME. 2014. Successful use of plasma exchange for profound hemolysis in a child with loxoscelism. *Pediatrics* 134:e1464-7

25. Levin C, Bonstein L, Lauterbach R, Mader R, Rozemman D, Koren A. 2014. Immune-mediated mechanism for thrombocytopenia after Loxosceles spider bite. *Pediatric blood & cancer* 61:1466-8

26. Lane L, McCoppin HH, Dyer J. 2011. Acute generalized exanthematous pustulosis and Coombs-positive hemolytic anemia in a child following Loxosceles reclusa envenomation. *Pediatric dermatology* 28:685-8

27. Barcellini W, Fattizzo B. 2015. Clinical Applications of Hemolytic Markers in the Differential Diagnosis and Management of Hemolytic Anemia. *Disease markers* 2015:635670

28. Berlin NI, Berk PD. 1981. Quantitative aspects of bilirubin metabolism for hematologists. *Blood* 57:983-99

29. Wang Y, Kinzie E, Berger FG, Lim SK, Baumann H. 2001. Haptoglobin, an inflammation-inducible plasma protein. *Redox report : communications in free radical research* 6:379-85

30. Shih AW, McFarlane A, Verhovsek M. 2014. Haptoglobin testing in hemolysis: measurement and interpretation. *American journal of hematology* 89:443-7

31. Forks TP. 2000. Brown recluse spider bites. *J Am Board Fam Pract* 13:415-23

32. Lane DR, Youse JS. 2004. Coombs-positive hemolytic anemia secondary to brown recluse spider bite: a review of the literature and discussion of treatment. *Cutis* 74:341-7

33. Wang-Rodriguez J, Rearden A. 1996. Reduced frequency of HLA-DQ6 in individuals with a positive direct antiglobulin test. *Transfusion* 36:979-84

34. Gough SC, Simmonds MJ. 2007. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Current genomics* 8:453-65

35. Schenone H, Saavedra T, Rojas A, Villarroel F. 1989. [Loxoscelism in Chile. Epidemiologic, clinical and experimental studies]. *Revista do Instituto de Medicina Tropical de Sao Paulo* 31:403-15

36. Rios JC, Perez M, Sanchez P, Bettini M, Mieres JJ, Paris E. 2007. [Prevalence and epidemiology of Loxosceles laeta bite. Analysis of consultations to a poison control center]. *Revista medica de Chile* 135:1160-5

37. Zambrano A, Gonzalez J, Callejas G. 2005. [Severe loxoscelism with lethal outcome. Report of one case]. *Revista medica de Chile* 133:219-23

38. Barretto OC, Satake M, Nonoyama K, Cardoso JL. 2003. The calcium-dependent protease of Loxosceles gaucho venom acts preferentially upon red cell band 3 transmembrane protein. *Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas* 36:309-13

39. Schenone H, Prats F. 1961. Arachnidism by Loxosceles laeta. Report of 40 cases of necrotic arachnidism. *Arch Dermatol* 83:139-42

40. Veerreddy P. 2013. Hemoglobinuria misidentified as hematuria: review of discolored urine and paroxysmal nocturnal hemoglobinuria. *Clinical medicine insights. Blood disorders* 6:7-17

41. Bastarache L, Mosely J, Edwards T, Carroll R, Mo H, et al. Complex diseases are associated with variation in Mendelian genes: A phenome-wide study using Human Phenotype Ontology and a population genotyped on the Exome BeadChip. *Proc. American Society of Human Genetics*, *Baltimore, MD, 2015*:

42. Pauli I, Puka J, Gubert IC, Minozzo JC. 2006. The efficacy of antivenom in loxoscelism treatment. *Toxicon* 48:123-37

CHAPTER V

Association of Genetic Risk for Obesity with Postoperative Complications Using Mendelian Randomization

Jamie R. Robinson, M.D.[1, 2], Robert J. Carroll, Ph.D.[1], Lisa Bastarache, B.S.[1], Qingxia Chen, Ph.D. [3], Zongyang Mou, B.S.[1], Wei-Qi Wei, M.D., Ph.D.[1], John J. Connolly, Ph.D.[4], Frank Mentch, Ph.D.[4], Patrick Sleiman, Ph.D.[4], Paul K. Crane, M.D., MPH[5], Scott J. Hebbring, Ph.D.[6], Ian B. Stanaway, Ph.D.[7], David R. Crosslin, Ph.D.[7], Adam S. Gordon, Ph.D.[8], Elisabeth A. Rosenthal, Ph.D.[8], David Carrell, Ph.D.[9], M. Geoffrey Hayes, Ph.D.[10], Wei Wei, Ph.D.[11], Lynn Petukhova, Ph.D.[12], Bahram Namjou, M.D.[13], Ge Zhang, M.D., Ph.D.[14], Maya S. Safarova, M.D., Ph.D. [15], Nephi A. Walton, M.D., M.S.[16], Christopher Still, D.O.[16], Erwin P. Bottinger, M.D.[17], Ruth J. F. Loos, Ph.D.[17], Shawn N. Murphy, M.D., Ph.D.[18], Gretchen P. Jackson, M.D., Ph.D., FACS [1, 19], Iftikhar J. Kullo, M.D.[15], Hakon Hakonarson, M.D., Ph.D.[4], Gail P. Jarvik, M.D., Ph.D.[8], Eric B. Larson, M.D., MPH[9], Chunhua Weng, Ph.D.[20], Dan M. Roden, M.D.[1, 21, 22], Joshua C. Denny, M.D., M.S.[1, 21]

Affiliations:

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[2] Department of Surgery, Vanderbilt University Medical Center, Nashville, TN, USA

[3] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[4] The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

[5] Department of Medicine, University of Washington, Seattle, WA, USA

[6] Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA.

[7] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA

[8] Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA

[9] Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

[10] Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[11] University of Pittsburgh Medical Center, Pittsburgh, PA, USA

[12] Departments of Dermatology and Epidemiology, Columbia University, New York, NY, USA

[13] Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[14] Division of Human Genetics, Cincinnati Children's Hospital Medical Center, and the Center for Prevention of Preterm Birth, Perinatal Institute, Cincinnati Children's Hospital Medical Center and March of Dimes Prematurity Research Center Ohio Collaborative, and Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

[15] Department of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA

[16] Department of Biomedical and Translational Informatics, Geisinger Health System, Danville, PA, USA

[17] The Charles Bronfman Institute for Personalized Medicine at Mount Sinai, The Mindich Child Health and Development Institute, New York, NY, USA

[18] Department of Neurology, Partners Healthcare, Boston, MA, USA

[19] Department of Pediatric Surgery, Vanderbilt University Medical Center, Nashville, TN, USA

[20] Department of Biomedical Informatics, Columbia University, New York, NY, US

[21] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[22] Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA

This manuscript was published in the journal World Journal of Surgery as follows:

Robinson JR, Carroll RJ, Bastarache L, Chen Q, Mou Z, Wei WQ, Connolly J, Mentch F, Sleiman P, Crane PK, Hebbring SJ, Stanaway IB, Crosslin DR, Gordon AS, Rosenthal EA, Carrell D, Hayes MG, Wei W, Petukhova L, Namjou B, Zhang G, Safarova MS, Walton NA, Still C, Bottinger EP, Loos RJF, Murphy SN, Jackson GP, Kullo IJ, Hakonarson H, Jarvik GP, Larson EB, Weng C, Roden DM, Denny

JC. Association of Genetic Risk for Obesity with Postoperative Complications Using Mendelian Randomization. *World J Surg.* 2019 Oct. PMID 31605180.

Abstract

Background: The extent to which obesity and genetics determine post-operative complications is incompletely understood.

Methods: We performed a retrospective study using two population cohorts with electronic health record (EHR) data. The first included 736,726 adults with body mass index (BMI) recorded between 1990-2017 at Vanderbilt University Medical Center. The second cohort consisted of 65,174 individuals from 12 institutions contributing EHR and genome-wide genotyping data to the Electronic Medical Records & Genomics (eMERGE) Network. Pairwise logistic regression analyses were used to measure the association of BMI categories with postoperative complications derived from International Classification of Disease-9 codes, including postoperative infection, incisional hernia, and intestinal obstruction. A genetic risk score (GRS) was constructed from 97 obesity-risk single nucleotide polymorphisms for a Mendelian randomization study to determine the association of genetic risk for obesity on postoperative complications. Logistic regression analyses were adjusted for sex, age, site, and race/principal components.

Results: Individuals with overweight or obese BMI ($\geq$25 kg/m^2) had increased risk for incisional hernia (Odds ratio [OR] 1.7-5.5, $p<3.1\text{x}10^{-20}$), and people with obesity (BMI$\geq$30 kg/m^2) had increased risk for postoperative infection (OR 1.2-2.3, $p<2.5\text{x}10^{-5}$). In the eMERGE cohort, genetically-predicted BMI was associated with incisional hernia (OR 2.1 [95% CI 1.8-2.5], $p=1.4\text{x}10^{-6}$) and postoperative infection (OR 1.6 [95% CI 1.4-1.9], $p=3.1\text{x}10^{-6}$). Association findings were similar after limitation of the cohorts to those who underwent abdominal procedures.

Conclusions: Clinical and Mendelian randomization studies suggest that obesity, as measured by BMI, is associated with the development of postoperative incisional hernia and infection.

Introduction

Obesity, defined as a body-mass index (BMI) of 30.0 kg/m^2 or greater, is known to be a strong predictor of cardiovascular morbidity and mortality.(1–3) Over two-thirds of the adult population in the United States have an overweight or obese BMI,(4) and there is significant burden of obesity on healthcare worldwide.(5, 6) In addition to the known cardiovascular morbidity associated with obesity, it is generally regarded that obesity is a risk factor for increased postoperative complications. This risk has growing significance in surgery, as the obesity epidemic has resulted in a rising prevalence of obesity-related diseases that require operative intervention, thus increasing the number of patients with obesity undergoing surgery.(7) Bariatric surgery has also become increasingly common and safe to perform with very low reported immediate post-operative complications.(8, 9) However, prior cohort studies have suggested an increased incidence of surgical site infections in individuals with obesity undergoing non-bariatric procedures.(10–23) The majority of these studies consist of cohorts undergoing a limited set of procedures such as vascular surgeries,(12, 13) oncologic resections,(14) gynecologic procedures,(15, 16) or colorectal resections.(17) Therefore, we aim to determine the influence obesity has on postoperative outcomes and if genetic risk for obesity impacts long-term surgical complications. This information can provide surgeons with more definitive data on a patient's operative risk stratification.

Mendelian randomization (MR) is a method that uses single or sets of genetic variants associated with a phenotype of interest as an instrumental variable for association studies.(24) Prior studies have used MR to determine the association of obesity-risk single nucleotide polymorphisms (SNPs) with medical conditions such as ischemic heart disease,(25, 26) hypertension,(26) type 2 diabetes,(26) symptomatic cholelithiasis,(27) deep venous thrombosis,(28) and others.(29–37) However, prior studies have not investigated the association of obesity-risk SNPs with postoperative outcomes.

We leveraged a large electronic health record (EHR) population to identify specific postoperative complications associated with BMI. In a second cohort, we then used MR for obesity by estimating BMI-risk using 97 SNPs known to strongly correlate with BMI to investigate the relationship between genetic risk for obesity and postoperative complications.(38)

Methods

Vanderbilt Cohort

We conducted a retrospective study of all adult (≥18 years of age) individuals using the Vanderbilt University Medical Center (VUMC) Synthetic Derivative, a de-identified version of over 2.4 million VUMC patient health records.(39, 40) Inclusion criteria were at least one documented BMI, calculated as weight in kilograms divided by height in meters squared (kg/m^2), where both weight and height were measured at a single encounter. The study protocol was designated as non-human subject research by the Institutional Review Board at VUMC.

All measured BMI values were extracted for each individual, with BMI data obtained during pregnancy or with clinically implausible values (less than 10 kg/m^2 or greater than 70 kg/m^2) excluded. Each individual was classified by his or her median BMI into one of 6 BMI categories, as defined by the World Health Organization (WHO), including underweight (<18.5 kg/m^2), normal (18.5-24.9 kg/m^2), overweight (25.0-29.9 kg/m^2), and obesity class 1 (30.0-34.9 kg/m^2), class 2 (35.0-39.9 kg/m^2), and class 3 (≥40.0 kg/m^2).(41)

Evaluating the Vanderbilt Cohort for Postoperative Complications

We evaluated for three of the most prevalent postoperative outcomes in abdominal surgery (e.g., postoperative infection, incisional hernia, and intestinal obstruction) using International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes. These outcomes were chosen to not only capture immediate postoperative outcomes, but also potential long-term consequences of surgical interventions, including those not present on the initial admission. To do so, all distinct ICD-9-CM codes from each individuals' record were captured and translated into PheWAS codes (phecodes).(42, 43)

Phecodes are a hierarchical classification system for ICD-9-CM codes and have been previously shown to appropriately categorize diseases in clinical practice.(42, 44) Existing phecodes were reviewed by a team member with surgical expertise, and these three phecodes were selected as outcomes for this study because they represent well-defined surgical complications with explicit ICD-9 billing codes. To improve the accuracy of mapping ICD-9 codes to phenotypes, a minimum of 2 instances of a matching ICD-9 code on separate days was required to be translated to a phecode. In order to capture long-term sequelae and patients who underwent surgery at a separate institution, the full cohort was not limited to patients who had undergone surgery at Vanderbilt and no specific timepoint for the postoperative complication following abdominal surgery was defined.

We performed a sequence of logistic regression models adjusted for age, sex, and self-reported race, with the predictor being the BMI category and outcomes being each postoperative complication. Effect sizes for associations of BMI categories with postoperative outcomes are determined by comparison to those individuals with BMI in the normal range. All analyses were performed using the PheWAS code map version 1.2.(45) and PheWAS package for R statistical software, version 3.4.3.(46) Bonferroni correction for analyses with multiple comparisons was used to adjust the significance threshold to a two-sided p-value <0.003.

eMERGE Mendelian Randomization Cohort

The cohort utilized for MR analyses was obtained through the Electronic Medical Records and Genomics (eMERGE) Consortium, a national network organized and funded by the National Human Genome Research Institute (NHGRI).(47) This cohort included all individuals from institutions contributing data to the eMERGE network phases I-III. Inclusion criteria were age ≥18 years with extant genome-wide genotyping data and ICD-9-CM codes.

Genotyping and Imputation in the eMERGE Mendelian Randomization Cohort

Genotyping and imputation was performed to coalesce genetic results across 12 different sites and 78 genotype array batches in the eMERGE Consortium using the Michigan Imputation Server (48) and Haplotype Reference Consortium (HRC1.1).(49, 50) The resulting imputed genome wide set consists of approximately 40 million single nucleotide variant marker allele doses down to 0.1% minor allele frequency (MAF). Genotype array files were referenced to the build 37 genome position using the forward genome strand. Quality control included filtering for sample missingness <2.0% and SNP missingness <2.0% in data preprocessing. For duplicate samples on differing arrays, the sample with the most genotyped variants for that subject was selected for the merged dataset. Principal component analysis (PCA) using the first 10 principal components was performed to determine genetic ancestry using PLINK (51) with variants having >5% MAF. Single nucleotide variants with a missing rate >10% or not meeting the linkage disequilibrium threshold $r^2 < 0.7$ were excluded in PCA. We performed identity by descent (IBD) analysis to identify related individuals using probability of zero alleles IBD (Z0) < 0.83 and the probability of having one allele IBD (Z1) > 0.10 to capture first through third-degree relatives. The oldest family member from each family was included in the cohort analysis. We excluded suspected monozygotic twins or duplicates.

Construction of the Obesity Genetic Risk Score (GRS)

The GRS was calculated from 97 SNPs (Supplementary Table 1) associated with BMI at genome-wide significance in a prior meta-analysis of genome-wide association studies conducted by the Genetic Investigation of ANthropometric Traits (GIANT) Consortium.(38) For the 97 SNPs, the minimum mean imputation r^2 for any single SNP was 0.83 with an overall mean r^2 of 0.95. Using the all-ancestry beta-coefficients reported by GIANT, we calculated a GRS for obesity for each individual in our cohort. This obesity GRS was calculated as a sum of risk allele dosages weighted by the effect estimates. The effect estimates are described by the GIANT consortium as beta-coefficients per 1-SD unit of BMI (4.8 kg/m^2).

We measured the association of the GRS with BMI by calculation of the BMI variance explained (adjusted R^2) by the associated SNPs using linear regression models adjusted for site, age, sex, and the first 10 principal components.

Mendelian Randomization Analyses

We used MR to assess for association of genetic risk for obesity with the postoperative outcomes. We performed logistic regression, adjusted for site, age, sex, and the first 10 principal components, to calculate causal effect estimates for genetically-determined BMI on the postoperative complications. To adjust for multiple comparison analyses, we used a Bonferroni correction for association, giving a conservative significance threshold of p=0.017. Effect estimates are reported per 1-SD difference in BMI (derived from beta estimates and SD of 4.8 kg/m^2 in a prior cohort of 449,472 individuals).(26) Among individuals with both a calculated GRS and reported BMI, the logistic regression analyses were performed with additional adjustment for median BMI to assess for residual association between the instrumental variable (GRS) with the outcomes (postoperative complications) through a pathway external to BMI. Such associations could indicate pleiotropic genes, or genes that can affect multiple, distinct phenotypes, were included in the BMI GRS.

Surgical Cohort Analyses

The analyses were also performed in subsets of the Vanderbilt and eMERGE populations who had documentation of undergoing a procedure. These two separate cohorts of surgical patients were captured by extracting Current Procedural Terminology (CPT) codes and mapping them to aggregated procedure categories including general, urologic, or gynecologic abdominal operations.(52) To further ensure that surgical patients undergoing hernia repair were not driving the associations with incisional hernia, the analysis was performed with exclusion of individuals with a CPT code corresponding to hernia repair. We also evaluated for the difference in complications in patients who underwent exploratory laparotomy versus laparoscopy.

Lastly, we evaluated for the association of obesity with 90-day postoperative mortality among all individuals at Vanderbilt who had undergone an abdominal surgical procedure. We performed a logistic regression to measure the association of BMI category with 90-day postoperative mortality, adjusting for age, gender, and race.

Results

Demographics of the Vanderbilt and eMERGE Cohorts

After exclusion of adult individuals with only BMI values recorded in pregnancy (12,588 individuals) or BMI values out of range (354 individuals), there were 736,726 individuals in the Vanderbilt cohort. Of these, 68,266 had undergone an abdominal surgical procedure for inclusion in the Vanderbilt surgical cohort (Figure 1A). Median follow-up of individuals who underwent a surgical procedure was 6.9 years (range 0-30.4 years). In the eMERGE MR cohort, 65,174 individuals had extant genotyping and ICD-9 codes for analysis in the entire cohort, of which 15,355 had a CPT code for abdominal surgery for inclusion in the eMERGE surgical cohort (Figure 1B). Table 1 describes the institutions contributing patients to the eMERGE cohorts. The majority of individuals in all cohorts were female and white or European ancestry (Table 2). Median BMI of the Vanderbilt individuals was 27.3 kg/m^2 (IQR 23.6-32.0), which was similar to that in the eMERGE cohort and surgical subpopulations. Individuals with overweight or obese BMI comprised 65.2% (480,530 individuals) of the Vanderbilt cohort and 67.9% (35,722 individuals) of the eMERGE cohort.

Table 1. eMERGE Sites and Numbers of Individuals Contributing Adult Data

Site	Entire Cohort	Surgical Cohort
	Number of subjects, n (%) (Total n = 65,174)	Number of subjects, n (%) (Total n = 15,355)
Boston Children's Hospital	252 (0.4)	0
Children's Hospital of Philadelphia	4,649 (7.1)	24 (0.2)
Cincinnati Children's Hospital Medical Center	1,331 (2.0)	45 (0.3)
Columbia University	1,680 (2.6)	686 (4.5)
Geisinger	2,772 (4.3)	974 (6.3)
Harvard University	9,689 (14.9)	2,141 (13.9)
Kaiser Permanente Washington with the University of Washington and the Fred Hutchinson Cancer Research Center	3,197 (4.9)	763 (5.0)
Marshfield Clinic	3,683 (5.7)	1,711 (11.1)
Mayo Clinic	8,199 (12.6)	2,053 (13.4)
Mount Sinai	5,701 (8.7)	758 (4.9)
Northwestern University	4,431 (6.8)	848 (5.5)
Vanderbilt University	19,590 (30.1)	5,352 (34.9)

Abbreviations: eMERGE, Electronic Medical Records and Genomics consortium

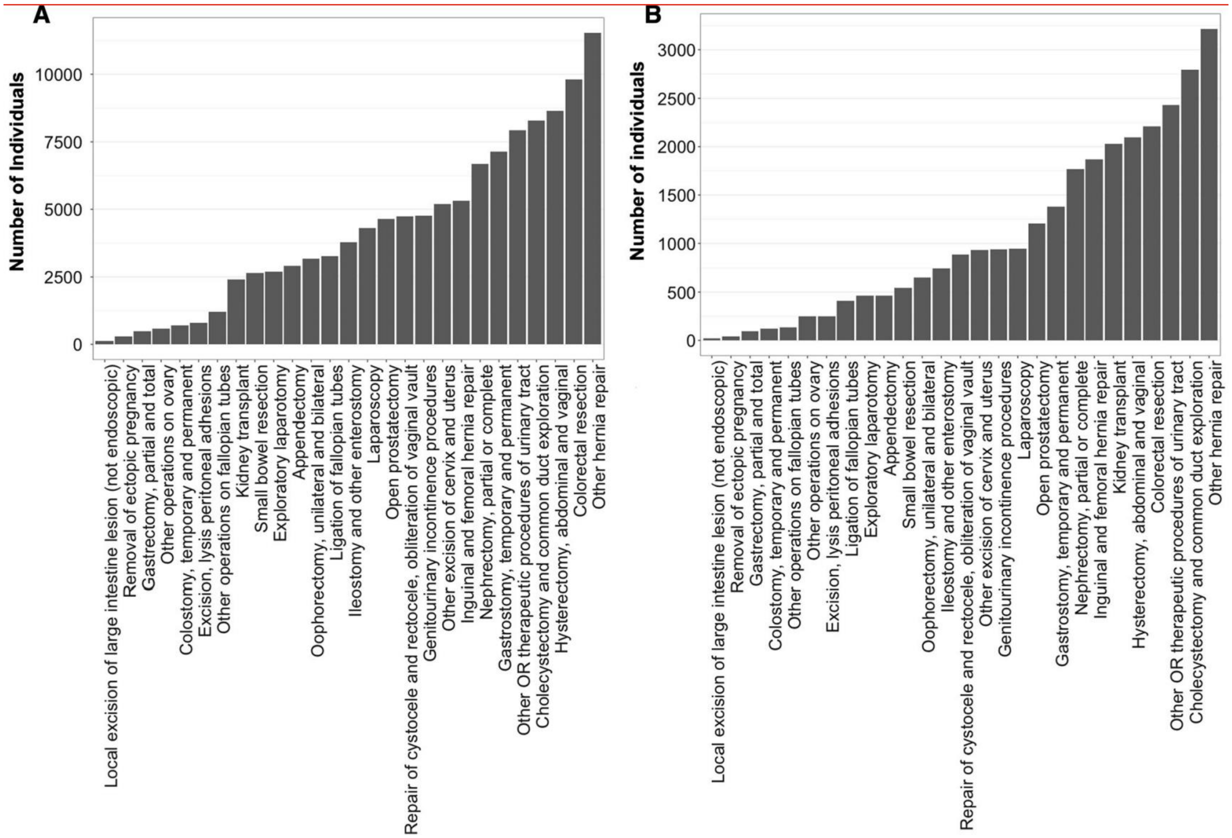


Figure 1. Frequency of Surgical Procedures in Vanderbilt (A) and eMERGE (B) Surgical Cohorts.

Table 2. Demographics for Vanderbilt and eMERGE cohorts

Clinical Variable	**Vanderbilt Cohort (n = 736,726)**	**Vanderbilt Surgical Cohort (n = 68,266)**	**eMERGE Cohort (n = 65,174)**	**eMERGE Surgical Cohort (n = 15,355)**
Age, median (IQR), years	49.0 (33.0 – 63.0)	54.0 (40.0 – 66.0)	67.0 (51.0 – 79.0)	69.0 (56.0-80.0)
Sex, No. (%)				
Female	434,266 (58.9)	41,077 (60.2)	35,997 (55.2)	8,701 (56.7)
Unknown	57 (0.1)	0	0	0
Race (Vanderbilt) or Genetic Ancestry (eMERGE), No. (%)				
White/European ancestry	553,368 (75.1)	55,694 (81.6)	52,760 (81.0)	13,068 (85.1)
Black/African ancestry	70,409 (9.6)	8,656 (12.7)	11,323 (17.4)	2,017 (13.1)
Asian	11,998 (1.6)	798 (1.2)	1,091 (1.7)	270 (1.8)
Other	18,332 (2.5)	1,942 (2.8)	0	0
Unknown	82,619 (11.2)	1,176 (1.7)	0	0
BMI (kg/m^2), median (IQR)	27.3 (23.6 – 32.0)	27.8 (24.1 – 32.7)	27.6 (23.9 – 32.1)	28.3 (24.8 – 33.1)
BMI (kg/m^2), mean (SD)	28.5 (7.0)	29.1 (7.3)	28.6 (7.0)	29.67 (7.0)
BMI category, No. (%)				
Underweight (<18.5)	15,509 (2.1)	1,564 (2.3)	1,961 (3.7)	221 (1.6)
Normal (18.5 – 24.9)	240,676 (32.7)	19,630 (28.8)	14,926 (28.4)	3,414 (25.1)
Overweight (25.0 – 29.9)	229,630 (31.2)	21,590 (31.6)	17,088 (32.5)	4,576 (33.7)
Obesity Class 1 (30.0 – 34.9)	135,488 (18.4)	13,317 (19.5)	10,425 (19.8)	2,827 (20.8)
Obesity Class 2 (35.0 – 39.9)	64,539 (8.8)	6,702 (9.8)	4,802 (9.1)	1,407 (10.4)
Obesity Class 3 (≥40.0)	50,873 (6.9)	5,463 (8.0)	3,407 (6.5)	1,146 (8.4)
No BMI reported	0	0	12,565	1,764

Abbreviations: eMERGE, Electronic Medical Records and Genomics consortium; BMI, body mass index; IQR, interquartile range; SD, standard deviation

BMI Associations with Postoperative Complications

In the Vanderbilt cohort, we found that overweight or obesity was associated with incisional hernia and postoperative infection in both the full and surgical cohorts (Table 3). There was an increased association with these postoperative complications with increasing BMI (Figure 2A-B) in both the complete and surgical Vanderbilt cohorts. In the entire Vanderbilt cohort, OR for incisional hernia in individuals with overweight BMI was 1.7 (95% confidence interval [CI] 1.5-1.8, p=3.1x10^{-20}) and increased to an OR of 5.5 (95% CI 5.4-5.6, p=2.2x10^{-172}) in class 3 obesity, a 3.2-fold increase. The association of obesity with incisional hernia persisted in the Vanderbilt subpopulation of individuals who had undergone general, urologic, or gynecologic abdominal surgery, with OR in surgical patients with overweight BMI of 1.6 (95% CI 1.5-1.7, $p=1.2x10^{-15}$) and surgical patients with class 3 obesity BMI of 4.9 (95% CI 4.8-5.1, $p=2.5x10^{-117}$). Further exclusion of individuals who had undergone hernia repair showed persistent associations of obesity with incisional hernia in class 1 (OR 1.7 [95% CI 1.4-2.1], $p=1.2x10^{-3}$), 2 (OR 3.5 [95% CI 3.1-3.8], $p=1.6x10^{-12}$) and 3 (OR 3.9 [95% CI 3.5-4.3], $p=1.2x10^{-12}$) obesity in the surgical patient population. In patients with both a low (<30 mg/kg^2) and high BMI (≥30 mg/kg^2), the large majority of incisional hernias presented themselves in the first 2 years following the index operation (Supplementary Figure 1).

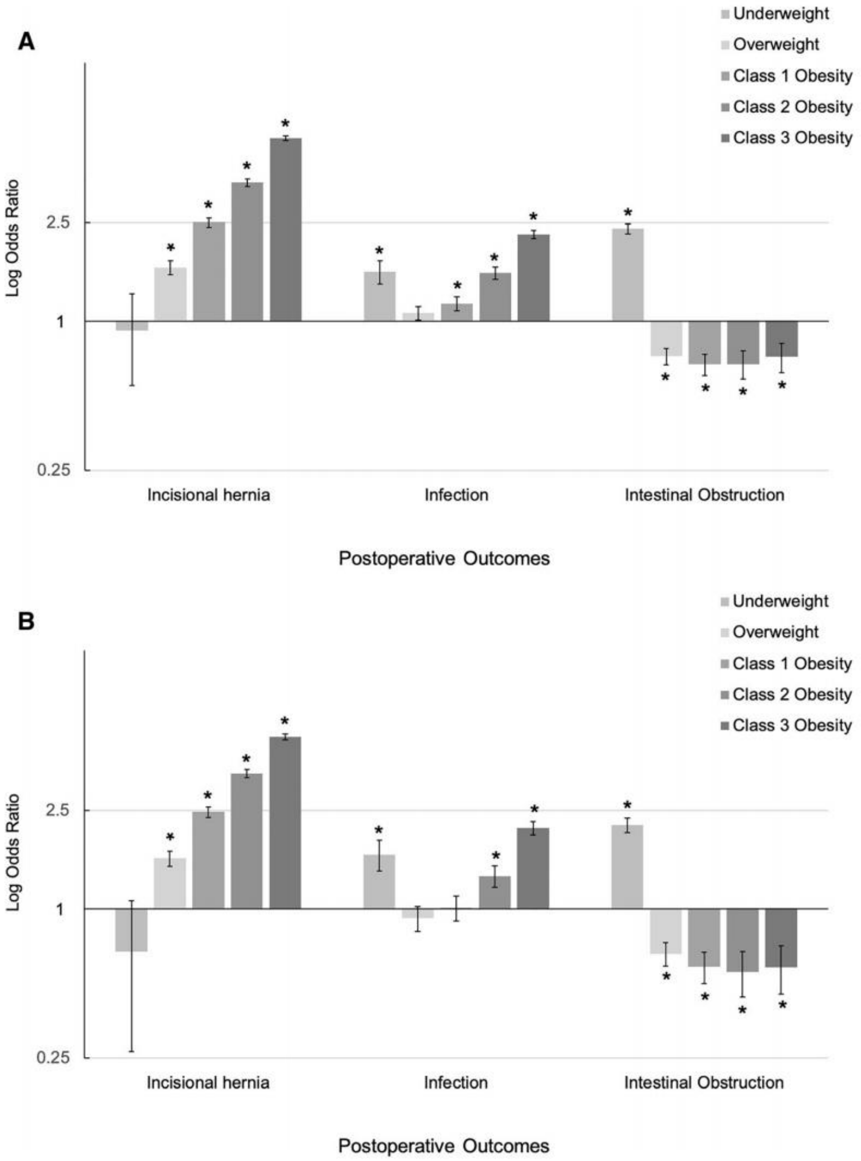


Figure 2. Association of BMI with Postoperative Complications in Vanderbilt General (A) and Surgical (B) Cohorts. Error bars represent 95% confidence interval. Significance threshold of p <0.003. BMI, body mass index; SD, standard deviation.

In the clinical cohort, both underweight (BMI <18.5 kg/m^2) and class 1-3 obesity (BMI ≥ 30.0 kg/m^2) demonstrated an association with postoperative infection (p<2.5x10^{-5}). The strongest association of postoperative infection with BMI class was with class 3 obesity (OR 2.3 [95% CI 2.2-2.3], p=2.3x10^{-71}) and this relationship persisted in the abdominal surgery subpopulation (OR 2.1 [95% CI 2.0-2.6], p=3.0x10^{-29}).

In the full cohort, patients with underweight BMI had an increased risk of intestinal obstruction compared to patients with a BMI within the normal range (OR 2.4 [95% CI 2.3-2.5], $p=4.6 \times 10^{-57}$). In contrast, patients with a BMI in the overweight or obese range showed a decreased risk of intestinal obstruction in comparison to patients with a normal BMI (Table 3). The finding of an increase in obstruction in patients with an underweight BMI and decrease in obstruction in patients with BMI over the normal range persisted in the subset of individuals who had undergone an abdominal surgical procedure.

Table 3. Association of BMI with postoperative complications[*]

Phenotype	Underweight <18.5 kg/m^2 OR (95% CI)	Overweight 25.0-29.9 kg/m^2 OR (95% CI)	Obesity Class 1 30.0-34.9 kg/m^2 OR (95% CI)	Obesity Class 2 35.0-39.9 kg/m^2 OR (95% CI)	Obesity Class 3 ≥40.0 kg/m^2 OR (95% CI)
Entire Vanderbilt Cohort (n = 736,726)					
Postoperative infection (n = 6,228)	1.59 (1.42-1.76)[†]	1.08 (1.01-1.15)	1.18 (1.10-1.26)[†]	1.57 (1.48-1.66)[†]	2.25 (2.16-2.33)[†]
Incisional hernia (n = 3,580)	0.92 (0.55-1.29)	1.65 (1.54-1.76)[†]	2.51 (2.40-2.62)[†]	3.62 (3.51-3.75)[†]	5.48 (5.36-5.60)[†]
Intestinal obstruction (n = 8,525)	2.37 (2.26-2.47)[†]	0.72 (0.67-0.78)[†]	0.67 (0.61-0.74)[†]	0.67 (0.58-0.76)[†]	0.72 (0.62-0.82)[†]
Vanderbilt Surgical Cohort (n = 68,266)					
Postoperative infection (n = 2,749)	1.66 (1.43-1.89)[†]	0.92 (0.81-1.02)	1.01 (0.89-1.13)	1.36 (1.22-1.49)[†]	2.13 (2.00-2.26)[†]
Incisional hernia (n = 3,120)	0.67 (0.27-1.08)	1.60 (1.49-1.72)[†]	2.47 (2.35-2.59)[†]	3.52 (3.39-3.65)[†]	4.95 (4.81-5.08)[†]
Intestinal obstruction (n = 5,389)	2.19 (2.03-2.33)[†]	0.66 (0.59-0.73)[†]	0.59 (0.50-0.67)[†]	0.56 (0.44-0.67)[†]	0.58 (0.45-0.71)[†]
Vanderbilt Exploratory Laparotomy Cohort (n = 2,410)					
Postoperative infection (n = 432)	0.95 (0.36-1.55)	1.25 (0.96-1.53)	1.14 (0.81-1.47)	1.39 (0.97-1.80)	2.95 (2.55-3.34)[†]
Incisional hernia (n = 296)	0.91 (0.08-1.73)	1.79 (1.42-2.15)[†]	2.00 (1.59-2.40)[†]	4.47 (4.03-4.92)[†]	4.63 (4.16-5.09)[†]
Intestinal obstruction (n = 688)	1.11 (0.63-1.59)	0.91 (0.66-1.16)	0.62 (0.32-0.92)[†]	1.14 (0.75-1.53)	0.66 (0.23-1.10)
Vanderbilt Laparoscopy Cohort (n = 3,841)					
Postoperative infection (n = 174)	1.06 (0.00-2.26)	1.06 (0.60-1.51)	1.17 (0.70-1.65)	1.19 (0.65-1.73)	1.56 (1.10-2.03)
Incisional hernia (n = 250)	0.89 (0.00-2.35)	1.35 (0.91-1.79)	2.15 (1.71-2.59)[†]	3.09 (2.63-3.55)[†]	3.54 (3.10-3.98)[†]
Intestinal obstruction (n = 327)	0.37 (0.80-2.26)	0.16 (0.44-1.05)	0.18 (0.21-0.91)[†]	0.25 (0.00-0.90)[†]	0.20 (0.15-0.95)

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); OR, odds ratio; CI, confidence interval

[*] Reference odds ratio 1.0 represents normal median BMI. Logistic regressions adjusted for sex, age, and reported race.[†] Results significant to Bonferroni corrected p-value of $p = 0.003$ compared to individuals with normal range BMI.

While patients with obesity who underwent exploratory laparotomy or laparoscopy both had increased risk for postoperative infection and incisional hernia, the risk was greatest in patients who underwent open laparotomy (Table 3). In class 3 obesity, the risk for postoperative infection for patients undergoing laparotomy was OR 3.0 (95% CI 2.6-3.3) compared to OR 1.6 (95% CI 1.1-2.0) in patients who underwent laparoscopy. Similarly, patients with class 3 obesity who underwent laparotomy had an OR of 4.6 (95% CI 4.2-5.1) for incisional hernia compared to OR 3.5 (95% CI 3.1-4.0) in those who underwent laparoscopy.

In comparison to individuals having normal BMI who underwent abdominal surgical procedure, those having an overweight or obese BMI had increased risk for mortality in the 90-day postoperative period. Increased BMI was associated with increased risk: overweight BMI had a OR of 1.02 (95% CI 1.0-1.0, p=0.04) while class 3 obesity had a OR of 1.12 (95% CI 1.1-1.2, p=2.5x10^{-11}).

Mendelian Randomization for Obesity Associations with Postoperative Complications

In the eMERGE cohort, the obesity GRS was strongly correlated with mean BMI (p<2.0x10^{-16}), aligning with findings from Locke et al.(38) Using a conservative p-value threshold of 0.017, the obesity GRS was associated with incisional hernia (OR 2.1 [95% CI 1.8-2.4], p=1.4x10^{-6}) and postoperative infection (OR 1.6 [95% CI 1.4-1.9], p=3.1x10^{-6}) in the entire eMERGE cohort (Table 4). Limiting to only those individuals who had undergone a general, urologic, or gynecologic abdominal surgery, the obesity GRS remained associated with both incisional hernia (OR 2.0 [95% CI 1.7-2.4], p=9.4x10^{-5}) and postoperative infection (OR 1.5 [95% CI 1.2-1.8], p=0.01).

The obesity GRS was not associated with intestinal obstruction in the complete (OR 1.1 [95% CI 0.9-1.2], p = 0.59) or surgical cohort (OR 1.0 [95% CI 0.7-1.2], p = 0.80).

Adjustment for median BMI in MR analyses to assess for residual association not attributable to BMI exposure showed attenuation of the associations with postoperative infection (p=0.126) and incisional hernia (p=0.038), suggesting that the association of the obesity-risk GRS with these postoperative complications is through BMI.

Table 4. Mendelian randomization genetic risk for obesity association with postoperative complications [*]

Entire eMERGE Cohort (n = 65,174)		
Phenotype	**OR per 1-SD BMI (95% CI)**	**p-value**
Incisional hernia [†] (n = 1,620)	2.14 (1.83-2.45)	1.4 x 10^{-6}
Postoperative infection [†] (n = 3,709)	1.64 (1.42-1.86)	3.1 x10^{-6}
Intestinal obstruction (n = 4,523)	1.05 (0.86-1.24)	0.595
eMERGE Surgical Cohort (n = 15,355)		
Phenotype	**OR per 1-SD BMI (95% CI)**	**p-value**
Incisional hernia [†] (n =1,356)	1.82 (1.66-2.36)	9.4 x 10^{-5}
Postoperative infection (n = 1,938)	2.01 (1.19-1.79)	0.009
Intestinal obstruction (n = 2,792)	0.97 (0.71-1.23)	0.801

Abbreviations: eMERGE, Electronic Medical Records and Genomics consortium; SE, standard error; OR, odds ratio

[*] Logistic regression adjusted for site, sex, age, and first ten principal components. Odds ratio report per 1-SD (4.8 kg/m^2) of BMI

[†] Results significant to Bonferroni corrected p-value = 0.017.

Discussion

This study found that obesity as measured by both BMI and genetic risk is associated with postoperative infections and incisional hernias in separate cohorts. The findings from this study are supported by prior reports in which overweight and obesity demonstrated an observed association with surgical site infections (10–23) and incisional hernias.(17, 53, 54) While these clinical associations have been demonstrated previously, the use of Mendelian randomization in this study suggests a possible causal role for obesity in the development of postoperative infections and incisional hernias.

It has long been studied whether it is obesity itself or the comorbidities found in obese patients, such as diabetes mellitus, are the driver for postoperative complications. There are many potential explanations for the association between obesity and postoperative infections and incisional hernias with the mechanism likely being multifactorial.(55) An increase in subcutaneous adipose tissue and local tissue trauma related to retraction could play a role. Subcutaneous tissue oxygenation is reduced in obese patients (56) and thus may reduce wound perfusion and predispose to wound infection and decreased healing, leading to both postoperative infections and incisional hernias. Lengthened operative time may also contribute to the increased incidence of surgical-site infections caused by obesity,(57) and surgical site infection itself is known to be a strong risk factor for incisional hernia formation.(58)

Because BMI itself is a strong predictor of postoperative complications, genetic variants are clinically unnecessary for estimation of the risk obesity plays in operative interventions. However, as genetic testing becomes less expensive and more common, it is another piece of data that can be leveraged in both research and clinical settings to provide for more accurate and validated predictions. Further, the use of genetic data to confirm clinical findings as we have demonstrated in this study substantiates the role obesity plays in development of postoperative incisional hernias and surgical site infections.

Interestingly, intestinal obstruction showed no association with obesity and, in the clinical cohort, was associated underweight BMI. This association is unclear and should be further investigated.

MR can be particularly useful because genetic variants are not subject to the same biases as traditional observational studies due to their random assortment during meiosis, thus allowing for potential causal inferences.(24, 59) Despite these advantages, MR and this study has several potential limitations. The method relies on BMI recorded in the EHR; however, this measure may not fully capture the true causal exposure of lifetime obesity exposure. Another limitation is that while the sensitivity analysis with adjustment of BMI suggested that pleiotropy of genetic variants did not play a significant role, pleiotropy is common and cannot be fully excluded. Lastly, the main limitation of this retrospective study is that it relies on both medical and procedural codes within the EHR, which can change over time, be inaccurate, and are often incomplete. While the use of ICD and CPT codes captures diagnoses and procedures at the study institution, we were unable to capture individuals who had surgery elsewhere or who presented to outside institutions with postoperative complications.

Conclusions

Genetic determinants of BMI suggest that obesity, aside from confounders or other metabolic diseases, is associated with the development of postoperative infection and incisional hernia. Thus, BMI represents an important risk factor for postoperative complication, warranting appropriate preoperative consideration and postoperative awareness.

Acknowledgements

Funding: JR Robinson receives support by the 5T15LM007450 training grant from the National Library of Medicine. Support for the research and personnel was also provided by the R01LM010685 grant from the National Library of Medicine. The eMERGE sites were funded through several series of grants from the National Human Genome Research Institute: U01HG8657, U01HG006375, U01HG004610 (Kaiser Permanente Washington/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672, U01HG006378, U01HG004608 (Vanderbilt University Medical Center); U01HG8666, U01HG006828 (Cincinnati Children's Hospital Medical Center); U01HG6379, U01HG04599 (Mayo Clinic); U01HG8679, U01HG006382 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG8684, U01HG006830 (Children's Hospital of Philadelphia); U01HG8673, U01HG006388, U01HG004609 (Northwestern University); U01HG8676 (Partners Healthcare/Broad Institute); U01HG8664 (Baylor College of Medicine); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG8701, U01HG006385, U01HG04603 (Vanderbilt University

Medical Center serving as the Coordinating Center); eMERGE Genotyping Centers were also funded through U01HG004438 (CIDR) and U01HG004424 (the Broad Institute). Vanderbilt University Medical Center’s Synthetic Derivative and BioVU are supported by institutional funding and by the CTSA grant ULTR000445 from NCATS/NIH.

References

1. Global BMI Mortality Collaboration null, Di Angelantonio E, Bhupathiraju S, Wormser D, Gao P, et al. 2016. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet Lond. Engl.* 388(10046):776–86

2. Prospective Studies Collaboration, Whitlock G, Lewington S, Sherliker P, Clarke R, et al. 2009. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet Lond. Engl.* 373(9669):1083–96

3. Emerging Risk Factors Collaboration, Wormser D, Kaptoge S, Di Angelantonio E, Wood AM, et al. 2011. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *Lancet Lond. Engl.* 377(9771):1085–95

4. Ogden CL, Carroll MD, Kit BK, Flegal KM. 2014. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA.* 311(8):806–14

5. GBD 2015 Obesity Collaborators, Afshin A, Forouzanfar MH, Reitsma MB, Sur P, et al. 2017. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N. Engl. J. Med.* 377(1):13–27

6. NCD Risk Factor Collaboration (NCD-RisC). 2016. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *Lancet Lond. Engl.* 387(10026):1377–96

7. Hawn MT, Bian J, Leeth RR, Ritchie G, Allen N, et al. 2005. Impact of obesity on resource utilization for general surgical procedures. *Ann. Surg.* 241(5):821–26; discussion 826-828

8. Surve A, Cottam D, Zaveri H, Cottam A, Belnap L, et al. 2018. Does the future of laparoscopic sleeve gastrectomy lie in the outpatient surgery center? A retrospective study of the safety of 3162 outpatient sleeve gastrectomies. *Surg. Obes. Relat. Dis. Off. J. Am. Soc. Bariatr. Surg.* 14(10):1442–47

9. Poelemeijer YQM, Marang-van de Mheen PJ, Wouters MWJM, Nienhuijs SW, Liem RSL. 2019. Textbook Outcome: an Ordered Composite Measure for Quality of Bariatric Surgery. *Obes. Surg.* 29(4):1287–94

10. Dindo D, Muller MK, Weber M, Clavien P-A. 2003. Obesity in general elective surgery. *Lancet Lond. Engl.* 361(9374):2032–35

11. Mullen JT, Moorman DW, Davenport DL. 2009. The obesity paradox: body mass index and outcomes in patients undergoing nonbariatric general surgery. *Ann. Surg.* 250(1):166–72

12. Giles KA, Hamdan AD, Pomposelli FB, Wyers MC, Siracuse JJ, Schermerhorn ML. 2010. Body mass index: surgical site infections and mortality after lower extremity bypass from the National Surgical Quality Improvement Program 2005-2007. *Ann. Vasc. Surg.* 24(1):48–56

13. Giles KA, Wyers MC, Pomposelli FB, Hamdan AD, Ching YA, Schermerhorn ML. 2010. The impact of body mass index on perioperative outcomes of open and endovascular abdominal aortic aneurysm repair from the National Surgical Quality Improvement Program, 2005-2007. *J. Vasc. Surg.* 52(6):1471–77

14. Mullen JT, Davenport DL, Hutter MM, Hosokawa PW, Henderson WG, et al. 2008. Impact of body mass index on perioperative outcomes in patients undergoing major intra-abdominal cancer surgery. *Ann. Surg. Oncol.* 15(8):2164–72

15. Bouwman F, Smits A, Lopes A, Das N, Pollard A, et al. 2015. The impact of BMI on surgical complications and outcomes in endometrial cancer surgery--an institutional study and systematic review of the literature. *Gynecol. Oncol.* 139(2):369–76

16. Wloch C, Wilson J, Lamagni T, Harrington P, Charlett A, Sheridan E. 2012. Risk factors for surgical site infection following caesarean section in England: results from a multicentre cohort study. *BJOG Int. J. Obstet. Gynaecol.* 119(11):1324–33

17. He Y, Wang J, Bian H, Deng X, Wang Z. 2017. BMI as a Predictor for Perioperative Outcome of Laparoscopic Colorectal Surgery: a Pooled Analysis of Comparative Studies. *Dis. Colon Rectum*. 60(4):433–45

18. Thelwall S, Harrington P, Sheridan E, Lamagni T. 2015. Impact of obesity on the risk of wound infection following surgery: results from a nationwide prospective multicentre cohort study in England. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* 21(11):1008.e1-8

19. Holley JL, Shapiro R, Lopatin WB, Tzakis AG, Hakala TR, Starzl TE. 1990. Obesity as a risk factor following cadaveric renal transplantation. *Transplantation*. 49(2):387–89

20. Thomas EJ, Goldman L, Mangione CM, Marcantonio ER, Cook EF, et al. 1997. Body mass index as a correlate of postoperative complications and resource utilization. *Am. J. Med.* 102(3):277–83

21. Jeschke E, Citak M, Günster C, Halder AM, Heller K-D, et al. 2018. Obesity Increases the Risk of Postoperative Complications and Revision Rates Following Primary Total Hip Arthroplasty: An Analysis of 131,576 Total Hip Arthroplasty Cases. *J. Arthroplasty*

22. Galyfos G, Geropapas GI, Kerasidis S, Sianou A, Sigala F, Filis K. 2017. The effect of body mass index on major outcomes after vascular surgery. *J. Vasc. Surg.* 65(4):1193–1207

23. Tjeertes EKM, Tjeertes EEKM, Hoeks SE, Hoeks SSE, Beks SBJ, et al. 2015. Obesity--a risk factor for postoperative complications in general surgery? *BMC Anesthesiol.* 15:112

24. Smith GD, Ebrahim S. 2003. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32(1):1–22

25. Nordestgaard BG, Palmer TM, Benn M, Zacho J, Tybjaerg-Hansen A, et al. 2012. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. *PLoS Med.* 9(5):e1001212

26. Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, et al. 2017. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. *JAMA Cardiol.* 2(8):882–89

27. Stender S, Nordestgaard BG, Tybjaerg-Hansen A. 2013. Elevated body mass index as a causal risk factor for symptomatic gallstone disease: a Mendelian randomization study. *Hepatol. Baltim. Md*. 58(6):2133–41

28. Lindström S, Germain M, Crous-Bou M, Smith EN, Morange P-E, et al. 2017. Assessing the causal relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum. Genet.* 136(7):897–902

29. Vimaleswaran KS, Berry DJ, Lu C, Tikkanen E, Pilz S, et al. 2013. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. *PLoS Med.* 10(2):e1001383

30. Huang Y, Xu M, Xie L, Wang T, Huang X, et al. 2016. Obesity and peripheral arterial disease: A Mendelian Randomization analysis. *Atherosclerosis*. 247:218–24

31. Mokry LE, Ross S, Timpson NJ, Sawcer S, Davey Smith G, Richards JB. 2016. Obesity and Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Med.* 13(6):e1002053

32. Gianfrancesco MA, Glymour MM, Walter S, Rhead B, Shao X, et al. 2017. Causal Effect of Genetic Variants Associated With Body Mass Index on Multiple Sclerosis Susceptibility. *Am. J. Epidemiol.* 185(3):162–71

33. Thrift AP, Shaheen NJ, Gammon MD, Bernstein L, Reid BJ, et al. 2014. Obesity and risk of esophageal adenocarcinoma and Barrett’s esophagus: a Mendelian randomization study. *J. Natl. Cancer Inst.* 106(11):

34. Jarvis D, Mitchell JS, Law PJ, Palin K, Tuupanen S, et al. 2016. Mendelian randomisation analysis strongly implicates adiposity with risk of developing colorectal cancer. *Br. J. Cancer*. 115(2):266–72

35. Dixon SC, Nagle CM, Thrift AP, Pharoah PD, Pearce CL, et al. 2016. Adult body mass index and risk of ovarian cancer by subtype: a Mendelian randomization study. *Int. J. Epidemiol.* 45(3):884–95

36. Chatterjee NA, Giulianini F, Geelhoed B, Lunetta KL, Misialek JR, et al. 2017. Genetic Obesity and the Risk of Atrial Fibrillation: Causal Estimates from Mendelian Randomization. *Circulation*. 135(8):741–54

37. Panoutsopoulou K, Metrustry S, Doherty SA, Laslett LL, Maciewicz RA, et al. 2014. The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a Mendelian randomisation study. *Ann. Rheum. Dis.* 73(12):2082–86

38. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 518(7538):197–206

39. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, et al. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* 84(3):362–69

40. Robinson JR, Wei W-Q, Roden DM, Denny JC. 2018. Defining Phenotypes from Clinical Data to Drive Genomic Research. *Annu Rev Biomed Data Sci*. In Publication:

41. *Defining Adult Overweight and Obesity | Overweight & Obesity | CDC*. www.cdc.gov

42. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* 26(9):1205–10

43. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86(4):560–72

44. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS One*. 12(7):e0175508

45. Carroll RJ, Bastarache L, Denny JC. 2014. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma. Oxf. Engl.* 30(16):2375–76

46. *R: The R Project for Statistical Computing*. www.r-project.org

47. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, et al. 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 15(10):761–71

48. *Michigan Imputation Server*. https://imputationserver.sph.umich.edu

49. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48(10):1279–83

50. Stanaway IB, Hall TO, Rosenthal EA, Palmer M, Naranbhai V, et al. 2019. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet. Epidemiol.* 43(1):63–81

51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81(3):559–75

52. *HCUP-US Tools and Software Page CCS-Services and Procedures*. www.hcup-us.ahrq.gov

53. Weissler JM, Lanni MA, Hsu JY, Tecce MG, Carney MJ, et al. 2017. Development of a Clinically Actionable Incisional Hernia Risk Model after Colectomy Using the Healthcare Cost and Utilization Project. *J. Am. Coll. Surg.* 225(2):274-284.e1

54. Ooms LS, Verhelst J, Jeekel J, Ijzermans JN, Lange JF, Terkivatan T. 2016. Incidence, risk factors, and treatment of incisional hernia after kidney transplantation: An analysis of 1,564 consecutive patients. *Surgery*. 159(5):1407–11

55. Falagas ME, Kompoti M. 2006. Obesity and infection. *Lancet Infect. Dis.* 6(7):438–46

56. Kabon B, Nagele A, Reddy D, Eagon C, Fleshman JW, et al. 2004. Obesity Decreases Perioperative Tissue Oxygenation. *Anesthesiology*. 100(2):274–80

57. Cheng H, Chen BP-H, Soleas IM, Ferko NC, Cameron CG, Hinoul P. 2017. Prolonged Operative Duration Increases Risk of Surgical Site Infections: A Systematic Review. *Surg. Infect.* 18(6):722–35

58. Murray BW, Cipher DJ, Pham T, Anthony T. 2011. The impact of surgical site infection on the development of incisional hernia and small bowel obstruction in colorectal surgery. *Am. J. Surg.* 202(5):558–60

59. Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. 2006. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am. J. Epidemiol.* 163(5):397–403

CHAPTER VI

Quantifying the Phenome-wide Disease Burden of Obesity using Electronic Health Records and Genomics

Jamie R. Robinson, M.D., M.S.[1,19], Robert J. Carroll, Ph.D.[1], Lisa Bastarache, M.S.[1], Qingxia Chen, Ph.D. [1,2], James Pirruccello, M.D.[3], Zongyang Mou, M.D.[4], Wei-Qi Wei, M.D., Ph.D.[1], John Connolly, Ph.D.[5], Frank Mentch, Ph.D.[5], Paul K. Crane, M.D., MPH[6], Scott J. Hebbring, Ph.D.[7], David R. Crosslin, Ph.D.[8], Adam S. Gordon, Ph.D.[9], Elisabeth A. Rosenthal, Ph.D.[10], Ian B. Stanaway, Ph.D.[8], M. Geoffrey Hayes, Ph.D.[11], Wei Wei, Ph.D.[12], Lynn Petukhova, Ph.D.[13], Bahram Namjou-Khales, M.D.[14], Ge Zhang, M.D., Ph.D.[14], Mayya S. Safarova, M.D., Ph.D. [15], Nephi A. Walton, M.D., M.S.[16], Christopher Still, D.O.[16], Erwin P. Bottinger, M.D.[17], Ruth J. F. Loos, Ph.D.[17], Shawn N. Murphy, M.D., Ph.D.[18], Gretchen P. Jackson, M.D., Ph.D.[1,19], Naji Abumrad, M.D. [19], Iftikhar J. Kullo, M.D.[15], Gail P. Jarvik, M.D., Ph.D.[10], Eric B. Larson, M.D., MPH[20], Chunhua Weng, Ph.D.[21], Dan Roden, M.D.[1], Amit V. Khera, M.D., M.Sc. [3,22], Joshua C. Denny, M.D., M.S.[23*]

Affiliations:

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[2] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[3] Center for Genomics Medicine, Massachusetts General Hospital, Boston, MA, USA

[4] Department of Surgery, University of California, San Diego, CA, USA

[5] The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

[6] Department of Medicine, University of Washington, Seattle, WA, USA

[7] Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA.

[8] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA

[9] Department of Pharmacology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[10] Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington Medical Center, Seattle, WA, USA

[11] Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[12] University of Pittsburgh Medical Center, Pittsburgh, PA, USA

[13] Department of Epidemiology, Columbia University, New York, NY, USA

[14] Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[15] Department of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA

[16] Department of Biomedical and Translational Informatics, Geisinger Health System, Danville, PA, USA

[17] The Charles Bronfman Institute for Personalized Medicine at Mount Sinai, The Mindich Child Health and Development Institute, New York, NY, USA

[18] Department of Neurology, Partners Healthcare, Boston, MA, USA

[19] Department of Surgery, Vanderbilt University Medical Center, Nashville, TN, USA

[20] Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

[21] Department of Biomedical Informatics, Columbia University, New York, NY, USA

[22] Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[23]All of Us Research Program, National Institutes of Health, Bethesda, MD.

Abstract

High body mass index (BMI) is associated with many comorbidities and mortality; however, the overall phenomic burden of obesity remains unknown. We performed a phenome-wide association study (PheWAS) of BMI using electronic health records (EHRs) from a clinical cohort of 736,726 adults and then followed it with genetic association studies using 2 separate cohorts with genome-wide genotyping data, one consisting of 65,174 adults in the eMERGE Network and another with 405,532 participants in the UK Biobank. In the clinical cohort PheWAS, class 3 obesity (BMI ≥40 kg/m^2) was associated with 433 phenotypes, representing 59.3% of all billed ICD9 codes in class 3 obese individuals. A polygenic risk score for BMI was also associated with 296 (68.4%) of the associations with class 3 obesity in the primary cohort, including type 2 diabetes, sleep apnea, hypertension, and chronic liver disease. In all 3 cohorts, 199 phenotypes were associated with class 3 obesity and polygenic risk for obesity. A predicted 17.1% of disease was attributable to obesity, including 87% of sleep apnea and 72% of type 2 diabetes. High BMI is a potentially modifiable risk factor associated with a broad range of diseases.

Introduction

Over two-thirds of the adult population in the United States is overweight or obese.(1) The prevalence of obesity, defined as having a body-mass index (BMI) of 30.0 kg/m^2 or greater, has doubled in over 70 countries in the last 3 decades.(2, 3) Prospective large-scale observational studies have shown that BMI above the normal range (overweight, BMI $\geq$ 25.0 kg/m^2) is associated with significant disease morbidity and increased overall mortality.(4–6) However, most prior studies focus on a single disease or a set of related diseases, leaving the overall disease burden associated with obesity unknown.

Obesity has a strong genetic predilection. It is known that rare genetic variants in *MC4R*(7) and *LEP*(8) and other genes can strongly influence obesity. However, for the large majority of individuals with obesity, their genetic risk stems from the cumulative effect of numerous more common genetic risk factors with fairly modest effect sizes.(9, 10) To date, the largest genome-wide association studies (GWAS) evaluating the genetic basis of obesity evaluated the relationship between 2.1 million common variants and BMI in 300,000 individuals, finding 97 single nucleotide polymorphisms (SNPs) associated with BMI and accounting for approximately 2.7% of the variation in BMI.(11)

Prior studies have suggested that obesity-risk SNPs are associated with ischemic heart disease,(12, 13) hypertension,(13) type 2 diabetes,(13) atrial fibrillation,(14) symptomatic cholelithiasis,(15) osteoarthritis,(16) and deep venous thrombosis,(17) among others.(18–25) These studies are limited by the use of either a single or limited number of genetic polymorphisms associated with BMI and evaluation for an association with a single comorbid phenotype. Genome-wide polygenic risk scores (PRS) have demonstrated the ability to predict disease occurrence and earlier onset of disease.(26) A recent study by Khera et al. has shown that incorporation of 2.1 million common variants into a quantitative genome-wide PRS for BMI can identify individuals with genetic risk for obesity comparable to that of rare monogenic mutations in *MC4R.*(10) Further, they also found that those individuals with a high PRS were at increased risk for six common cardiometabolic diseases, including coronary artery disease, diabetes mellitus, hypertension, congestive heart failure, ischemic stroke, and venous thromboembolism. Here, we aim to extend these prior studies by systematically evaluating the association of genetic risk for obesity with diseases in both a genome- and phenome-wide approach.

We leveraged a large electronic health record (EHR) population to perform a phenome-wide association study (PheWAS)(27, 28) to provide insights into patterns of disease associated with BMI. We then used genomic risk to predict BMI and obesity in two cohorts, using both known common genetic variants associated with obesity(11) and a genome-wide polygenic score (Figure 1).(10) These studies demonstrated that genetic risk for obesity is associated with increased risk for almost 200 diseases across the phenome, 42% of which were attributable to obesity, accounting for 17% of the total disease burden in individuals with obesity.

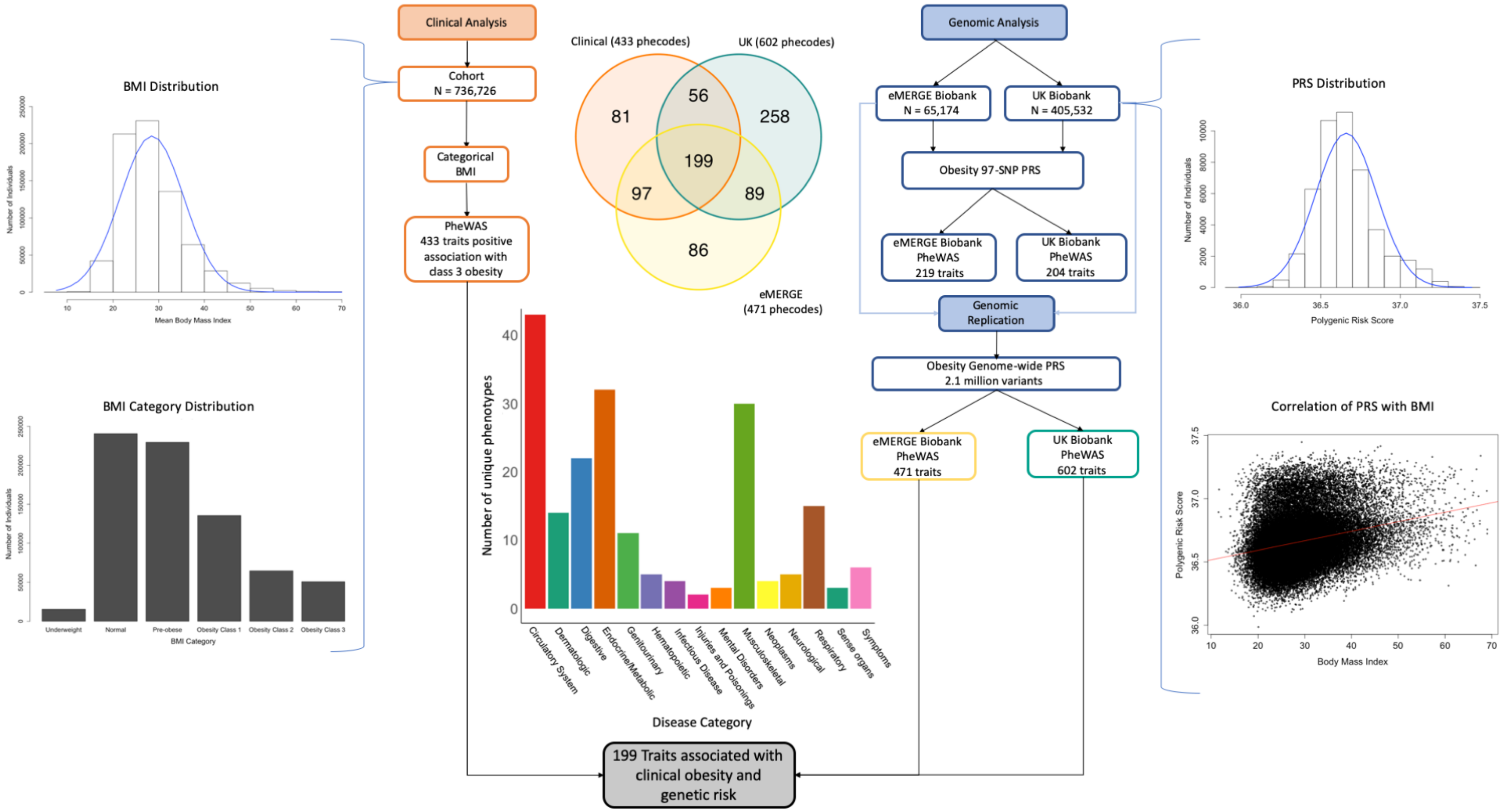

BMI Distribution
Number of Individuals
0
50000
100000
150000
200000
250000
10
20
30
40
50
60
70
Mean Body Mass Index
BMI Category Distribution
Number of Individuals
0
50000
100000
150000
200000
250000
Underweight
Normal
Pre-obese
Obesity Class 1
Obesity Class 2
Obesity Class 3
BMI Category
Clinical Analysis
Cohort
N = 736,726
Categorical
BMI
PheWAS
433 traits positive
association with
class 3 obesity
Clinical (433 phecodes)
UK (602 phecodes)
eMERGE
(471 phecodes)
81
56
258
199
97
89
86
Number of unique phenotypes
0
10
20
30
40
Circulatory System
Dermatologic
Digestive
Endocrine/Metabolic
Genitourinary
Hematopoietic
Infectious Disease
Injuries and Poisonings
Mental Disorders
Musculoskeletal
Neoplasms
Neurological
Respiratory
Sense organs
Symptoms
Disease Category
199 Traits associated with
clinical obesity and
genetic risk
Genomic Analysis
eMERGE Biobank
N = 65,174
UK Biobank
N = 405,532
Obesity 97-SNP PRS
eMERGE Biobank
PheWAS
219 traits
UK Biobank
PheWAS
204 traits
Genomic
Replication
Obesity Genome-wide PRS
2.1 million variants
eMERGE Biobank
PheWAS
471 traits
UK Biobank
PheWAS
602 traits
PRS Distribution
Number of Individuals
0
2000
4000
6000
8000
10000
36.0
36.5
37.0
37.5
Polygenic Risk Score
Correlation of PRS with BMI
Polygenic Risk Score
36.0
36.5
37.0
37.5
10
20
30
40
50
60
70
Body Mass Index

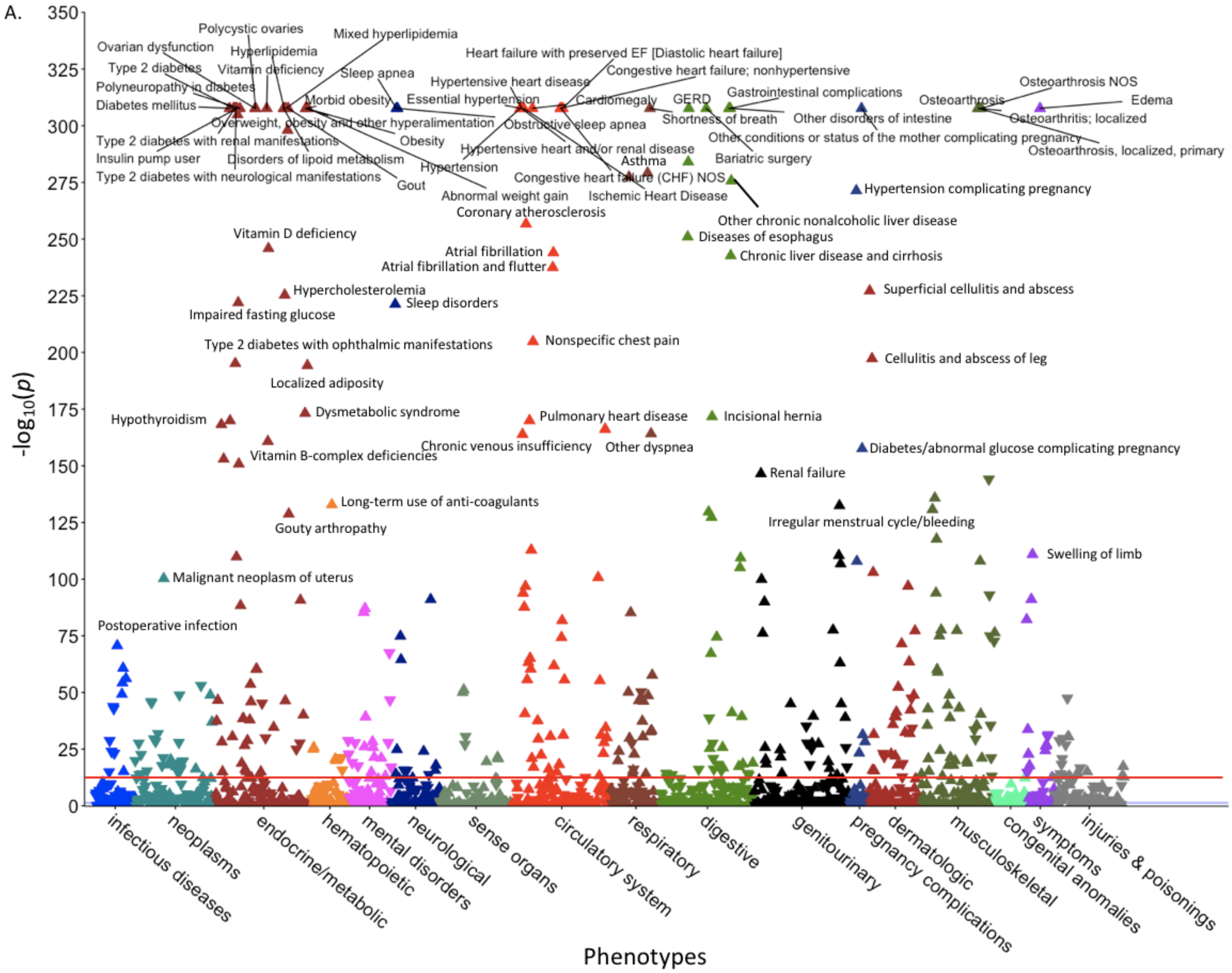
Figure 1. Clinical and Genomic Analysis Flow. Clinical analysis performed with PheWAS using BMI against 1816 traits. Genomic analysis performed in 2 separate cohorts with PheWAS against a 97-SNP PRS and genome-wide PRS. BMI distribution in the clinical cohort is demonstrated. Genome-wide PRS distribution and correlation with BMI in the eMERGE cohort is demonstrated. 199 disease phenotypes across all disease categories were associated with class 3 obesity in the clinical cohort and the genome-wide PRS in the eMERGE and UK Biobank cohorts.

Results

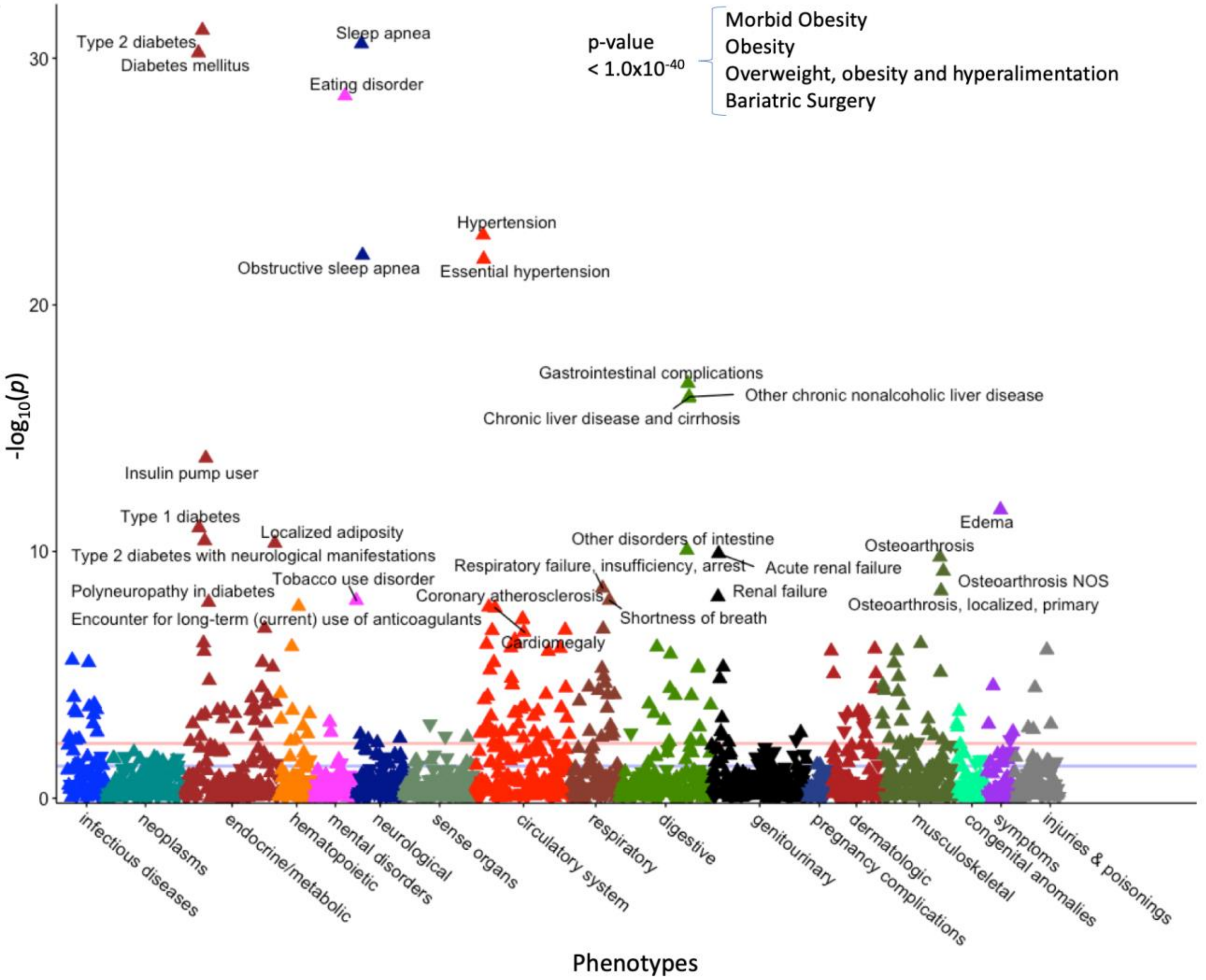
Phenotypes Across BMI Categories in Clinical Cohort

We first performed a PheWAS in 736,726 adults (age >18 years) in a clinical cohort (Supplemental Table 1). Of these individuals, 434,266 (58.9%) were female and the majority reported race as Caucasian (553,368, 75.1%). A median of 5 BMI assessments (IQR 2-13) were available per individual with a median BMI of 27.3 kg/m^2 (IQR 23.6-32.0). Median BMI was classified into one of 6 BMI categories for analysis, as defined by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO), including underweight (<18.5 kg/m^2), normal (18.5-24.9 kg/m^2), overweight (25.0-29.9 kg/m^2), and obesity class 1 (30.0-34.9 kg/m^2), class 2 (35.0-39.9 kg/m^2), and class 3 (≥40.0 kg/m^2).(29) Overweight or obese individuals comprised 65.2% of the clinical cohort population.

The PheWAS was performed using separate models with both mean BMI and categorical BMI as the predictor using Bonferroni significance thresholds. Mean BMI was associated with risk of 504 phenotypes in the clinical cohort. Obesity class 3 was associated with risk of 433 clinical phenotypes (23.8% of 1816 phenotypes included in the analysis; Figure 2A) across all broad disease categories compared to normal BMI; 170 phenotypes also demonstrated positive association with overweight and lesser obesity classes (Figure 3). The strongest associations of overweight and obesity classes with phenotypes of chronic diseases included essential hypertension (odds ratio [OR] 1.48 [95% CI 1.46-1.50) - 4.97 [4.94-5.00], $p < 1.0 \times 10^{-300}$), type 2 diabetes (OR 1.68 [1.65-1.71] - 8.28 [8.25-8.32], $p<2.1\text{x}10^{-289}$), obstructive sleep apnea (OR 2.68 [2.62-2.74] - 27.36 [27.30-27.42], $p<4.0\text{x}10^{-221}$), and polycystic ovarian disorder (OR 3.07 [2.94-3.19] - 23.18 [23.06-23.29], $p<1.1\text{x}10^{-72}$). Across phenotypes, effect sizes were higher for increasing BMI (Table 1). 59.3% of all billed diagnosis codes in individuals with class 3 obesity corresponded to a disease associated with BMI.

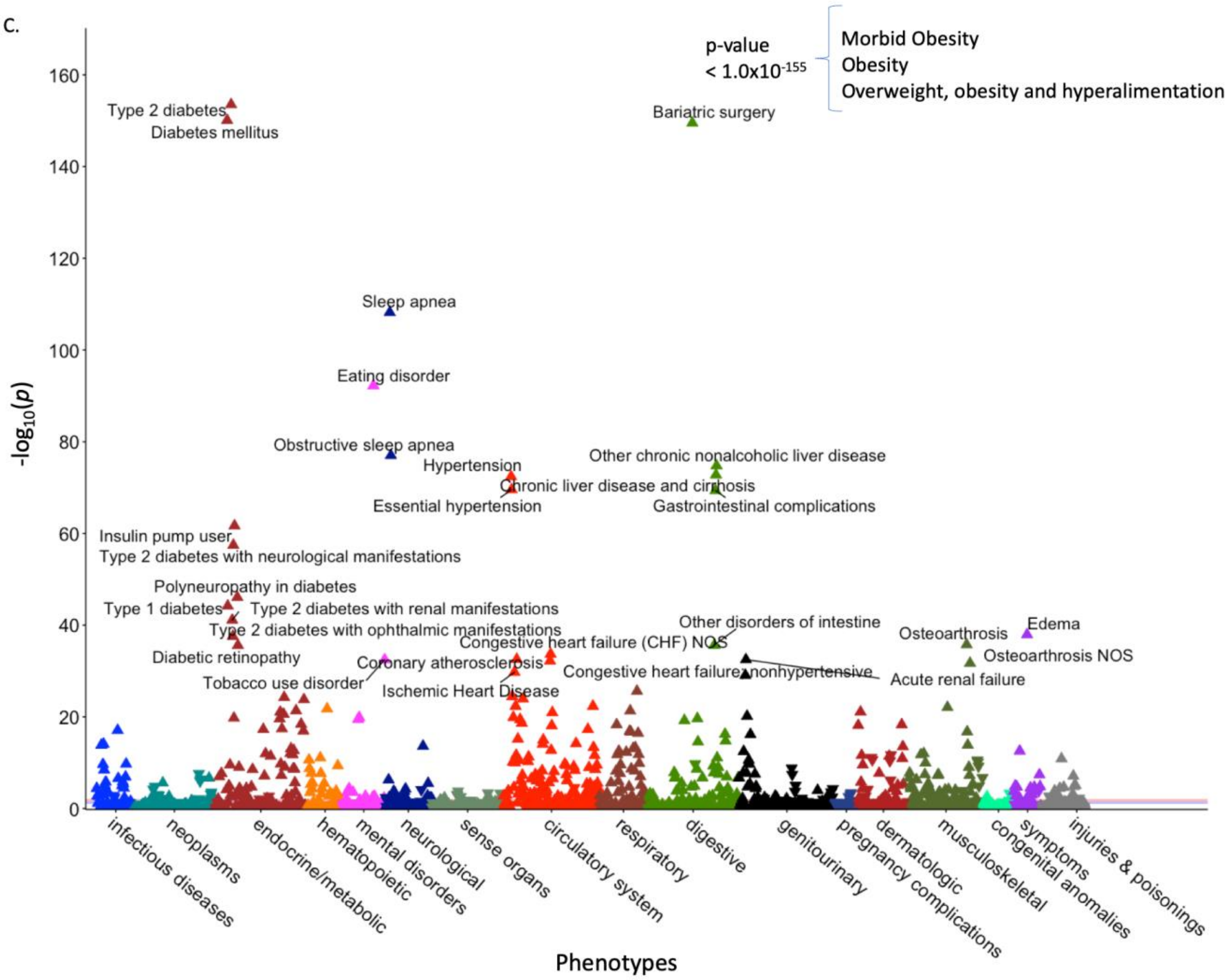

A.
350
325
300
275
250
225
200
175
150
125
100
75
50
25
0
$-\log_{10}(p)$
Polycystic ovaries
Mixed hyperlipidemia
Ovarian dysfunction
Hyperlipidemia
Type 2 diabetes
Vitamin deficiency
Sleep apnea
Polyneuropathy in diabetes
Diabetes mellitus
Morbid obesity
Overweight, obesity and other hyperalimentation
Type 2 diabetes with renal manifestations
Obesity
Insulin pump user
Disorders of lipoid metabolism
Type 2 diabetes with neurological manifestations
Gout
Heart failure with preserved EF [Diastolic heart failure]
Congestive heart failure; nonhypertensive
Hypertensive heart disease
Essential hypertension
Cardiomegaly
GERD
Gastrointestinal complications
Obstructive sleep apnea
Shortness of breath
Other disorders of intestine
Hypertensive heart and/or renal disease
Other conditions or status of the mother complicating pregnancy
Hypertension
Asthma
Bariatric surgery
Congestive heart failure (CHF) NOS
Abnormal weight gain
Ischemic Heart Disease
Coronary atherosclerosis
Osteoarthrosis NOS
Osteoarthrosis
Edema
Osteoarthritis; localized
Osteoarthrosis, localized, primary
Hypertension complicating pregnancy
Other chronic nonalcoholic liver disease
Diseases of esophagus
Chronic liver disease and cirrhosis
Vitamin D deficiency
Atrial fibrillation
Atrial fibrillation and flutter
Superficial cellulitis and abscess
Hypercholesterolemia
Sleep disorders
Impaired fasting glucose
Nonspecific chest pain
Type 2 diabetes with ophthalmic manifestations
Cellulitis and abscess of leg
Localized adiposity
Dysmetabolic syndrome
Pulmonary heart disease
Incisional hernia
Hypothyroidism
Chronic venous insufficiency
Other dyspnea
Diabetes/abnormal glucose complicating pregnancy
Vitamin B-complex deficiencies
Renal failure
Long-term use of anti-coagulants
Gouty arthropathy
Irregular menstrual cycle/bleeding
Swelling of limb
Malignant neoplasm of uterus
Postoperative infection
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & poisonings
Phenotypes

B.



p-value
< 1.0x10-40
Morbid Obesity
Obesity
Overweight, obesity and hyperalimentation
Bariatric Surgery
Type 2 diabetes
Diabetes mellitus
Sleep apnea
Eating disorder
Hypertension
Obstructive sleep apnea
Essential hypertension
Gastrointestinal complications
Other chronic nonalcoholic liver disease
Chronic liver disease and cirrhosis
Insulin pump user
Type 1 diabetes
Localized adiposity
Type 2 diabetes with neurological manifestations
Tobacco use disorder
Polyneuropathy in diabetes
Encounter for long-term (current) use of anticoagulants
Other disorders of intestine
Respiratory failure, insufficiency, arrest
Acute renal failure
Renal failure
Coronary atherosclerosis
Shortness of breath
Cardiomegaly
Edema
Osteoarthrosis
Osteoarthrosis NOS
Osteoarthrosis, localized, primary
$-\log_{10}(p)$
30
20
10
0
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & poisonings
Phenotypes

C.


p-value
< 1.0x10-155
Morbid Obesity
Obesity
Overweight, obesity and hyperalimentation
Type 2 diabetes
Diabetes mellitus
Bariatric surgery
Sleep apnea
Eating disorder
Obstructive sleep apnea
Hypertension
Other chronic nonalcoholic liver disease
Chronic liver disease and cirrhosis
Essential hypertension
Gastrointestinal complications
Insulin pump user
Type 2 diabetes with neurological manifestations
Polyneuropathy in diabetes
Type 1 diabetes
Type 2 diabetes with renal manifestations
Type 2 diabetes with ophthalmic manifestations
Other disorders of intestine
Edema
Osteoarthrosis
Congestive heart failure (CHF) NOS
Osteoarthrosis NOS
Diabetic retinopathy
Coronary atherosclerosis
Congestive heart failure nonhypertensive
Acute renal failure
Tobacco use disorder
Ischemic Heart Disease
-log10(p)
160
140
120
100
80
60
40
20
0
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & poisonings
Phenotypes

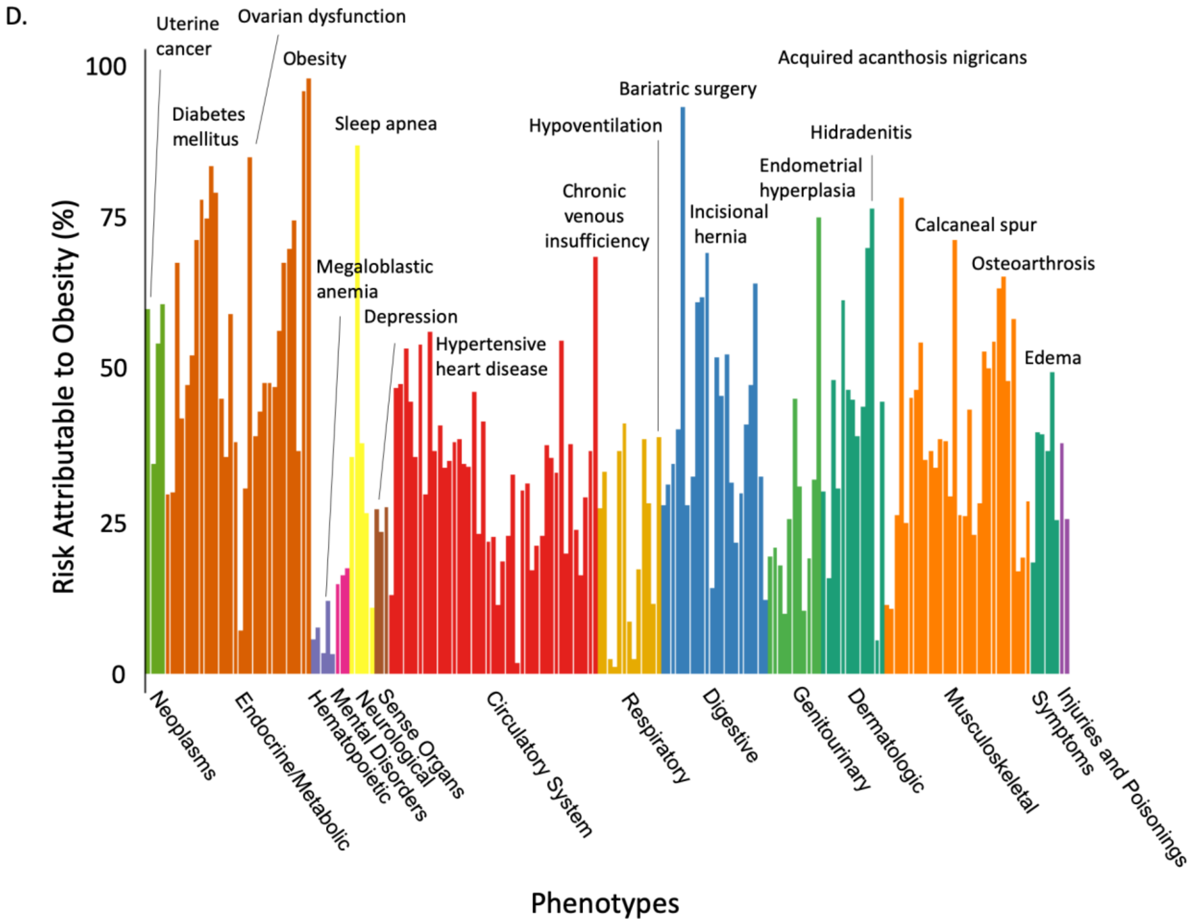


Figure 2. A. Association of Class 3 Obesity with Diseases in PheWAS. B. Association of Obesity 97-SNP PRS with Diseases in PheWAS in eMERGE Cohort. C. Association of Genome-wide PRS with Diseases in PheWAS in eMERGE Cohort. Blue horizontal line represents *p* = 0.05. Red horizontal line represents Bonferroni significance threshold (*p* = 5.6 x 10^{-6}) for clinical analysis [A] and false discovery rate significance threshold = 0.05 for genomic analysis [B-C]). Point direction relates to directionality of odds ratio: upward triangles are associated with increased risk for patients while downward triangles are associated with decreased risk. **D. Risk Attributable to Obesity with Normalization of Obesity Classes 1-3 to Normal BMI.** Phenotypes shown are 199 phenotypes that were associated with class 3 obesity in the clinical cohort and obesity polygenic risk score in both eMERGE and UK biobank cohorts. Example phenotypes are annotated.

Table 1. Associations Between BMI Categories and Common Phenotypes [a]

Phenotype	Underweight <18.5 kg/m^2 OR (95% CI)	Overweight 25.0-29.9 kg/m^2 OR (95% CI)	Obesity Class 1 30.0-34.9 kg/m^2 OR (95% CI)	Obesity Class 2 35.0-39.9 kg/m^2 OR (95% CI)	Obesity Class 3 ≥40.0 kg/m^2 OR (95% CI)
Type 2 Diabetes Mellitus	0.81 (0.71-0.91)	1.68 (1.65-1.71) [b]	3.06 (3.04-3.10) [b]	5.24 (5.21-5.27) [b]	8.28 (8.25-8.32) [b]
Polycystic Ovaries	0.45 (-0.05-0.95)	3.07 (2.94-3.19) [b]	6.44 (6.32-6.56) [b]	13.36 (13.24-13.48) [b]	23.18 (23.06-23.29) [b]
Vitamin Deficiency	1.26 (1.15-1.37)	1.09 (1.05-1.13) [b]	1.33 (1.29-1.37) [b]	1.76 (1.71-1.81) [b]	2.65 (2.61-2.70) [b]
Hyperlipidemia	0.49 (0.40-0.57) [b]	1.64 (1.62-1.67) [b]	2.17 (2.15-2.19) [b]	2.66 (2.63-2.69) [b]	3.06 (3.03-3.09) [b]
Gout	0.60 (0.28-0.92)	1.80 (1.73-1.97) [b]	2.84 (2.77-2.91) [b]	4.01 (3.92-4.09) [b]	5.38 (5.30-5.47) [b]
Obstructive Sleep Apnea	0.68 (0.41-0.95)	2.67 (2.62-2.74) [b]	6.00 (5.94-6.06) [b]	12.29 (12.23-12.36) [b]	27.36 (27.30-27.42) [b]
Essential Hypertension	0.89 (0.83-0.95)	1.48 (1.46-1.50) [b]	2.23 (2.21-2.25) [b]	3.24 (3.22-3.26) [b]	4.97 (4.94-5.00) [b]
Ischemic Heart Disease	1.05 (0.97-1.13)	1.24 (1.21-1.26) [b]	1.60 (1.57-1.63) [b]	1.95 (1.91-1.98) [b]	2.19 (2.15-2.23) [b]
Heart Failure with Preserved EF	1.17 (0.93-1.41)	1.35 (1.27-1.43) [b]	2.40 (2.32-2.49) [b]	4.52 (4.44-4.62) [b]	9.05 (8.95-9.14) [b]
GERD	1.25 (1.18-1.33) [b]	1.23 (1.20-1.26) [b]	1.43 (1.40-1.45) [b]	1.67 (1.63-1.70) [b]	2.05 (2.02-2.09) [b]
Osteoarthrosis	0.58 (0.48-0.69) [b]	1.53 (1.50-1.56) [b]	2.10 (2.07-2.13) [b]	2.71 (2.67-2.74) [b]	3.71 (3.67-3.75) [b]
Asthma	1.27 (1.17-1.36) [b]	1.11 (1.08-1.15) [b]	1.33 (1.29-1.37) [b]	1.64 (1.59-1.69) [b]	2.29 (2.24-2.33) [b]
Nonalcoholic Liver Disease	1.04 (0.87-1.20)	1.27 (1.22-1.32) [b]	1.73 (1.68-1.79) [b]	2.44 (2.37-2.50) [b]	3.23 (3.18-3.30) [b]
Atrial Fibrillation	1.23 (1.13-1.33)	1.08 (1.04-1.11)	1.33 (1.29-1.36) [b]	1.67 (1.62-1.72) [b]	2.46 (2.41-2.51) [b]
Superficial Cellulitis/Abscess	1.18 (1.08-1.28)	1.00 (0.97-1.04)	1.19 (1.15-1.23) [b]	1.52 (1.47-1.57) [b]	2.14 (2.10-2.19) [b]

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); EF, ejection fraction; OR, odds ratio; CI, confidence interval

[a] Information shown in the table includes the most significant associations to class 3 obesity (by p-value) with exclusion of phenotypes definitive for obesity. For redundant phenotypes, those with strongest OR are shown. Reference odds ratio 1.0 represents normal median BMI.

[b] Results significant to Bonferroni corrected p-value of p = 5.6 x 10^{-6} compared to individuals with normal range BMI

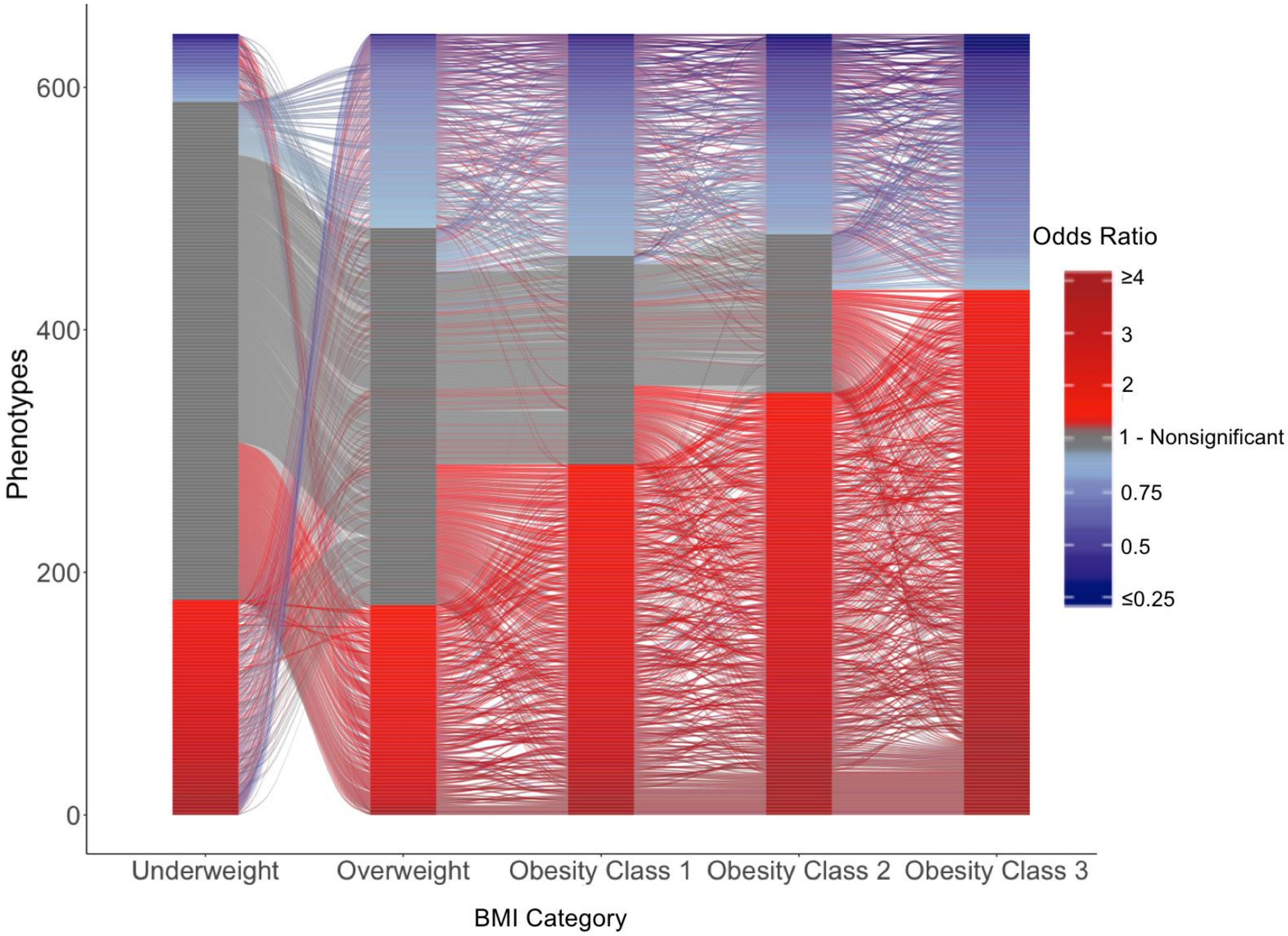


Figure 3. Trends of Odds Ratios in Phenotypes Significantly Associated with Class 3 Obesity in PheWAS. All phenotypes with significance in class 3 obesity are visualized (433 phenotypes with OR >1.0). Gray represents non-significant findings. Increasing odds ratios are seen with higher BMI for many phenotypes.

To assess the robustness of the clinical associations with obesity, we performed a tipping points analysis.(30) We found that 57-82% of all obesity PheWAS associations would still be present assuming presence of a hypothetical unmeasured binary confounder with OR=2 (Supplemental Figure 1).

Genetic Risk Score for Obesity

For the genomic analysis, we used 2 separate cohorts (Supplemental Table 1). The first consisted of 65,174 individuals from 12 institutions within the Electronic Medical Records and Genomics (eMERGE) network (Supplemental Table 2). The second cohort consisted of 405,532 participants within the UK Biobank. Overweight or obese individuals comprised 67.9% of the eMERGE cohort and 66.7% of the UK Biobank. A majority of both cohorts were of European ancestry (81.0% of eMERGE and 96.3% of the UK Biobank).

To evaluate the difference using SNPs known to be associated with BMI compared to a genome-wide association score, we performed the PheWAS analyses using both a limited PRS of 97 SNPs (Supplemental Table 3) and a genome-wide PRS of 2.1 million SNPs in both of the genomic cohorts (Figure 2B-C, Supplemental Figure 2-3). In the eMERGE cohort, the 97-SNP PRS explained 1.92% [95% CI 1.67-2.17] of the variance in mean BMI ($p<2.0x10^{-16}$) (Supplemental Figure 4A). The genome-wide PRS explained 9.51% [9.01-10.11] of the variance in mean BMI ($p<2.0x10^{-16}$) (Supplemental Figure 4B). Pearson correlation coefficient also showed a much stronger correlation between observed BMI and the genome-wide PRS (0.26 [95% CI 0.25-0.27]) compared to the 97-SNP PRS (0.11 [95% CI 0.10-0.12]).

In the eMERGE cohort, the 97-SNP obesity PRS was significantly associated with 161 (37.2 %) of the phenotypes showing association with class 3 obesity in the clinical cohort (Supplemental Table 4) with OR demonstrating positive direction of effect (i.e., risk with increasing BMI). Excluding phenotypes definitive for obesity (e.g., bariatric surgery, morbid obesity, and localized adiposity), some of the most significant associations were with type 2 diabetes (OR 1.99 [95% CI 1.87-2.11], p=7.35x10^{-32}), sleep apnea (OR 2.24 [2.10-2.38], p=2.59x10^{-31}), and hypertension (OR 1.82 [1.70-1.94], p=1.46x10^{-23}). Some of the strongest causal effect sizes were seen for panniculitis (OR 4.30 [3.73-4.87], p=5.3x10^{-7}), non-healing surgical wounds (OR 3.16 [2.71-3.61], p=9.7x10^{-7}), and polycystic ovaries (OR 2.7 [2.17-3.23], p=2.4x10^{-4}). Only 8 phenotypes positively associated with the obesity 97-SNP PRS were not clinically associated with class 3 obesity (Supplemental Table 5).

In the UK Biobank cohort, the 97-SNP obesity PRS was positively associated with 124 (28.6 %) of the phenotypes showing association with class 3 obesity in the clinical cohort. We replicated 77 of the associations with both class 3 obesity and the 97-SNP obesity PRS in the eMERGE cohort using the UK Biobank cohort (Supplemental Table 6).

The PheWAS analysis performed in the eMERGE cohort utilizing the genome-wide PRS (Figure 2C) showed a positive association with 296 (68.4%) of the phenotypes associated with class 3 obesity in the clinical cohort (Supplemental Table 7, replicating 135 more phenotype associations than the PRS with only 97 SNPs. PheWAS using the genome-wide PRS in the UK Biobank (Supplemental Figure 3) replicated 255 of the phenotype associations seen in class 3 obesity, replicating 131 more phenotypes that then 97-SNP PRS.

There were 199 phenotype associations replicated in the clinical data set and both genomic data sets using the genome-wide PRS (Supplemental Table 8) compared to 77 phenotypes replicated in all 3 cohorts using the 97-SNP PRS. Disease associations with obesity were replicated in all data sets across all predefined PheWAS disease classes: infections (bacterial infection, septicemia), neoplastic (uterine and renal cancer), endocrine (diabetes, hypothyroidism), hematologic (anemia), psychiatric (major depressive disorder), cardiovascular (hypertension, ischemic heart disease, chronic venous insufficiency), respiratory (sleep apnea, pulmonary hypertension), digestive (cholelithiasis, esophagitis, gastroesophageal reflux, liver disease), urologic (renal failure), rheumatologic (rheumatoid arthritis, gout), musculoskeletal (osteoarthritis, lumbar disc displacement), and dermatologic (psoriasis, hidradenitis).

Measured effect sizes for the genomic associations strongly correlated with observed BMI effect sizes in the clinical cohort ($p < 2.2 \times 10^{-16}$, $R^2 = 0.54$; Figure 4, Supplemental Table 9).

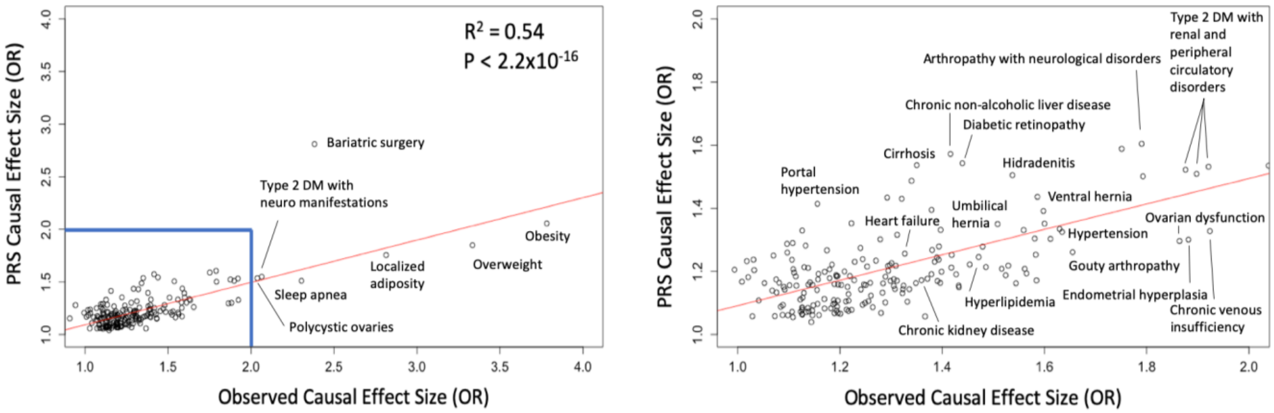


Figure 4. Clinically observed versus Genome-wide Obesity PRS PheWAS Causal Effect Sizes. Each dot represents a phenotype significantly associated with both class 3 obesity in clinical cohort and calculated genome-wide obesity polygenic risk score for obesity in eMERGE and UK biobank cohorts. 199 total phenotypes. Red line represents linear regression. Adjusted $R^2 = 0.544$.

Population Attributable Risk Due to Elevated BMI

For the 199 phenotypes associated with obesity in genetic association analysis, normalization of BMI for obese individuals predicted there were 607,430 cases in the clinical cohort (41.5% of all occurrences of the 199 significant phenotypes and 17.1% of all phenotypes in obese individuals) that were attributable to obesity (Supplemental Figure 5). The proportions of disease attributable to obesity in the in the clinical cohort for the 199 phenotypes associated with both clinical and genetic risk for obesity is shown in Figure 2D. There were 87.4% (17,624 cases) of sleep apnea, 71.7% (25,511 cases) of type 2 diabetes, 70.2% (3,349 cases) of gout, and 20.9% (3,332 cases) of renal failure predicted to be attributable to obesity. Our analysis suggested over half of the cases for 47 different diagnoses were attributable to obesity (Supplemental Table 10).

Discussion

Through a combined genome- and phenome- wide approach in both clinical and genomic cohorts, this study confirms that obesity is associated with a considerable burden of disease across all disease classes. Nearly one-quarter of disease phenotypes, across all major disease domains, were associated with class 3 obesity. Almost 200 phenotypes were associated with both class 3 obesity and genetic risk for obesity in 2 separate cohorts. The phenotypes associated with class 2 and 3 obesity resulted in over 50% of billed diagnosis codes in those individuals. Our analysis suggested that at least 17% of all phenotypes in the obese population where attributable to obesity.

The findings from this study are further supported by prior reports in which genetic risk for obesity was associated with other comorbidities.(12–18, 31–34) For example, Lyall et al. described similar effect estimates of obesity on coronary heart disease, hypertension, and diabetes using an obesity PRS comprised of 93 of the 97 SNPs used in construction of the PRS for the current study.(13) Zhu et al. also demonstrated the association of a PRS for BMI with disease several diseases including a summation of disease count; however, this study limited the PRS calculation to SNPs meeting GWAS significance.(35) The replication of findings in prior studies further validates the phenome-wide approach coupled with polygenic risk scores in this study. The highly stringent criteria for our final association results, which required significance with class 3 obesity in a clinical cohort along with the obesity PRS in two separate genotyped cohorts likely resulted in exclusion of some phenotypes with smaller effect sizes for association with obesity. For example, Lindstrom et al. demonstrated an association of genetic risk for obesity and venous thromboembolism(17), which we identified in the clinical and eMERGE cohorts, but not the UK Biobank cohort, possibly as a result of the differing demographics of the UK Biobank with fewer individuals with class 2 and 3 obesity.

The genome-wide approach in this study demonstrated 2.5x the power to detect significant associations compared to the limited PRS. Thus, despite the conservative study methods requiring 3 cohorts and family-wise significance thresholds, the use of a phenome-wide approach coupled to a genome-wide PRS allowed this study to identify novel associations of genetically-determined BMI with diseases, including increased risk of renal failure, urinary calculus, bundle branch block, cardiomyopathy, venous insufficiency, gastroesophageal reflux, spinal stenosis, tendon rupture, and rheumatoid arthritis.

Other interesting associations with little previous data were also supported by our study including asthma(36), cholelithiasis(15), postoperative complications, and major depression.

Resolution of obesity in some individuals can reduce disease burden for specific phenotypes. It is well-described that bariatric surgery can induce rapid reduction and cure of diabetes, hyperlipidemia, hypertension, and obstructive sleep apnea.(38) However, in individuals with obesity, many of the associated phenotypes will have already developed and are unlikely to be cured with weight-loss alone, and our analysis does not elucidate which diseases may regress with weight loss. For example, obesity was strongly associated with end-organ dysfunction, including cardiac, renal, or liver failure, evidence of long-standing effects from obesity. These data suggest that treatment of obesity may be a crucial component to ameliorate disease progression for a broader range of diseases than previously considered. The breadth of disease associated with obesity substantiates the principle that primary obesity prevention could have an enormous impact on healthcare, surpassing that of medical or surgical weight-loss alone.

This study has limitations. The method relies on BMI and billing codes recorded in the EHR; however, this measure may not fully capture the true causal exposure for some phenotypes (e.g., lifetime exposure of obesity). For some individuals, their entrance into a tertiary care health system occurs following a change in their health state that may also have affected his or her presenting BMI. As we included all recorded BMI values, we were unable to assess the temporal relationship between observations and BMI. As for the phenotypes, for the clinical and eMERGE analyses, ICD10 codes were not yet available for use; however, they were minimally used at the time of this data collection. ICD10 codes were however used in the phenotype mapping in the analysis using data from the UK Biobank. The phenotype coding system also can have a significant amount of overlap as it is a hierarchical structure; however, we were able to confirm that at least 95 of the phenotype disease associations were clinically unique. Although in the eMERGE cohort 17% of individuals were of African ancestry, the majority of the 3 cohorts primarily consisted individuals of European descent. It has been well-established that the predictive power of many PRSs are improved in European populations, and as such our PRS may underestimate the associations of phenotypes with genetic risk for obesity in non-European populations.(39) Nonlinearity of some associations may also bias towards the null hypothesis, limited the ability to identity those associations with elevated BMI or high genomic risk for elevated BMI.

This study is among the first to demonstrate the significant role that genetic risk for obesity plays in a systematic spectrum of diseases and the overall healthcare burden imposed by obesity. This comprehensive evidence on disease risk emphasizes that population-level reduction in BMI with efforts toward prevention could have a major impact on the incidence of disease globally. Future studies should assess the influence that environmental modifications, such as diet and exercise starting at a young age, in the setting of strong polygenic risk for obesity could have on the development of subsequent comorbidities.

Methods

Clinical cohort

We conducted a retrospective observational study using the Vanderbilt University Medical Center (VUMC) Synthetic Derivative, a de-identified version of over 3 million VUMC patient health records dating back several decades, depending on data type.(40, 41) The primary site study population consisted of all adult individuals (≥18 years of age) with at least one documented BMI. The study protocol was designated as non-human subject research by the Institutional Review Board.

Body Mass Index (BMI) Extraction and Categorization

BMI was calculated as weight in kilograms divided by height in meters squared, where both weight and height were measured at a single encounter. All measured BMI values were extracted for each adult individual (9,573,624 BMI observations), with BMI data obtained during pregnancy (649,442 observations) or with clinically implausible values (less than 10 kg/m^2 or greater than 70 kg/m^2, 6316 observations) excluded. The median BMI for each individual was classified into one of 6 BMI categories, as defined by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO), including underweight (<18.5 kg/m^2), normal (18.5-24.9 kg/m^2), overweight (25.0-29.9 kg/m^2), and obesity class 1 (30.0-34.9 kg/m^2), class 2 (35.0-39.9 kg/m^2), and class 3 (≥40.0 kg/m^2).(29)

Clinical Cohort BMI Phenome-wide Association Study

All distinct International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes from each individuals' record were captured and translated into PheWAS codes (phecodes), a hierarchical classification system for ICD-9-CM codes.(27, 42) A minimum of 2 instances of a matching ICD-9 code on separate days was required to be translated to a phecode using PheWAS code map version 1.2. For all PheWAS analyses in this study, as many phenotypes occur rarely, we analyzed only those that occurred in a minimum of 20 cases.

PheWAS was performed using logistic regression models adjusted for age at last BMI recorded, sex, and self-reported race to determine the association of BMI categories with phenotypes. Using categorical BMI, effect sizes are determined by comparison to those individuals with BMI in the normal range. We also used mean BMI as the predictor in the PheWAS model to calculate effect sizes per standard deviation (SD)-difference in BMI. All PheWAS analyses were performed using the PheWAS package for R statistical software, version 3.4.3(44) and using PheWAS code map version 1.2.(43) Two-sided p-value $<5.6\text{x}10^{-6}$ was considered as statistically significant using Bonferroni correction for multiple comparisons.

Sensitivity Analysis for Unmeasured Confounders

As a sensitivity analysis, a tipping point analysis(30) was performed to assess the impact of unmeasured confounders on the conclusions. For diseases in which abnormal BMI was statistically significant, we calculated the minimum effect size of a hypothetical binary unmeasured confounder such that the upper (if OR<1) or lower (if OR>1) bound of the (1-5.6x10^{-6}) x100% confidence interval [CI] for BMI would cross 1 for various conditions.

Genomic Cohort Analyses

Data for the genomic analyses were derived from two separate cohorts, the first being the eMERGE Consortium, a national network organized and funded by the National Human Genome Research Institute (NHGRI).(45) eMERGE combines DNA biorepositories with EHRs for large scale, high-throughput genetic research. Both the genomic and phenomic data (ICD9 diagnosis codes and demographics) were coalesced into a central repository. Of the eMERGE cohort, 19,590 (30.1%) individuals were from Vanderbilt University Medical Center and also likely contributed data to the clinical cohort, although deidentification limits the ability to confirm overlap. The second genomic cohort was derived from a maximal subset of unrelated UK Biobank participants with both genomic and phenomic data available.

Genotyping and Imputation in the eMERGE Mendelian Randomization Cohort

The eMERGE population in this study consisted of individuals from institutions contributing data to the eMERGE network phases I-III (Supplemental Table 1). Inclusion criteria were age ≥18 years with extant genome-wide genotyping data and ICD-9-CM codes. The eMERGE Consortium has unified genetic results from 12 different sites across 78 genotype array batches through imputation using the Michigan Imputation Server (46) and Haplotype Reference Consortium (HRC1.1).(47) This pipeline has resulted in an imputed genome wide set of approximately 40 million single nucleotide variant marker allele doses down to 0.1% minor allele frequency (MAF). Genotype array files were referenced to build 37 genome position using the forward genome strand. Quality control included filtering for sample missingness <2.0% and SNP missingness <2.0% in the preprocessing of the data before imputation. For duplicate samples on differing arrays, the sample with the most genotyped variants for that subject was selected for the merged dataset. Principal component analysis (PCA) using the first 10 principal components was performed to determine genetic ancestry using PLINK(48) with variants having >5% MAF. Single nucleotide variants with a missing rate >10% or not meeting the linkage disequilibrium threshold $r^2 < 0.7$ were excluded in PCAs. We performed identity by descent (IBD) analysis to identify related individuals. One individual from suspected monozygotic twins or duplicates were excluded randomly. Subject relatedness was determined using probability of zero alleles IBD (Z0) < 0.83 and the probability of having one allele IBD (Z1) > 0.10 to capture first through third-degree relatives. The oldest family member from each family was included in the cohort analysis. Minimum mean imputation r^2 was 0.83 with a mean r^2 of 0.95 across imputed SNPs.

UK Biobank sample selection

A maximal subset of unrelated UK Biobank participants after application of quality control was selected as detailed in Bycroft, et al, Supplemental section 3.3.2.(49) Individuals in this subset where chosen to have no other related individuals within 3 degrees within the subset, to have genotyping missingness < 2%, to have no mismatch between genetically inferred and reported sex, and to not be outliers for heterozygosity or genotype missingness. We additionally removed individuals without a BMI measurement at the time of enrollment, as well as those who revoked consent after enrollment. This left 405,432 UK Biobank participants for analysis.

Construction of Obesity Polygenic Risk Scores (PRS)

The limited PRS in each genomic cohort was calculated from 97 SNPs (Supplemental Table 3) associated with BMI at genome-wide significance in a prior meta-analysis of genome-wide association studies conducted by the Genetic Investigation of ANthropometric Traits (GIANT) Consortium.(11) The 97-SNP polygenic score was computed for each participant by multiplying the effect estimate at each allele by the genetic dosage of the effect allele, summing the values across all SNPs for each participant.

The genome-wide PRS was computed for each participant with the same procedure, using the best performing LDPred-adjusted values—from a model built assuming that 3% of variants are causal, and constructed with 2,100,302 variants—as described previously.(10) In this approach, each variant's posterior mean effect is calculated based on the prior effect estimate and a shrinkage based on the variant's correlation structure with other variants from the reference population.(50) As all sets were imputed to the same reference standard, we were able to use all SNPs for calculation of the genome-wide PRS. Pearson correlation was used to compare the correlation between the PRS and observed BMI. The BMI variance explained (adjusted R^2) by the associated SNPs was calculated with individual-level genotype and phenotype data using linear regression models adjusted for site, age, sex, and the first 10 principal components.

eMERGE PheWAS

To calculate effect estimates for genetically-determined BMI on disease phenotypes, PheWAS was performed as described above using logistic regression models, adjusted for site, age, sex, and the first 10 principal components. For phenotypes already passing a Bonferroni significance threshold for association with class 3 obesity in the primary cohort, a false discovery rate (FDR) significance threshold <0.05 was used to assess for replication of obesity associations with the genomic score.(44) Effect estimates for the 97-SNP score are reported per standard deviation (SD) difference in BMI (derived from beta estimates and SD of 4.8 kg/m^2 in a prior GIANT cohort of 449,472 individuals).

The genome-wide PRS was scaled to a mean of 0 and SD of 1 prior to PheWAS analysis. Effect estimates were compared using correlation coefficient analysis to determine the similarity between clinical and genomic effect sizes.

UK Biobank PheWAS

ICD10 codes for hospitalizations were reported by the UK Biobank. These were translated into phecodes using mappings described previously.(51) For each phecode, a logistic regression model was computed, predicting the presence or absence of a phecode as a function of the polygenic score, sex, age at enrollment, the UK Biobank genotyping array, and the first ten principal components of ancestry. For phenotypes already passing a Bonferroni significance threshold for association with class 3 obesity in the primary cohort as well as showing significant association in the eMERGE genomic cohort, an FDR significance threshold <0.05 was used to assess for replication of obesity associations with the risk score.(44) These models were computed separately for the 97-SNP score and the genome-wide polygenic score. PRS was scaled to a mean of 0 and SD of 1 prior to PheWAS analysis. Effect estimates between cohorts were compared using Pearson correlation analysis.

Prediction of Burden of Disease Attributable to Extremes of BMI

To estimate the disease phenotypes attributable to BMI in the clinical cohort, for phenotypes showing association to obesity both clinically and genetically, we performed logistic regression, adjusted for age, sex, self-reported race, and BMI category, and then normalized BMI for the individuals with class 1-3 obesity by changing their BMI category to be normal and calculating the predicted probability of events. The positive differences between the number of phenotypes in the clinical cohort and the number of predicted events after normalizing the BMI were reported as the population risk attributable to obesity.

Acknowledgements

JR Robinson receives support by the 5T15LM007450 training grant from the National Library of Medicine. Support for the research and personnel was provided by the R01LM010685 grant from the National Library of Medicine and R01GM114128 from the National Institutes of Health. The eMERGE sites were funded through several series of grants from the National Human Genome Research Institute:

U01HG8657, U01HG006375, U01HG004610 (Kaiser Permanente Washington/University of Washington); U01HG8685 (Brigham and Women’s Hospital); U01HG8672, U01HG006378, U01HG004608 (Vanderbilt University Medical Center); U01HG8666, U01HG006828 (Cincinnati Children’s Hospital Medical Center); U01HG6379, U01HG04599 (Mayo Clinic); U01HG8679, U01HG006382 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG8684, U01HG006830 (Children’s Hospital of Philadelphia); U01HG8673, U01HG006388, U01HG004609 (Northwestern University); U01HG8676 (Partners Healthcare/Broad Institute); U01HG8664 (Baylor College of Medicine); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG8701, U01HG006385, U01HG04603 (Vanderbilt University Medical Center serving as the Coordinating Center); eMERGE Genotyping Centers were also funded through U01HG004438 (CIDR) and U01HG004424 (the Broad Institute). Vanderbilt University Medical Center’s Synthetic Derivative and BioVU are supported by institutional funding and by the CTSA grant ULTR000445 from NCATS/NIH. AV Khera received support from grants 1K08HG010155 and 5UM1HG008895 from the National Human Genome Research Institute, a Hassenfeld Scholar Award from Massachusetts General Hospital, a Merkin Institute Fellowship from the Broad Institute of MIT and Harvard, and a sponsored research agreement from IBM Research.

References

1. Ogden CL, Carroll MD, Kit BK, Flegal KM. 2014. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA*. 311(8):806–14

2. GBD 2015 Obesity Collaborators, Afshin A, Forouzanfar MH, Reitsma MB, Sur P, et al. 2017. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N. Engl. J. Med.* 377(1):13–27

3. NCD Risk Factor Collaboration (NCD-RisC). 2016. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *Lancet Lond. Engl.* 387(10026):1377–96

4. Global BMI Mortality Collaboration null, Di Angelantonio E, Bhupathiraju S, Wormser D, Gao P, et al. 2016. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet Lond. Engl.* 388(10046):776–86

5. Prospective Studies Collaboration, Whitlock G, Lewington S, Sherliker P, Clarke R, et al. 2009. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet Lond. Engl.* 373(9669):1083–96

6. Emerging Risk Factors Collaboration, Wormser D, Kaptoge S, Di Angelantonio E, Wood AM, et al. 2011. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *Lancet Lond. Engl.* 377(9771):1085–95

7. Yeo GS, Farooqi IS, Aminian S, Halsall DJ, Stanhope RG, O’Rahilly S. 1998. A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nat. Genet.* 20(2):111–12

8. Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, et al. 1997. Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature*. 387(6636):903–8

9. O’Rahilly S. 2009. Human genetics illuminates the paths to metabolic disease. *Nature*. 462(7271):307–14

10. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, et al. 2019. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*. 177(3):587-596.e9

11. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 518(7538):197–206

12. Nordestgaard BG, Palmer TM, Benn M, Zacho J, Tybjaerg-Hansen A, et al. 2012. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. *PLoS Med.* 9(5):e1001212

13. Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, et al. 2017. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. *JAMA Cardiol.* 2(8):882–89

14. Chatterjee NA, Giulianini F, Geelhoed B, Lunetta KL, Misialek JR, et al. 2017. Genetic Obesity and the Risk of Atrial Fibrillation: Causal Estimates from Mendelian Randomization. *Circulation*. 135(8):741–54

15. Stender S, Nordestgaard BG, Tybjaerg-Hansen A. 2013. Elevated body mass index as a causal risk factor for symptomatic gallstone disease: a Mendelian randomization study. *Hepatol. Baltim. Md*. 58(6):2133–41

16. Panoutsopoulou K, Metrustry S, Doherty SA, Laslett LL, Maciewicz RA, et al. 2014. The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a Mendelian randomisation study. *Ann. Rheum. Dis.* 73(12):2082–86

17. Lindström S, Germain M, Crous-Bou M, Smith EN, Morange P-E, et al. 2017. Assessing the causal relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum. Genet.* 136(7):897–902

18. Vimaleswaran KS, Berry DJ, Lu C, Tikkanen E, Pilz S, et al. 2013. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. *PLoS Med.* 10(2):e1001383

19. Huang Y, Xu M, Xie L, Wang T, Huang X, et al. 2016. Obesity and peripheral arterial disease: A Mendelian Randomization analysis. *Atherosclerosis*. 247:218–24

20. Mokry LE, Ross S, Timpson NJ, Sawcer S, Davey Smith G, Richards JB. 2016. Obesity and Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Med.* 13(6):e1002053

21. Gianfrancesco MA, Glymour MM, Walter S, Rhead B, Shao X, et al. 2017. Causal Effect of Genetic Variants Associated With Body Mass Index on Multiple Sclerosis Susceptibility. *Am. J. Epidemiol.* 185(3):162–71

22. Thrift AP, Shaheen NJ, Gammon MD, Bernstein L, Reid BJ, et al. 2014. Obesity and risk of esophageal adenocarcinoma and Barrett's esophagus: a Mendelian randomization study. *J. Natl. Cancer Inst.* 106(11):

23. Jarvis D, Mitchell JS, Law PJ, Palin K, Tuupanen S, et al. 2016. Mendelian randomisation analysis strongly implicates adiposity with risk of developing colorectal cancer. *Br. J. Cancer*. 115(2):266–72

24. Dixon SC, Nagle CM, Thrift AP, Pharoah PD, Pearce CL, et al. 2016. Adult body mass index and risk of ovarian cancer by subtype: a Mendelian randomization study. *Int. J. Epidemiol.* 45(3):884–95

25. Millard LAC, Davies NM, Timpson NJ, Tilling K, Flach PA, Davey Smith G. 2015. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci. Rep.* 5:16645

26. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, et al. 2020. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* 26(4):549–57

27. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* 26(9):1205–10

28. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86(4):560–72

29. *Defining Adult Overweight and Obesity | Overweight & Obesity | CDC*. www.cdc.gov

30. Lin DY, Psaty BM, Kronmal RA. 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 54(3):948–63

31. Dale CE, Fatemifar G, Palmer TM, White J, Prieto-Merino D, et al. 2017. Causal Associations of Adiposity and Body Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis. *Circulation*. 135(24):2373–88

32. Skaaby T, Taylor AE, Thuesen BH, Jacobsen RK, Friedrich N, et al. 2018. Estimating the causal effect of body mass index on hay fever, asthma and lung function using Mendelian randomization. *Allergy*. 73(1):153–64

33. Hägg S, Fall T, Ploner A, Mägi R, Fischer K, et al. 2015. Adiposity as a cause of cardiovascular disease: a Mendelian randomization study. *Int. J. Epidemiol.* 44(2):578–86

34. Censin JC, Nowak C, Cooper N, Bergsten P, Todd JA, Fall T. 2017. Childhood adiposity and risk of type 1 diabetes: A Mendelian randomization study. *PLoS Med.* 14(8):e1002362

35. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, et al. 2018. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* 9(1):224

36. Zhu Z, Guo Y, Shi H, Liu C-L, Panganiban RA, et al. 2020. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J. Allergy Clin. Immunol.* 145(2):537–49

37. Robinson JR, Carroll RJ, Bastarache L, Chen Q, Mou Z, et al. 2020. Association of Genetic Risk of Obesity with Postoperative Complications Using Mendelian Randomization. *World J. Surg.* 44(1):84–94

38. *Bariatric surgery: a systematic review and meta-analysis. - PubMed - NCBI*. www.ncbi.nlm.nih.gov

39. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51(4):584–91

40. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, et al. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* 84(3):362–69

41. Robinson JR, Wei W-Q, Roden DM, Denny JC. 2018. Defining Phenotypes from Clinical Data to Drive Genomic Research. *Annu Rev Biomed Data Sci*. In Publication:

42. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS One*. 12(7):e0175508

43. Carroll RJ, Bastarache L, Denny JC. 2014. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma. Oxf. Engl.* 30(16):2375–76

44. *R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* www.r-project.org

45. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, et al. 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 15(10):761–71

46. *Michigan Imputation Server*. https://imputationserver.sph.umich.edu

47. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48(10):1279–83

48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81(3):559–75

49. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 562(7726):203–9

50. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, et al. 2015. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* 97(4):576–92

51. Wu P, Gifford A, Meng X, Li X, Campbell H, et al. 2019. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* 7(4):e14325

Supplemental Table 1. Demographics for Clinical and Genotyped (eMERGE and UK Biobank) Cohorts

Clinical Variable	**Clinical cohort (n = 736,726)**	**eMERGE cohort (n = 65,174)**	**UK Biobank cohort (n = 405,432)**
Age, median (IQR), years	49.0 (33.0 – 63.0)	67.0 (51.0 – 79.0)	58.2 (50.5 – 63.6)
Sex, No. (%)			
Female	434,266 (58.9)	35,997 (55.2)	218,818 (54.0)
Unknown	57 (0.1)	0	0
Race (Vanderbilt) or Genetic Ancestry (eMERGE), No. (%)			
White/European ancestry	553,368 (75.1)	52,760 (81.0)	390,459 (96.3)
Black/African ancestry	70,409 (9.6)	11,323 (17.4)	3,018 (0.7)
Asian	11,998 (1.6)	1,091 (1.7)	8,169 (2.0)
Other	18,332 (2.5)	0	3,786 (0.9)
Unknown	82,619 (11.2)	0	0
BMI (kg/m^2), median (IQR)	27.3 (23.6 – 32.0)	27.6 (23.9 – 32.1)	26.7 (24.1 – 29.9)
BMI (kg/m^2), mean (SD)	28.5 (7.0)	28.6 (7.0)	27.4 (4.8)
BMI category, No. (%)			
Underweight (<18.5)	15,509 (2.1)	1,961 (3.7)	2,091 (0.5)
Normal (18.5 – 24.9)	240,676 (32.7)	14,926 (28.4)	132,710 (32.7)
Overweight (25.0 – 29.9)	229,630 (31.2)	17,088 (32.5)	172.314 (42.5)
Obesity Class 1 (30.0 – 34.9)	135,488 (18.4)	10,425 (19.8)	70,662 (17.4)
Obesity Class 2 (35.0 – 39.9)	64,539 (8.8)	4,802 (9.1)	19,961 (4.9)
Obesity Class 3 (≥40.0)	50,873 (6.9)	3,407 (6.5)	7,694 (1.9)
No BMI reported	0	12,565 (19.3)	0

Abbreviations: BMI, body mass index; IQR, interquartile range; SD, standard deviation

Supplemental Table 2. Sites and Number of Unique Individuals Contributing Adult Data for Genotyped eMERGE Cohort

Site	Number of unique subjects, n (%) (Total n = 65,174)
Boston Children's Hospital	252 (0.4%)
Children's Hospital of Philadelphia	4,649 (7.1%)
Cincinnati Children's Hospital Medical Center	1,331 (2.0%)
Columbia University	1,680 (2.6%)
Geisinger	2,772 (4.3%)
Harvard University	9,689 (14.9%)
Kaiser Permanente Washington with the University of Washington and the Fred Hutchinson Cancer Research Center	3,197 (4.9%)
Marshfield Clinic	3,683 (5.7%)
Mayo Clinic	8,199 (12.6%)
Mount Sinai	5,701 (8.7%)
Northwestern University	4,431 (6.8%)
Vanderbilt University	19,590 (30.1%)

Supplemental Table 3. Single Nucleotide Polymorphisms for Obesity 97-SNP Polygenic Risk Score Calculation [a]

SNP	Chrom	Pos1	Pos2	Gene	Effect Allele	Beta [b]
rs1558902	16	53803574	53803574	FTO	A	0.0809
rs6567160	18	57829135	57829135	MC4R	C	0.0562
rs13021737	2	632348	632348	TMEM18	G	0.0604
rs10938397	4	45182527	45182527	GNPDA2, GABRG1	G	0.0399
rs543874	1	177889480	177889480	SEC16B	G	0.0497
rs2207139	6	50845490	50845490	TFAP2B	G	0.0448
rs11030104	11	27684517	27684517	BDAF	A	0.0416
rs3101336	1	72751185	72751185	NEGR1	C	0.0319
rs7138803	12	50247468	50247468	BCDIN3D; FAIM2	A	0.032
rs10182181	2	25150296	25150296	ADCY3; POMC; NCOA1; SH2B1; ABOBR	G	0.0309
rs3888190	16	28889486	28889486	ATXN2L; SBK1; SULT1A2; TUFM	A	0.0311
rs1516725	3	185824004	185824004	E7V5	C	0.0448
rs12446632	16	19935389	19935389	GPRC5B; IQCK	G	0.0399
rs2287019	19	46202172	46202172	QPCTL; GIPR	C	0.0354
rs16951275	15	68077168	68077168	M4P2K5; LBXCOR1	T	0.0304
rs3817334	11	47650993	47650993	MTCH2; C1QTNF4; SPI1; CELF1	T	0.0256
rs2112347	5	75015242	75015242	PCO5; HMGCR; COL4A3BP	T	0.0254
rs12566985	1	75002193	75002193	FPGT-TNNI3K	G	0.0237
rs3810291	19	47569003	47569003	ZC3H4	A	0.0285
rs7141420	14	79899454	79899454	NRXN3	T	0.0227
rs13078960	3	85807590	85807590	CADM2	G	0.029
rs10968576	9	28414339	28414339	LINGO2	G	0.0247
rs17024393	1	110154688	110154688	GNAT2; AMPD2	C	0.0611
rs657452	1	49589847	49589847	AGBL4	A	0.0227
rs12429545	13	54102206	54102206	OLFM4	A	0.0324
rs12286929	11	115022404	115022404	CADM1	G	0.0211
rs13107325	4	103188709	103188709	SLC39A8	T	0.0472
rs11165643	1	96924097	96924097	PTBP2	T	0.0221
rs7903146	10	114758349	114758349	TCF7L2	C	0.0235
rs10132280	14	25928179	25928179	STXBP6	C	0.0221
rs17405819	8	76806584	76806584	HNF4G	T	0.0221
rs6091540	20	51087862	51087862	ZFP64	C	0.0185
rs1016287	2	59305625	59305625	LINC01122	T	0.0228
rs4256980	11	8673939	8673939	TRIM66; TUB	G	0.0205
rs17094222	10	102395440	102395440	HIF1AN	C	0.0249

rs12401738	1	78446761	78446761	FUBP1; USP33	A	0.0202
rs7599312	2	213413231	213413231	ERBB4	G	0.0214
rs2365389	3	61236462	61236462	FHIT	C	0.0195
rs205262	6	34563164	34563164	C6orf106; SNRPC	G	0.021
rs2820292	1	201784287	201784287	NAV1	C	0.0181
rs12885454	14	29736838	29736838	PRKD1	C	0.0204
rs9641123	7	93197732	93197732	CALCR; has-miR-653	C	0.0193
rs12016871	13	28017782	28017782	MTIF3; GTF3A	T	0.0298
rs16851483	3	141275436	141275436	RASA2	T	0.0478
rs1167827	7	75163169	75163169	HIP1; PMS2L3; PMS2P5; WBSCR16	G	0.02
rs758747	16	3627358	3627358	NLRC3	T	0.0225
rs1928295	9	120378483	120378483	TLR4	T	0.0182
rs9925964	16	31129895	31129895	KAT8; ZNF646; VKORC1; ZNF668; STX1B; FBXL19	A	0.0198
rs11126666	2	26928811	26928811	KCNK3	A	0.0201
rs2650492	16	28333411	28333411	SBK1; APOBR	A	0.0205
rs6804842	3	25106437	25106437	RARB	G	0.0183
rs12940622	17	78615571	78615571	RPTOR	G	0.0183
rs7164727	15	73093991	73093991	LOC100287559; BBS4	T	0.0189
rs11847697	14	30515112	30515112	PRKD1	T	0.0374
rs4740619	9	15634326	15634326	C9or93	T	0.017
rs492400	2	219349752	219349752	PLCD4; CYP27A1; USP37; TTLL4; STK36; ZNF142; RQCD1	C	0.015
rs13191362	6	163033350	163033350	PARK2	A	0.0285
rs3736485	15	51748610	51748610	SCG3; DMXL2	A	0.016
rs17001654	4	77129568	77129568	NUP54; SCARB2	G	0.0304
rs11191560	10	104869038	104869038	NT5C2; CYP17A1; SFXN2	C	0.031
rs2080454	16	49062590	49062590	CLBLN1	C	0.0171
rs7715256	5	153537893	153537893	GALNT10	G	0.0168
rs2176040	2	227092802	227092802	LOC646736; IRS1	A	0.0147
rs1528435	2	181550962	181550962	UBE2E3	T	0.0175
rs2075650	19	45395619	45395619	TOMM40; APO3; APOC1	A	0.0256
rs1000940	17	5283252	5283252	RABEP1	G	0.0184
rs2033529	6	40348653	40348653	TDRG1; LRFN2	G	0.0183
rs11583200	1	50559820	50559820	ELAVL4	C	0.0174
rs7239883	18	40147671	40147671	LOC284260; RIT2	G	0.0152
rs2836754	21	40291740	40291740	ETS2	C	0.0169
rs9400239	6	108977663	108977663	FOXO3; HSS00296402	C	0.0173
rs10733682	9	129460914	129460914	LMX1B	A	0.0188
rs11688816	2	63053048	63053048	EHBP1	G	0.0148
rs11057405	12	122781897	122781897	CLIP1	G	0.0304

rs9914578	17	2005136	2005136	SMG6; N29617	G	0.0201
rs977747	1	47684677	47684677	TAL1	T	0.0168
rs2121279	2	143043285	143043285	LRP1B	T	0.0242
rs29941	19	34309532	34309532	KCTD15	G	0.0177
rs11727676	4	145659064	145659064	HHIP	T	0.0365
rs3849570	3	81792112	81792112	GBE1	A	0.0183
rs9374842	6	120185665	120185665	LOC285762	T	0.0196
rs6477694	9	111932342	111932342	EPB41L4B; C9orf4	C	0.0169
rs4787491	16	30015337	30015337	MAPK3; KCTD13; INO80E; TAOK2; YPEL3; DOC2A; FAM57B	G	0.0151
rs1441264	13	79580919	79580919	MIR548A2	A	0.0172
rs7899106	10	87410904	87410904	GRID1	G	0.0379
rs2176598	11	43864278	43864278	HSD17B12	T	0.0185
rs2245368	7	76608143	76608143	PMS2L11	C	0.0288
rs17203016	2	208255518	208255518	CREB1; KLF7	G	0.0211
rs17724992	19	18454825	18454825	GDF15; PGPEP1	A	0.0196
rs7243357	18	56883319	56883319	GRP	T	0.0219
rs16907751	8	81375457	81375457	ZBRB10	C	0.0326
rs1808579	18	21104888	21104888	NPC1; C18orf8	C	0.016
rs13201877	6	137675541	137675541	IFNGR1; OLIG3	G	0.0236
rs2033732	8	85079709	85079709	RALYL	C	0.0176
rs9540493	13	66205704	66205704	MIR548X2; PCDH9	A	0.0182
rs1460676	2	164567689	164567689	FIGN	C	0.0209
rs6465468	7	95169514	95169514	ASB4	T	0.016

Abbreviations: SNP, single nucleotide polymorphism

[a] Information shown in the table includes significant associations in prior GWAS meta-analysis at 5 x10^{-8} (Locke et al, Nature, 2015).

[b] Beta-coefficients are derived from the all-ancestry meta-analysis.

Supplemental Table 4. Phenotype Associations of 97-SNP PRS for Obesity in eMERGE that Replicate Associations with Class 3 Obesity [a]

Phenotype	Cases	Controls	Beta	SE	OR	p-value
Morbid obesity	7275	57899	1.80	0.08	6.08	5.67x10^{-112}
Obesity	16524	48650	1.25	0.06	3.48	1.51x10^{-102}
Overweight, obesity and other hyperalimentation	21061	44113	1.10	0.05	3.02	6.16x10^{-93}
Bariatric surgery	1892	63282	2.06	0.15	7.85	9.47x10^{-45}
Type 2 diabetes	17364	47810	0.69	0.06	1.99	7.35x10^{-32}
Sleep apnea	9819	55355	0.81	0.07	2.24	2.59x10^{-31}
Diabetes mellitus	18620	46554	0.67	0.06	1.95	6.01x10^{-31}
Eating disorder	1465	63709	1.86	0.17	6.44	3.32x10^{-29}
Hypertension	39206	25968	0.60	0.06	1.82	1.46x10^{-23}
Obstructive sleep apnea	7599	57575	0.76	0.08	2.13	9.66x10^{-23}
Essential hypertension	38654	26520	0.58	0.06	1.79	1.40x10^{-22}
Gastrointestinal complications	2981	62193	1.00	0.12	2.72	1.51x10^{-17}
Other chronic nonalcoholic liver disease	5029	60145	0.77	0.09	2.17	5.32x10^{-17}
Chronic liver disease and cirrhosis	5471	59703	0.74	0.09	2.10	5.96x10^{-17}
Insulin pump user	4489	60685	0.77	0.10	2.15	1.62x10^{-14}
Edema	13670	51504	0.44	0.06	1.56	2.02x10^{-12}
Type 1 diabetes	4985	60189	0.64	0.09	1.90	1.07x10^{-11}
Type 2 diabetes with neurological manifestations	4352	60822	0.67	0.10	1.96	3.62x10-^{11}
Localized adiposity	642	64532	1.63	0.25	5.12	4.54x10^{-11}
Other disorders of intestine	7062	58112	0.51	0.08	1.67	8.76x10^{-11}
Acute renal failure	9186	55988	0.47	0.07	1.61	1.20x10^{-10}
Osteoarthrosis	22861	42313	0.37	0.06	1.45	1.68x10^{-10}
Osteoarthrosis NOS	18362	46812	0.37	0.06	1.45	6.23x10^{-10}
Respiratory failure, insufficiency, arrest	6684	58490	0.49	0.08	1.63	3.04x10^{-09}
Osteoarthrosis, localized, primary	6124	59050	0.51	0.09	1.66	3.89x10^{-09}
Renal failure	15129	50045	0.36	0.06	1.43	6.86x10^{-09}
Shortness of breath	19815	45359	0.32	0.06	1.38	9.51x10^{-9}
Polyneuropathy in diabetes	3597	61577	0.63	0.11	1.87	1.14x10^{-8}
Encounter for long-term (current) use of anticoagulants	10310	54864	0.40	0.07	1.48	1.66x10^{-8}
Coronary atherosclerosis	17803	47371	0.35	0.06	1.42	1.77x10^{-8}
Cardiomegaly	9268	55906	0.41	0.07	1.51	1.90x10^{-8}
Congestive heart failure (CHF) NOS	9979	55195	0.40	0.07	1.49	5.38x10^{-8}
Congestive heart failure; nonhypertensive	11702	53472	0.37	0.07	1.45	5.75x10^{-8}
Respiratory failure	4502	60672	0.52	0.10	1.68	1.38x10^{-7}
Chronic venous insufficiency [CVI]	3708	61466	0.57	0.11	1.76	1.51x10^{-7}
Pulmonary heart disease	7098	58076	0.42	0.08	1.52	1.59x10^{-7}

Heart failure with reduced EF [Systolic or combined heart failure]	4181	60993	0.54	0.10	1.72	$1.83\text{x}10^{-7}$
Atrial fibrillation and flutter	10191	54983	0.37	0.07	1.45	$3.82\text{x}10^{-7}$
Type 2 diabetes with renal manifestations	3453	61721	0.57	0.11	1.76	$5.05\text{x}10^{-7}$
Panniculitis	461	64713	1.46	0.29	4.30	$5.28\text{x}10^{-7}$
Atrial fibrillation	9935	55239	0.37	0.07	1.45	$5.62\text{x}10^{-7}$
Ischemic Heart Disease	20357	44817	0.30	0.06	1.35	$5.73\text{x}10^{-7}$
Gastrojejunal ulcer	271	64903	1.86	0.38	6.42	$7.39\text{x}10^{-7}$
Cardiac pacemaker/device in situ	4020	61154	0.52	0.10	1.68	$7.96\text{x}10^{-7}$
Other venous embolism and thrombosis	7287	57887	0.39	0.08	1.47	$8.45\text{x}10^{-7}$
Chronic ulcer of skin	5649	59525	0.44	0.09	1.55	$8.75\text{x}10^{-7}$
Non-healing surgical wound	716	64458	1.15	0.23	3.16	$9.72\text{x}10^{-7}$
Arthropathy NOS	9057	56117	0.36	0.07	1.43	$1.06\text{x}10^{-6}$
Superficial cellulitis and abscess	13672	51502	0.30	0.06	1.35	$1.07\text{x}10^{-6}$
Type 2 diabetes with ophthalmic manifestations	2447	62727	0.65	0.13	1.91	$1.11\text{x}10^{-6}$
Incisional hernia	1620	63554	0.76	0.16	2.14	$1.43\text{x}10^{-6}$
Septicemia	5395	59779	0.43	0.09	1.53	$2.52\text{x}10^{-6}$
Chronic pulmonary heart disease	4900	60274	0.44	0.10	1.56	$3.10\text{x}10^{-6}$
Postoperative infection	3709	61465	0.49	0.11	1.64	$3.14\text{x}10^{-6}$
Other arthropathies	9647	55527	0.33	0.07	1.39	$3.31\text{x}10^{-6}$
Cholelithiasis	5031	60143	0.43	0.09	1.53	$4.68\text{x}10^{-6}$
Other disorders of the kidney and ureters	10875	54299	0.31	0.07	1.37	$4.80\text{x}10^{-6}$
Dysmetabolic syndrome X	1084	64090	0.88	0.19	2.40	$4.82\text{x}10^{-6}$
Cholelithiasis and cholecystitis	5727	59447	0.40	0.09	1.49	$5.37\text{x}10^{-6}$
Pulmonary collapse; interstitial and compensatory emphysema	10304	54870	0.32	0.07	1.38	$5.45\text{x}10^{-6}$
Other chronic ischemic heart disease, unspecified	8263	56911	0.35	0.08	1.43	$6.30\text{x}10^{-6}$
Osteoarthritis; localized	12113	53061	0.30	0.07	1.35	$7.76\text{x}10^{-6}$
Chronic ulcer of leg or foot	3479	61695	0.49	0.11	1.64	$8.97\text{x}10^{-6}$
Cellulitis and abscess of leg, except foot	4227	60947	0.45	0.10	1.57	$9.00\text{x}10^{-6}$
Spinal stenosis of lumbar region	4915	60259	0.42	0.10	1.52	$1.17\text{x}10^{-5}$
Cardiac pacemaker in situ	3163	62011	0.51	0.12	1.66	$1.35\text{x}10^{-5}$
Chronic renal failure [CKD]	11487	53687	0.30	0.07	1.34	$1.43\text{x}10^{-5}$
Diabetic retinopathy	2674	62500	0.55	0.13	1.74	$1.68\text{x}10^{-5}$
Dependence on respirator [Ventilator] or supplemental oxygen	1443	63731	0.71	0.17	2.04	$2.20\text{x}10^{-5}$
Cardiac defibrillator in situ	1877	63297	0.63	0.15	1.87	$2.47\text{x}10^{-5}$
Cardiac dysrhythmias	27872	37302	0.22	0.05	1.25	$2.54\text{x}10^{-5}$
Osteomyelitis	1739	63435	0.65	0.15	1.91	$2.54\text{x}10^{-5}$
Swelling of limb	10189	54985	0.29	0.07	1.34	$2.71\text{x}10^{-5}$
Asthma	11096	54078	0.28	0.07	1.32	$3.11\text{x}10^{-5}$

Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	7013	58161	0.34	0.08	1.41	$3.31x10^{-5}$
Ventral hernia	1887	63287	0.61	0.15	1.84	$3.54x10^{-5}$
Decubitus ulcer	2278	62896	0.56	0.14	1.75	$3.65x10^{-5}$
Pulmonary congestion and hypostasis	3727	61447	0.44	0.11	1.56	$4.22x10^{-5}$
Spinal stenosis	6326	58848	0.35	0.09	1.41	$4.74x10^{-5}$
Iron deficiency anemias	8970	56204	0.29	0.07	1.34	$5.62x10^{-5}$
Wheezing	3946	61228	0.42	0.10	1.52	$6.23x10^{-5}$
Intestinal malabsorption (non-celiac)	1168	64006	0.73	0.18	2.08	$6.73x10^{-5}$
Portal hypertension	724	64450	0.94	0.24	2.55	$6.86x10^{-5}$
Myocardial infarction	8758	56416	0.30	0.08	1.35	$7.23x10^{-5}$
Fluid overload	3837	61337	0.42	0.11	1.52	$7.53x10^{-5}$
Hyperlipidemia	34558	30616	0.22	0.06	1.25	$8.68x10^{-5}$
Hypertensive heart and/or renal disease	11631	53543	0.26	0.07	1.30	$1.01x10^{-4}$
Chronic tonsillitis and adenoiditis	1295	63879	0.70	0.18	2.02	$1.15x10^{-4}$
Abnormal weight gain	3618	61556	0.41	0.11	1.51	$1.26x10^{-4}$
Dermatophytosis of nail	6482	58692	0.32	0.09	1.38	$1.50x10^{-4}$
Barrett's esophagus	925	64249	0.78	0.21	2.19	$1.51x10^{-4}$
Degeneration of intervertebral disc	9858	55316	0.27	0.07	1.30	$1.72x10^{-4}$
Infection/inflammation of internal prosthetic device; implant; and graft	2852	62322	0.45	0.12	1.57	$1.93x10^{-4}$
Heart failure with preserved EF [Diastolic heart failure]	3819	61355	0.40	0.11	1.49	$1.95x10^{-4}$
Deep vein thrombosis	3311	61863	0.42	0.11	1.52	$2.15x10^{-4}$
Other symptoms of respiratory system	35281	29893	0.19	0.05	1.21	$2.22x10^{-4}$
Gout	4917	60257	0.35	0.10	1.42	$2.36x10^{-4}$
Polycystic ovaries	578	64596	0.99	0.27	2.70	$2.39x10^{-4}$
Sinoatrial node dysfunction (Bradycardia)	2471	62703	0.48	0.13	1.62	$2.42x10^{-4}$
Bacterial infection NOS	8419	56755	0.27	0.07	1.31	$2.74x10^{-4}$
Disorders of adrenal glands	2252	62922	0.49	0.13	1.63	$2.75x10^{-4}$
Disorders of lipoid metabolism	35419	29755	0.21	0.06	1.23	$2.78x10^{-4}$
Mixed hyperlipidemia	8620	56554	0.27	0.08	1.31	$2.92x10^{-4}$
Corns and callosities	3552	61622	0.40	0.11	1.49	$3.01x10^{-4}$
Acquired acanthosis nigricans	437	64737	1.16	0.32	3.19	$3.21x10^{-4}$
Paroxysmal ventricular tachycardia	3201	61973	0.41	0.12	1.51	$3.50x10^{-4}$
Atrial flutter	3193	61981	0.41	0.12	1.51	$3.55x10^{-4}$
Vitamin deficiency	10640	54534	0.24	0.07	1.28	$3.63x10^{-4}$
Diabetes type 2 with peripheral circulatory disorders	1446	63728	0.61	0.17	1.84	$3.74x10^{-4}$
Osteomyelitis, periostitis, and other infections involving bone	2244	62930	0.48	0.14	1.62	$3.92x10^{-4}$
Dermatophytosis	9887	55287	0.25	0.07	1.29	$4.05x10^{-4}$
Type 1 diabetes with neurological manifestations	1067	64107	0.69	0.20	1.99	$4.70x10^{-4}$

Other forms of chronic heart disease	6657	58517	0.29	0.08	1.34	4.71x10^{-4}
Chronic Kidney Disease, Stage IV	2754	62420	0.43	0.13	1.54	5.49x10^{-4}
Encounter for long-term (current) use of aspirin	6484	58690	0.29	0.09	1.34	5.59x10^{-4}
Ovarian dysfunction	807	64367	0.78	0.23	2.18	5.95x10^{-4}
Acquired foot deformities	7020	58154	0.28	0.08	1.32	6.02x10^{-4}
Iron deficiency anemias, unspecified or not due to blood loss	7682	57492	0.27	0.08	1.30	6.20x10^{-4}
Gout and other crystal arthropathies	5801	59373	0.30	0.09	1.35	7.04x10^{-4}
Intervertebral disc disorders	12542	52632	0.22	0.06	1.24	7.42x10^{-4}
Mood disorders	16737	48437	0.19	0.06	1.21	8.20x10^{-4}
Hypothyroidism NOS	10724	54450	0.23	0.07	1.25	9.37x10^{-4}
Arrhythmia (cardiac) NOS	9842	55332	0.23	0.07	1.26	9.41x10^{-4}
Dermatophytosis / Dermatomycosis	10419	54755	0.23	0.07	1.26	9.48x10^{-4}
Unstable angina (intermediate coronary syndrome)	5570	59604	0.30	0.09	1.35	9.55x10^{-4}
Thoracic or lumbosacral neuritis or radiculitis, unspecified	6140	59034	0.28	0.09	1.33	9.89x10^{-4}
Other dyspnea	19187	45987	0.18	0.06	1.20	0.001
Abnormal coagulation profile	2503	62671	0.42	0.13	1.52	0.001
Cholelithiasis with other cholecystitis	1878	63296	0.47	0.15	1.61	0.001
Dislocation	4381	60793	0.31	0.10	1.37	0.001
Vitamin D deficiency	8786	56388	0.23	0.07	1.26	0.002
Hypertensive chronic kidney disease	8311	56863	0.24	0.08	1.28	0.002
Flat foot	1773	63401	0.48	0.15	1.61	0.002
Hypertensive heart disease	5029	60145	0.30	0.10	1.35	0.002
Chronic obstructive asthma	2077	63097	0.45	0.14	1.56	0.002
Respiratory abnormalities	4122	61052	0.32	0.10	1.37	0.002
Acute pulmonary heart disease	2762	62412	0.38	0.12	1.46	0.002
Shock	1881	63293	0.46	0.15	1.58	0.002
Candidiasis of skin and nails	1357	63817	0.53	0.17	1.70	0.002
Depression	15718	49456	0.18	0.06	1.20	0.002
Nonspecific chest pain	28569	36605	0.16	0.05	1.17	0.002
Nonrheumatic aortic valve disorders	6091	59083	0.27	0.09	1.31	0.002
Ingrowing nail	2743	62431	0.38	0.12	1.46	0.002
Sleep disorders	13053	52121	0.19	0.06	1.21	0.003
Chronic kidney disease, Stage I or II	2369	62805	0.40	0.13	1.49	0.003
Hypoventilation	1481	63693	0.49	0.17	1.64	0.003
Gouty arthropathy	2091	63083	0.42	0.14	1.52	0.003
Heart failure NOS	4067	61107	0.30	0.10	1.35	0.004
Hereditary and idiopathic peripheral neuropathy	5570	59604	0.26	0.09	1.30	0.004
Arthropathy associated with neurological disorders	328	64846	1.01	0.35	2.76	0.004
Coagulation defects	6539	58635	0.24	0.08	1.27	0.004

Osteoarthrosis involving more than one site, but not specified as generalized	2277	62897	0.40	0.14	1.50	0.004
Liver abscess and sequelae of chronic liver disease	1339	63835	0.50	0.17	1.65	0.004
Hypothyroidism	11256	53918	0.19	0.07	1.21	0.005
Bursitis	3396	61778	0.31	0.11	1.37	0.005
Umbilical hernia	1708	63466	0.43	0.15	1.54	0.005
Cardiomyopathy	5291	59883	0.26	0.09	1.29	0.005
Calcaneal spur; Exostosis NOS	2988	62186	0.33	0.12	1.39	0.005
Pulmonary embolism and infarction, acute	2547	62627	0.35	0.13	1.43	0.005
Chronic pain	4649	60525	0.27	0.10	1.31	0.006
Spondylosis without myelopathy	11736	53438	0.19	0.07	1.20	0.006

Abbreviations: PRS, polygenic risk score; SE, standard error; OR, odds ratio

[a] Information shown in the table includes significant associations (Bonferroni corrected p-value of $p = 5.6 \times 10^{-6}$ compared to individuals with normal range BMI) with class 3 obese BMI that are replicated 97-SNP polygenic risk score for obesity (results significant to FDR-adjusted p-value 0.006). Logistic regression models are adjusted for age, sex, site, and 10 principal components.

Supplemental Table 5. Phenotype Associations with Obesity 97-SNP Polygenic Risk Score not Clinically Associated with Class 3 Obesity [a]

Phenotype	Cases	Controls	beta	SE	OR	p-value
Tobacco use disorder	15683	49491	0.34	0.06	1.41	$9.61x10^{-9}$
Hypopotassemia	9712	55462	0.38	0.07	1.46	$1.30x10^{-7}$
Other anemias	23113	42061	0.27	0.05	1.31	$7.06x10^{-7}$
Peripheral vascular disease, unspecified	7270	57904	0.40	0.08	1.49	$1.10x10^{-6}$
Electrolyte imbalance	16621	48553	0.28	0.06	1.32	$3.07x10^{-6}$
Respiratory insufficiency	1611	63563	0.70	0.16	2.01	$1.14x10^{-5}$
Debility unspecified	4141	61033	0.45	0.10	1.57	$1.22x10^{-5}$
Pain in limb	26025	39149	0.22	0.05	1.25	$1.85x10^{-5}$

Abbreviations: SE, standard error; OR, odds ratio

[a] Information shown in the table includes phenotypes having association with obesity 97-SNP PRS meeting Bonferroni significance threshold ($p = 2.8 \times 10^{-5}$) with no clinical association to class 3 obesity. Logistic regression models are adjusted for age, sex, site, and 10 principal components.

Supplemental Table 6. Phenotype Associations with Obesity 97-SNP Polygenic Risk Score in UK Biobank that Replicate Associations in eMERGE cohort and Class 3 Obesity [a]

Phenotype	Cases	Controls	Beta	SE	OR	p-value
Obesity	16524	48650	1.2	0.1	3.5	1.51x10^{-102}
Overweight, obesity and other hyperalimentation	21061	44113	1.1	0.1	3.0	6.16x10^{-93}
Type 2 diabetes	17364	47810	0.7	0.1	2.0	7.35x10^{-32}
Sleep apnea	9819	55355	0.8	0.1	2.2	2.59x10^{-31}
Diabetes mellitus	18620	46554	0.7	0.1	1.9	6.01x10^{-31}
Hypertension	39206	25968	0.6	0.1	1.8	1.46x10^{-23}
Essential hypertension	38654	26520	0.6	0.1	1.8	1.40x10^{-22}
Other chronic nonalcoholic liver disease	5029	60145	0.8	0.1	2.2	5.32x10^{-17}
Chronic liver disease and cirrhosis	5471	59703	0.7	0.1	2.1	5.96x10^{-17}
Edema	13670	51504	0.4	0.1	1.6	2.02x10^{-12}
Type 1 diabetes	4985	60189	0.6	0.1	1.9	1.07x10^{-11}
Type 2 diabetes with neurological manifestations	4352	60822	0.7	0.1	2.0	3.62x10^{-11}
Other disorders of intestine	7062	58112	0.5	0.1	1.7	8.76x10^{-11}
Acute renal failure	9186	55988	0.5	0.1	1.6	1.20x10^{-10}
Osteoarthrosis	22861	42313	0.4	0.1	1.4	1.68x10^{-10}
Osteoarthrosis NOS	18362	46812	0.4	0.1	1.4	6.23x10^{-10}
Respiratory failure, insufficiency, arrest	6684	58490	0.5	0.1	1.6	3.04x10^{-9}
Osteoarthrosis, localized, primary	6124	59050	0.5	0.1	1.7	3.89x10^{-9}
Renal failure	15129	50045	0.4	0.1	1.4	6.86x10^{-9}
Shortness of breath	19815	45359	0.3	0.1	1.4	9.51x10^{-9}
Coronary atherosclerosis	17803	47371	0.3	0.1	1.4	1.77x10^{-8}
Cardiomegaly	9268	55906	0.4	0.1	1.5	1.90x10^{-8}
Congestive heart failure	9979	55195	0.4	0.1	1.5	5.38x10^{-8}
Congestive heart failure; nonhypertensive	11702	53472	0.4	0.1	1.5	5.75x10^{-8}
Respiratory failure	4502	60672	0.5	0.1	1.7	1.38x10^{-7}
Chronic venous insufficiency	3708	61466	0.6	0.1	1.8	1.51x10^{-7}
Pulmonary heart disease	7098	58076	0.4	0.1	1.5	1.59x10^{-7}
Atrial fibrillation and flutter	10191	54983	0.4	0.1	1.4	3.82x10^{-7}
Type 2 diabetes with renal manifestations	3453	61721	0.6	0.1	1.8	5.05x10^{-7}
Ischemic Heart Disease	20357	44817	0.3	0.1	1.4	5.73x10^{-7}
Chronic ulcer of skin	5649	59525	0.4	0.1	1.6	8.75x10^{-7}
Superficial cellulitis and abscess	13672	51502	0.3	0.1	1.3	1.07x10^{-6}
Type 2 diabetes with ophthalmic manifestations	2447	62727	0.6	0.1	1.9	1.11x10^{-6}

Chronic pulmonary heart disease	4900	60274	0.4	0.1	1.6	3.10x10^{-6}
Postoperative infection	3709	61465	0.5	0.1	1.6	3.14x10^{-6}
Cholelithiasis	5031	60143	0.4	0.1	1.5	4.68x10^{-6}
Cholelithiasis and cholecystitis	5727	59447	0.4	0.1	1.5	5.37x10^{-6}
Pulmonary collapse; interstitial and compensatory emphysema	10304	54870	0.3	0.1	1.4	5.45x10^{-6}
Other chronic ischemic heart disease, unspecified	8263	56911	0.4	0.1	1.4	6.30x10^{-6}
Osteoarthritis; localized	12113	53061	0.3	0.1	1.3	7.76x10^{-6}
Chronic ulcer of leg or foot	3479	61695	0.5	0.1	1.6	8.97x10^{-6}
Spinal stenosis of lumbar region	4915	60259	0.4	0.1	1.5	1.17x10^{-5}
Chronic renal failure	11487	53687	0.3	0.1	1.3	1.43x10^{-5}
Diabetic retinopathy	2674	62500	0.6	0.1	1.7	1.68x10^{-5}
Cardiac dysrhythmias	27872	37302	0.2	0.1	1.3	2.54x10^{-5}
Ventral hernia	1887	63287	0.6	0.1	1.8	3.54x10^{-5}
Pulmonary congestion and hypostasis	3727	61447	0.4	0.1	1.6	4.22x10^{-5}
Spinal stenosis	6326	58848	0.3	0.1	1.4	4.74x10^{-5}
Iron deficiency anemias	8970	56204	0.3	0.1	1.3	5.62x10^{-5}
Myocardial infarction	8758	56416	0.3	0.1	1.4	7.23x10^{-5}
Hyperlipidemia	34558	30616	0.2	0.1	1.2	8.68x10^{-5}
Hypertensive heart and/or renal disease	11631	53543	0.3	0.1	1.3	1.01x10^{-4}
Barrett's esophagus	925	64249	0.8	0.2	2.2	1.51x10^{-4}
Degeneration of intervertebral disc	9858	55316	0.3	0.1	1.3	1.72x10^{-4}
Other symptoms of respiratory system	35281	29893	0.2	0.1	1.2	2.22x10^{-4}
Gout	4917	60257	0.4	0.1	1.4	2.36x10^{-4}
Polycystic ovaries	578	64596	1.0	0.3	2.7	2.39x10^{-4}
Bacterial infection NOS	8419	56755	0.3	0.1	1.3	2.74x10^{-4}
Disorders of adrenal glands	2252	62922	0.5	0.1	1.6	2.75x10^{-4}
Disorders of lipoid metabolism	35419	29755	0.2	0.1	1.2	2.78x10^{-4}
Diabetes type 2 with peripheral circulatory disorders	1446	63728	0.6	0.2	1.8	3.74x10^{-4}
Type 1 diabetes with neurological manifestations	1067	64107	0.7	0.2	2.0	4.70x10^{-4}
Other forms of chronic heart disease	6657	58517	0.3	0.1	1.3	4.71x10^{-4}
Chronic Kidney Disease, Stage IV	2754	62420	0.4	0.1	1.5	5.49x10^{-4}
Ovarian dysfunction	807	64367	0.8	0.2	2.2	5.95x10^{-4}
Acquired foot deformities	7020	58154	0.3	0.1	1.3	6.02x10^{-4}
Iron deficiency anemias, unspecified or not due to blood loss	7682	57492	0.3	0.1	1.3	6.20x10^{-4}
Gout and other crystal arthropathies	5801	59373	0.3	0.1	1.4	7.04x10^{-4}
Intervertebral disc disorders	12542	52632	0.2	0.1	1.2	7.42x10^{-4}

Mood disorders	16737	48437	0.2	0.1	1.2	8.20×10^{-4}
Hypothyroidism	10724	54450	0.2	0.1	1.3	9.37×10^{-4}
Unstable angina (intermediate coronary syndrome)	5570	59604	0.3	0.1	1.4	9.55×10^{-4}
Cholelithiasis with other cholecystitis	1878	63296	0.5	0.1	1.6	0.001
Hypertensive chronic kidney disease	8311	56863	0.2	0.1	1.3	0.002
Acute pulmonary heart disease	2762	62412	0.4	0.1	1.5	0.002
Depression	15718	49456	0.2	0.1	1.2	0.002
Nonspecific chest pain	28569	36605	0.2	0.1	1.2	0.002
Nonrheumatic aortic valve disorders	6091	59083	0.3	0.1	1.3	0.002
Sleep disorders	13053	52121	0.2	0.1	1.2	0.003
Hypoventilation	1481	63693	0.5	0.2	1.6	0.003
Heart failure NOS	4067	61107	0.3	0.1	1.4	0.004
Hypothyroidism	11256	53918	0.2	0.1	1.2	0.005
Umbilical hernia	1708	63466	0.4	0.2	1.5	0.005
Cardiomyopathy	5291	59883	0.3	0.1	1.3	0.005
Calcaneal spur; Exostosis	2988	62186	0.3	0.1	1.4	0.005
Pulmonary embolism and infarction, acute	2547	62627	0.4	0.1	1.4	0.005
Spondylosis without myelopathy	11736	53438	0.2	0.1	1.2	0.006

Abbreviations: SE, standard error; OR, odds ratio

[a] Information shown in the table includes significant associations (Bonferroni corrected p-value of $p = 5.6 \times 10^{-6}$ compared to individuals with normal range BMI) with class 3 obese BMI that are replicated with 97-SNP risk score for obesity in UK Biobank and eMERGE cohorts (results significant to FDR-adjusted p-value 0.006). Logistic regression models are adjusted for age, sex, and 10 principal components.

Supplemental Table 7. Phenotype Associations of Genome-wide PRS for Obesity in eMERGE that Replicate Associations with Class 3 Obesity [a]

Phenotype	Cases	Controls	Beta	SE	OR	p-value
Overweight, obesity and other hyperalimentation	21061	44113	2.8	0.1	16.4	0
Obesity	16524	48650	3.3	0.1	26.7	0
Morbid obesity	7275	57899	4.3	0.1	77.0	0
Type 2 diabetes	17364	47810	1.9	0.1	6.8	$3.00x10^{-154}$
Diabetes mellitus	18620	46554	1.9	0.1	6.4	$7.84x10^{-151}$
Bariatric surgery	1892	63282	4.7	0.2	110.6	$3.01x10^{-150}$
Sleep apnea	9819	55355	1.9	0.1	6.5	$6.21x10^{-109}$
Eating disorder	1465	63709	4.1	0.2	61.2	$5.92x10^{-93}$
Obstructive sleep apnea	7599	57575	1.8	0.1	5.8	$9.40x10^{-78}$
Other chronic nonalcoholic liver disease	5029	60145	2.1	0.1	7.9	$1.73x10^{-75}$
Chronic liver disease and cirrhosis	5471	59703	2.0	0.1	7.1	$1.67x10^{-73}$
Hypertension	39206	25968	1.3	0.1	3.7	$3.94x10^{-73}$
Essential hypertension	38654	26520	1.3	0.1	3.6	$2.80x10^{-70}$
Gastrointestinal complications	2981	62193	2.5	0.1	12.3	$4.37x10^{-70}$
Insulin pump user	4489	60685	2.0	0.1	7.7	$2.03x10^{-62}$
Type 2 diabetes with neurological manifestations	4352	60822	2.0	0.1	7.3	$3.61x10^{-58}$
Polyneuropathy in diabetes	3597	61577	1.9	0.1	6.9	$8.04x10^{-47}$
Type 1 diabetes	4985	60189	1.6	0.1	5.1	$6.35x10^{-45}$
Type 2 diabetes with renal manifestations	3453	61721	1.9	0.1	6.5	$7.61x10^{-42}$
Edema	13670	51504	1.0	0.1	2.7	$1.17x10^{-38}$
Type 2 diabetes with ophthalmic manifestations	2447	62727	2.1	0.2	8.2	$2.69x10^{-38}$
Osteoarthrosis	22861	42313	0.9	0.1	2.4	$1.81x10^{-36}$
Other disorders of intestine	7062	58112	1.2	0.1	3.3	$2.27x10^{-36}$
Diabetic retinopathy	2674	62500	2.0	0.2	7.2	$2.43x10^{-36}$
Congestive heart failure (CHF) NOS	9979	55195	1.1	0.1	2.9	$2.05x10^{-34}$
Coronary atherosclerosis	17803	47371	0.9	0.1	2.5	$2.80x10^{-33}$
Acute renal failure	9186	55988	1.1	0.1	2.9	$3.29x10^{-33}$
Congestive heart failure; nonhypertensive	11702	53472	1.0	0.1	2.7	$6.08x10^{-33}$
Osteoarthrosis NOS	18362	46812	0.9	0.1	2.4	$2.00x10^{-32}$
Ischemic Heart Disease	20357	44817	0.8	0.1	2.3	$1.84x10^{-30}$
Renal failure	15129	50045	0.9	0.1	2.3	$7.93x10^{-30}$
Shortness of breath	19815	45359	0.7	0.1	2.1	$2.39x10^{-26}$
Hypertensive heart and/or renal disease	11631	53543	0.9	0.1	2.4	$3.47x10^{-25}$

Gout	4917	60257	1.2	0.1	3.3	5.57x10^{-25}
Cardiomegaly	9268	55906	0.9	0.1	2.5	1.13x10^{-24}
Abnormal weight gain	3618	61556	1.3	0.1	3.8	1.67x10^{-24}
Myocardial infarction	8758	56416	0.9	0.1	2.5	4.62x10^{-23}
Chronic venous insufficiency [CVI]	3708	61466	1.3	0.1	3.6	4.65x10^{-23}
Panniculitis	461	64713	3.4	0.4	31.2	8.51x10^{-23}
Encounter for long-term (current) use of anticoagulants	10310	54864	0.8	0.1	2.3	1.66x10^{-22}
Fluid overload	3837	61337	1.3	0.1	3.5	4.54x10^{-22}
Respiratory failure, insufficiency, arrest	6684	58490	1.0	0.1	2.6	4.76x10^{-22}
Cellulitis and abscess of leg, except foot	4227	60947	1.2	0.1	3.2	8.28x10^{-22}
Hyperlipidemia	34558	30616	0.7	0.1	1.9	8.68x10^{-22}
Heart failure with preserved ejection fraction [Diastolic heart failure]	3819	61355	1.3	0.1	3.5	1.11x10^{-21}
Gout and other crystal arthropathies	5801	59373	1.0	0.1	2.8	2.45x10^{-21}
Chronic renal failure [	11487	53687	0.8	0.1	2.2	7.09x10^{-21}
Hypertensive chronic kidney disease	8311	56863	0.9	0.1	2.4	1.32x10^{-20}
Depression	15718	49456	0.7	0.1	1.9	1.43x10^{-20}
Diabetes type 2 with peripheral circulatory disorders	1446	63728	1.9	0.2	7.0	1.79x10^{-20}
Ventral hernia	1887	63287	1.6	0.2	5.2	2.27x10^{-20}
Disorders of lipoid metabolism	35419	29755	0.6	0.1	1.9	2.41x10^{-20}
Mood disorders	16737	48437	0.6	0.1	1.9	3.29x10^{-20}
Pulmonary heart disease	7098	58076	0.9	0.1	2.5	3.76x10^{-20}
Gastrojejunal ulcer	271	64903	4.1	0.4	61.0	6.09x10^{-20}
Chronic pulmonary heart disease	4900	60274	1.0	0.1	2.8	2.22x10^{-19}
Dysmetabolic syndrome X	1084	64090	2.1	0.2	8.0	3.69x10^{-19}
Asthma	11096	54078	0.7	0.1	2.1	4.87x10^{-19}
Chronic ulcer of skin	5649	59525	1.0	0.1	2.6	4.89x10^{-19}
Superficial cellulitis and abscess	13672	51502	0.7	0.1	1.9	6.88x10^{-19}
Heart failure with reduced ejection fraction [Systolic or combined heart failure]	4181	60993	1.1	0.1	3.0	8.09x10^{-19}
Mixed hyperlipidemia	8620	56554	0.8	0.1	2.2	3.78x10^{-18}
Vitamin deficiency	10640	54534	0.7	0.1	2.0	5.18x10^{-18}
Other venous embolism and thrombosis	7287	57887	0.8	0.1	2.3	5.19x10^{-18}
Postoperative infection	3709	61465	1.1	0.1	3.0	8.45x10^{-18}
Localized adiposity	642	64532	2.6	0.3	13.0	1.11x10^{-17}
Pulmonary collapse; interstitial and compensatory emphysema	10304	54870	0.7	0.1	2.1	1.30x10^{-17}
Osteoarthritis; localized	12113	53061	0.7	0.1	2.0	1.71x10^{-17}
Respiratory failure	4502	60672	1.0	0.1	2.8	1.86x10^{-17}

Other dyspnea	19187	45987	0.6	0.1	1.8	3.60x10^{-17}
Cholelithiasis and cholecystitis	5727	59447	0.9	0.1	2.4	5.67x10^{-17}
Other disorders of the kidney and ureters	10875	54299	0.7	0.1	2.0	6.54x10^{-17}
Other chronic ischemic heart disease, unspecified	8263	56911	0.8	0.1	2.2	6.57x10^{-16}
Atrial fibrillation	9935	55239	0.7	0.1	2.0	1.03x10^{-15}
Cholelithiasis	5031	60143	0.9	0.1	2.4	1.67x10^{-15}
Incisional hernia	1620	63554	1.5	0.2	4.5	2.64x10^{-15}
Bacterial infection NOS	8419	56755	0.7	0.1	2.0	8.39x10^{-15}
Atrial fibrillation and flutter	10191	54983	0.7	0.1	2.0	8.52x10^{-15}
Osteoarthrosis, localized, primary	6124	59050	0.8	0.1	2.2	1.37x10^{-14}
Septicemia	5395	59779	0.8	0.1	2.3	1.43x10^{-14}
Other peripheral nerve disorders	10170	55004	0.6	0.1	1.9	2.41x10^{-14}
Chronic ulcer of leg or foot	3479	61695	1.0	0.1	2.8	2.92x10^{-14}
Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	7013	58161	0.7	0.1	2.1	4.58x10^{-14}
Wheezing	3946	61228	1.0	0.1	2.6	4.95x10^{-14}
Dependence on respirator [Ventilator] or supplemental oxygen	1443	63731	1.5	0.2	4.6	8.13x10^{-14}
Heart failure	4067	61107	0.9	0.1	2.5	1.88x10^{-13}
Swelling of limb	10189	54985	0.6	0.1	1.8	2.76x10^{-13}
Chronic airway obstruction	9300	55874	0.6	0.1	1.9	2.87x10^{-13}
Nephritis and nephropathy in diseases classified elsewhere	1849	63325	1.3	0.2	3.8	3.26x10^{-13}
Pulmonary congestion and hypostasis	3727	61447	0.9	0.1	2.6	4.97x10^{-13}
Arthropathy	9057	56117	0.6	0.1	1.9	9.59x10^{-13}
Vitamin D deficiency	8786	56388	0.6	0.1	1.9	1.03x10^{-13}
Respiratory abnormalities	4122	61052	0.9	0.1	2.4	1.53x10^{-12}
Chronic obstructive asthma	2077	63097	1.2	0.2	3.4	1.64x10^{-12}
Other arthropathies	9647	55527	0.6	0.1	1.8	1.65x10^{-12}
Angina pectoris	7300	57874	0.7	0.1	2.0	2.27x10^{-12}
Cellulitis and abscess of trunk	2431	62743	1.1	0.2	3.1	2.36x10^{-12}
Encounter for long-term (current) use of aspirin	6484	58690	0.7	0.1	2.1	2.50x10^{-12}
Proteinuria	2890	62284	1.0	0.1	2.8	3.60x10^{-12}
Acquired acanthosis nigricans	437	64737	2.8	0.4	16.3	3.63x10^{-12}
Primary/intrinsic cardiomyopathies	4871	60303	0.8	0.1	2.2	5.20x10^{-12}
Cirrhosis of liver without mention of alcohol	1523	63651	1.4	0.2	4.0	8.36x10^{-12}
Decubitus ulcer	2278	62896	1.1	0.2	3.1	1.10x10^{-11}
Iron deficiency anemias	8970	56204	0.6	0.1	1.8	2.33x10^{-11}

Chronic Kidney Disease, Stage III	6548	58626	0.7	0.1	2.0	2.57x10^{-11}
Other forms of chronic heart disease	6657	58517	0.7	0.1	2.0	2.90x10^{-11}
Cardiomyopathy	5291	59883	0.7	0.1	2.1	5.32x10^{-11}
Hypertensive heart disease	5029	60145	0.8	0.1	2.1	7.23x10^{-11}
Unstable angina (intermediate coronary syndrome)	5570	59604	0.7	0.1	2.0	9.59x10^{-11}
Spinal stenosis of lumbar region	4915	60259	0.7	0.1	2.1	1.05x10^{-10}
Hypoventilation	1481	63693	1.3	0.2	3.7	1.37x10^{-10}
Chronic Kidney Disease, Stage IV	2754	62420	1.0	0.2	2.7	1.65x10^{-10}
Hypercholesterolemia	18414	46760	0.4	0.1	1.6	1.80x10^{-10}
Candidiasis of skin and nails	1357	63817	1.3	0.2	3.8	1.88x10^{-10}
Obstructive chronic bronchitis	2752	62422	1.0	0.2	2.6	2.61x10^{-10}
Carbuncle and furuncle	1396	63778	1.3	0.2	3.7	2.88x10^{-10}
Type 1 diabetes with neurological manifestations	1067	64107	1.5	0.2	4.6	2.93x10^{-10}
Gouty arthropathy	2091	63083	1.1	0.2	2.9	8.58x10^{-10}
Hypoglycemia	1818	63356	1.1	0.2	3.1	9.04x10^{-10}
Osteoarthrosis, generalized	3751	61423	0.8	0.1	2.2	9.22x10^{-10}
Liver abscess and sequelae of chronic liver disease	1339	63835	1.3	0.2	3.7	9.26x10^{-10}
Deep vein thrombosis	3311	61863	0.8	0.1	2.3	1.05x10^{-9}
Iron deficiency anemias, unspecified or not due to blood loss	7682	57492	0.6	0.1	1.8	1.78x10^{-9}
Umbilical hernia	1708	63466	1.1	0.2	3.1	2.24x10^{-9}
Polycystic ovaries	578	64596	2.0	0.3	7.0	2.69x10^{-9}
Staphylococcus infections	3035	62139	0.8	0.1	2.3	3.49x10^{-9}
Paroxysmal ventricular tachycardia	3201	61973	0.8	0.1	2.3	5.29x10^{-9}
Other symptoms of respiratory system	35281	29893	0.4	0.1	1.4	6.09x10^{-9}
Acute bronchitis and bronchiolitis	8554	56620	0.5	0.1	1.7	6.64x10^{-9}
Abnormal function study of cardiovascular system	4760	60414	0.7	0.1	2.0	7.66x10^{-9}
Diseases of esophagus	23509	41665	0.4	0.1	1.4	1.34x10^{-8}
Peritoneal adhesions (postoperative) (postinfection)	1424	63750	1.2	0.2	3.2	1.52x10^{-8}
Osteomyelitis	1739	63435	1.1	0.2	2.9	1.70x10^{-8}
Nephritis and nephropathy without mention of glomerulonephritis	2760	62414	0.9	0.2	2.4	1.76x10^{-8}
Renal failure	3058	62116	0.8	0.1	2.3	1.81x10^{-8}
Other specified erythematous conditions	1984	63190	1.0	0.2	2.6	1.88x10^{-8}
Megaloblastic anemia	1427	63747	1.1	0.2	3.2	1.91x10^{-8}
Chronic bronchitis	3433	61741	0.8	0.1	2.1	2.09x10^{-8}
Hypothyroidism	10724	54450	0.5	0.1	1.6	2.12x10^{-8}

Portal hypertension	724	64450	1.6	0.3	4.8	3.04x10^{-8}
Spinal stenosis	6326	58848	0.6	0.1	1.8	3.63x10^{-8}
Bundle branch block	4474	60700	0.7	0.1	1.9	3.65x10^{-8}
Calcaneal spur; Exostosis	2988	62186	0.8	0.1	2.2	3.68x10^{-8}
Shock	1881	63293	1.0	0.2	2.7	4.10x10^{-8}
Primary pulmonary hypertension	1266	63908	1.2	0.2	3.3	5.18x10^{-8}
Osteoarthrosis involving more than one site, but not specified as generalized	2277	62897	0.9	0.2	2.5	5.32x10^{-8}
Bursitis	3396	61778	0.7	0.1	2.1	5.33x10^{-8}
Hyperglyceridemia	2924	62250	0.8	0.1	2.2	5.49x10^{-8}
Cardiac pacemaker/device in situ	4020	61154	0.7	0.1	2.0	6.22x10^{-8}
Cardiac defibrillator in situ	1877	63297	1.0	0.2	2.6	7.24x10^{-8}
Vitamin B-complex deficiencies	2607	62567	0.8	0.2	2.3	7.50x10^{-8}
Non-healing surgical wound	716	64458	1.5	0.3	4.6	7.72x10^{-8}
Orthopnea	576	64598	1.7	0.3	5.7	1.09x10^{-7}
Hypothyroidism	11256	53918	0.4	0.1	1.5	1.10x10^{-7}
Cholelithiasis with other cholecystitis	1878	63296	0.9	0.2	2.6	1.18x10^{-7}
Gastroesophageal reflux disease	20873	44301	0.3	0.1	1.4	1.54x10^{-7}
Intestinal malabsorption (non-celiac)	1168	64006	1.2	0.2	3.2	1.92x10^{-7}
Nephritis; nephrosis; renal sclerosis	4173	61001	0.7	0.1	1.9	1.95x10^{-7}
Benign neoplasm of adrenal gland	477	64697	1.8	0.3	6.1	2.15x10^{-7}
Esophagitis, Gastroesophageal reflux disease and related diseases	22561	42613	0.3	0.1	1.4	2.17x10^{-7}
Osteomyelitis, periostitis, and other infections involving bone	2244	62930	0.9	0.2	2.4	2.32x10^{-7}
Infection/inflammation of internal prosthetic device; implant; and graft	2852	62322	0.8	0.1	2.1	2.57x10^{-7}
Ill-defined descriptions and complications of heart disease	10680	54494	0.4	0.1	1.5	2.80x10^{-7}
Cardiac conduction disorders	15497	49677	0.4	0.1	1.5	2.99x10^{-7}
Arthropathy associated with neurological disorders	328	64846	2.2	0.4	8.6	4.72x10^{-7}
Sleep disorders	13053	52121	0.4	0.1	1.5	5.20x10^{-7}
Acute pulmonary heart disease	2762	62412	0.7	0.1	2.1	7.53x10^{-7}
Cardiac pacemaker in situ	3163	62011	0.7	0.1	2.0	7.74x10^{-7}
Left bundle branch block	2454	62720	0.8	0.2	2.2	8.68x10^{-7}
Asthma with exacerbation	2887	62287	0.7	0.2	2.1	1.12x10^{-7}
Dermatophytosis	9887	55287	0.4	0.1	1.5	1.45x10^{-7}
Cellulitis and abscess of foot, toe	1652	63522	0.9	0.2	2.5	1.64x10^{-6}
Other vitamin B12 deficiency anemia	906	64268	1.2	0.3	3.4	2.15x10^{-6}
Sinoatrial node dysfunction (Bradycardia)	2471	62703	0.7	0.2	2.1	2.22x10^{-6}

Malignant neoplasm of uterus	703	64471	1.4	0.3	3.9	2.42x10^{-6}
Coagulation defects	6539	58635	0.5	0.1	1.6	3.11x10^{-6}
Hereditary and idiopathic peripheral neuropathy	5570	59604	0.5	0.1	1.7	3.12x10^{-6}
Intervertebral disc disorders	12542	52632	0.4	0.1	1.4	3.33x10^{-6}
Chronic venous hypertension	377	64797	1.8	0.4	6.1	3.60x10^{-6}
Dermatophytosis of nail	6482	58692	0.5	0.1	1.6	3.75x10^{-6}
Chronic kidney disease, Stage I or II	2369	62805	0.8	0.2	2.1	4.19x10^{-6}
Phlebitis and thrombophlebitis	2869	62305	0.7	0.1	1.9	4.27x10^{-6}
Joint effusions	5967	59207	0.5	0.1	1.6	4.49x10^{-6}
Chronic ulcer of unspecified site	1914	63260	0.8	0.2	2.3	4.53x10^{-6}
Other deficiency anemia	2699	62475	0.7	0.2	2.0	5.17x10^{-6}
Hidradenitis	385	64789	1.9	0.4	6.4	7.96x10^{-6}
Pain	9394	55780	0.4	0.1	1.5	8.43x10^{-6}
Difficulty in walking	3926	61248	0.6	0.1	1.8	8.67x10^{-6}
Dermatophytosis / Dermatomycosis	10419	54755	0.4	0.1	1.4	1.01x10^{-5}
Thoracic or lumbosacral neuritis or radiculitis, unspecified	6140	59034	0.5	0.1	1.6	1.20x10^{-5}
Abdominal hernia	11852	53322	0.3	0.1	1.4	1.27x10^{-5}
Displacement of intervertebral disc	5319	59855	0.5	0.1	1.6	1.66x10^{-5}
Ovarian dysfunction	807	64367	1.2	0.3	3.3	1.70x10^{-5}
Neuralgia, neuritis, and radiculitis	3284	61890	0.6	0.1	1.8	1.84x10^{-5}
Postphlebitic syndrome	280	64894	1.9	0.4	6.8	1.87x10^{-5}
Nonspecific chest pain	28569	36605	0.3	0.1	1.3	1.90x10^{-5}
Varicose veins of lower extremity	4119	61055	0.5	0.1	1.7	2.04x10^{-5}
Complication of internal orthopedic device	2205	62969	0.7	0.2	2.0	2.16x10^{-5}
Pulmonary embolism and infarction, acute	2547	62627	0.7	0.2	1.9	2.33x10^{-5}
Other hypertensive complications	3536	61638	0.6	0.1	1.8	2.36x10^{-5}
Atrial flutter	3193	61981	0.6	0.1	1.8	2.75x10^{-5}
Varicose veins of lower extremity, symptomatic	2444	62730	0.7	0.2	1.9	3.13x10^{-5}
Athlete's foot	2929	62245	0.6	0.1	1.8	3.19x10^{-5}
Cholecystitis without cholelithiasis	1710	63464	0.8	0.2	2.2	3.43x10^{-5}
Disorders of parathyroid gland	2164	63010	0.7	0.2	2.0	3.56x10^{-5}
Disorders of adrenal glands	2252	62922	0.7	0.2	2.0	3.59x10^{-5}
Ingrowing nail	2743	62431	0.6	0.1	1.8	4.32x10^{-5}
Other local infections of skin and subcutaneous tissue	3436	61738	0.5	0.1	1.7	4.74x10^{-5}
Chronic pain	4649	60525	0.5	0.1	1.6	5.14x10^{-5}
Arrhythmia (cardiac) NOS	9842	55332	0.3	0.1	1.4	7.42x10^{-5}

Cardiac dysrhythmias	27872	37302	0.3	0.1	1.3	7.66x10^{-5}
Bronchitis	6016	59158	0.4	0.1	1.5	8.03x10^{-5}
Fasciitis	4089	61085	0.5	0.1	1.6	8.32x10^{-5}
Flat foot	1773	63401	0.7	0.2	2.1	9.17x10^{-5}
Endometrial hyperplasia	625	64549	1.2	0.3	3.3	1.01x10^{-4}
Phlebitis and thrombophlebitis of lower extremities	1838	63336	0.7	0.2	2.0	1.18x10^{-4}
Nonrheumatic aortic valve disorders	6091	59083	0.4	0.1	1.5	1.24x10^{-4}
Cardiac arrest and ventricular fibrillation	1490	63684	0.8	0.2	2.2	1.24x10^{-4}
Dislocation	4381	60793	0.5	0.1	1.6	1.30x10^{-4}
Degeneration of intervertebral disc	9858	55316	0.3	0.1	1.4	1.36x10^{-4}
Barrett's esophagus	925	64249	0.9	0.2	2.6	1.40x10^{-4}
Back pain	24723	40451	0.2	0.1	1.3	1.58x10^{-4}
Disorders of sweat glands	1107	64067	0.9	0.2	2.5	1.70x10^{-4}
Sleep related movement disorders	2593	62581	0.6	0.2	1.8	1.72x10^{-4}
Atherosclerotic cardiovascular disease	1821	63353	0.7	0.2	2.0	1.78x10^{-4}
Spondylosis and allied disorders	12335	52839	0.3	0.1	1.3	1.84x10^{-4}
Other abnormal blood chemistry	10148	55026	0.3	0.1	1.4	2.08x10^{-4}
Restless legs syndrome	1834	63340	0.7	0.2	1.9	2.16x10^{-4}
Synovitis and tenosynovitis	6449	58725	0.4	0.1	1.5	2.18x10^{-4}
Cardiac arrest	1056	64118	0.9	0.2	2.4	2.43x10^{-4}
Symptoms and disorders of the joints	13943	51231	0.3	0.1	1.3	2.45x10^{-4}
Methicillin resistant Staphylococcus aureus	1063	64111	0.9	0.2	2.4	2.49x10^{-4}
Hyperparathyroidism	1866	63308	0.7	0.2	1.9	2.65x10^{-4}
Varicose veins	4735	60439	0.4	0.1	1.5	2.78x10^{-4}
Spondylosis without myelopathy	11736	53438	0.3	0.1	1.3	3.22x10^{-4}
Other disorders of synovium, tendon, and bursa	11380	53794	0.3	0.1	1.3	3.59x10^{-4}
Arthropathy associated with other disorders classified elsewhere	506	64668	1.2	0.3	3.4	4.03x10^{-4}
Internal derangement of knee	4467	60707	0.4	0.1	1.5	4.15x10^{-4}
Noninfectious disorders of lymphatic channels	1432	63742	0.7	0.2	2.1	4.19x10^{-4}
Chronic tonsillitis and adenoiditis	1295	63879	0.8	0.2	2.2	4.20x10^{-4}
Secondary/extrinsic cardiomyopathies	1820	63354	0.6	0.2	1.9	5.77x10^{-4}
Infective connective tissue disorders	260	64914	1.6	0.5	5.2	5.87x10^{-4}
Other abnormal glucose	9056	56118	0.3	0.1	1.3	6.61x10^{-4}
Abnormal coagulation profile	2503	62671	0.5	0.2	1.7	7.20x10^{-4}
Elevated sedimentation rate	898	64276	0.8	0.3	2.3	8.93x10^{-4}
Spondylosis with myelopathy	1289	63885	0.7	0.2	2.0	9.14x10^{-4}

Acquired foot deformities	7020	58154	0.3	0.1	1.4	0.001
Adrenal hyperfunction	520	64654	1.1	0.3	3.0	0.001
Diabetes type 1 with peripheral circulatory disorders	339	64835	1.4	0.4	3.9	0.001
Corns and callosities	3552	61622	0.4	0.1	1.5	0.001
Acute pain	6249	58925	0.3	0.1	1.4	0.001
Abnormal glucose	11626	53548	0.3	0.1	1.3	0.002
Chronic pain syndrome	1118	64056	0.7	0.2	2.1	0.002
Acute and chronic tonsillitis	1774	63400	0.6	0.2	1.8	0.002
Other disorders of pancreatic internal secretion	884	64290	0.8	0.3	2.3	0.002
Acute sinusitis	8984	56190	0.3	0.1	1.3	0.002
Rupture of tendon, nontraumatic	1588	63586	0.6	0.2	1.8	0.002
Hemorrhagic disorder due to intrinsic circulating anticoagulants	885	64289	0.8	0.3	2.2	0.002
Sciatica	4206	60968	0.4	0.1	1.5	0.002
Urinary incontinence	7483	57691	0.3	0.1	1.3	0.002
Postlaminectomy syndrome	759	64415	0.9	0.3	2.4	0.002
Dysthymic disorder	4404	60770	0.4	0.1	1.4	0.002
Diverticulitis	2360	62814	0.5	0.2	1.6	0.002
Cancer of kidney and renal pelvis	1245	63929	0.7	0.2	1.9	0.002
Precordial pain	3106	62068	0.4	0.1	1.5	0.003
Cushing's syndrome	296	64878	1.3	0.4	3.7	0.003
Atrioventricular block	4901	60273	0.3	0.1	1.4	0.003
Urinary calculus	5298	59876	0.3	0.1	1.4	0.004
Hypersomnia	1359	63815	0.6	0.2	1.8	0.004
Otitis externa	3150	62024	0.4	0.1	1.5	0.004
Retinal edema	845	64329	0.8	0.3	2.2	0.004
Myopathy	1272	63902	0.6	0.2	1.8	0.005
Cellulitis and abscess of fingers/toes	3489	61685	0.4	0.1	1.5	0.005
Benign neoplasm of other endocrine glands and related structures	1477	63697	0.6	0.2	1.7	0.005
Hirsutism	617	64557	0.9	0.3	2.4	0.005
Suppurative and unspecified otitis media	4431	60743	0.3	0.1	1.4	0.006
Abnormal electrocardiogram	9068	56106	0.2	0.1	1.3	0.006
Other disorders of soft tissues	1898	63276	0.5	0.2	1.6	0.006
Psoriasis	2054	63120	0.5	0.2	1.6	0.006
Diaphragmatic hernia	6702	58472	0.3	0.1	1.3	0.006
Other symptoms involving abdomen and pelvis	7188	57986	0.3	0.1	1.3	0.008
Other disorders of lipoid metabolism	1556	63618	0.5	0.2	1.7	0.008

Intervertebral disc disorder with myelopathy	606	64568	0.8	0.3	2.3	0.008
Psoriasis vulgaris	1869	63305	0.5	0.2	1.6	0.008
Otitis media	5557	59617	0.3	0.1	1.3	0.010
Diabetes insipidus	180	64994	1.5	0.6	4.3	0.010
Pilonidal cyst	232	64942	1.3	0.5	3.7	0.010
Rheumatoid arthritis and other inflammatory polyarthropathies	4073	61101	0.3	0.1	1.4	0.01
Acquired spondylolisthesis	2036	63138	0.4	0.2	1.6	0.01
Rotator cuff (capsule) sprain	2399	62775	0.4	0.2	1.5	0.01
Posttraumatic stress disorder	1053	64121	0.6	0.2	1.9	0.01
Other nondiabetic retinopathy	1341	63833	0.5	0.2	1.7	0.01
Other alveolar and parietoalveolar pneumonopathy	825	64349	0.7	0.3	2.0	0.01
Peripheral enthesopathies and allied syndromes	17614	47560	0.2	0.1	1.2	0.01
Major depressive disorder	5339	59835	0.3	0.1	1.3	0.01
Personality disorders	1206	63968	0.5	0.2	1.7	0.01

Abbreviations: PRS, polygenic risk score; SE, standard error; OR, odds ratio

[a] Information shown in the table includes significant associations (Bonferroni corrected p-value of p = 5.6 x 10^{-6} compared to individuals with normal range BMI) with class 3 obese BMI that are replicated with genome-wide PRS for obesity (results significant to FDR-adjusted p-value 0.015). Logistic regression models are adjusted for age, sex, site, and 10 principal components.

Supplemental Table 8. Phenotype Associations of Genome-wide PRS for Obesity in UK Biobank that Replicate Associations in eMERGE cohort and Class 3 Obesity

Phenotype	Cases	Controls	Beta	SE	OR	p-value
Essential hypertension	89669	315763	0.2	0.0	1.2	0
Hypertension	89898	315534	0.2	0.0	1.2	0
Diabetes mellitus	27240	378192	0.3	0.0	1.3	0
Type 2 diabetes	25470	379962	0.3	0.0	1.4	0
Overweight, obesity and other hyperalimentation	15258	390174	0.6	0.0	1.8	0
Obesity	15138	390294	0.6	0.0	1.8	0
Osteoarthrosis	46919	358513	0.2	0.0	1.2	7.63×10^{-233}
Osteoarthritis; localized	39428	366004	0.2	0.0	1.2	9.48×10^{-204}
Hyperlipidemia	42442	362990	0.1	0.0	1.2	1.14×10^{-109}
Disorders of lipoid metabolism	42579	362853	0.1	0.0	1.2	1.76×10^{-109}
Sleep apnea	5483	399949	0.4	0.0	1.4	3.06×10^{-105}
Hypercholesterolemia	39485	365947	0.1	0.0	1.2	8.85×10^{-104}
Ischemic Heart Disease	34324	371108	0.1	0.0	1.2	5.91×10^{-93}
Sleep disorders	6746	398686	0.3	0.0	1.3	7.26×10^{-91}
Other peripheral nerve disorders	13266	392166	0.2	0.0	1.2	3.21×10^{-90}
Cholelithiasis and cholecystitis	17141	388291	0.2	0.0	1.2	5.05×10^{-81}
Osteoarthrosis	12600	392832	0.2	0.0	1.2	4.12×10^{-80}
Other chronic ischemic heart disease, unspecified	17061	388371	0.2	0.0	1.2	1.15×10^{-78}
Osteoarthrosis, localized, primary	13382	392050	0.2	0.0	1.2	3.49×10^{-71}
Cholelithiasis	14538	390894	0.2	0.0	1.2	8.81×10^{-70}
Chronic airway obstruction	15454	389978	0.2	0.0	1.2	7.15×10^{-63}
Angina pectoris	17877	387555	0.2	0.0	1.2	9.92×10^{-63}
Renal failure	17488	387944	0.1	0.0	1.2	2.03×10^{-53}
Hypothyroidism NOS	16456	388976	0.1	0.0	1.2	3.51×10^{-50}
Hypothyroidism	18439	386993	0.1	0.0	1.1	6.55×10^{-50}
Cardiac dysrhythmias	28797	376635	0.1	0.0	1.1	1.10×10^{-44}
Myocardial infarction	13690	391742	0.1	0.0	1.2	2.19×10^{-44}
Coronary atherosclerosis	21097	384335	0.1	0.0	1.1	8.29×10^{-44}
Atrial fibrillation and flutter	17559	387873	0.1	0.0	1.1	1.08×10^{-43}
Osteoarthrosis, generalized	7278	398154	0.2	0.0	1.2	2.06×10^{-42}
Chronic renal failure	6697	398735	0.2	0.0	1.2	2.34×10^{-40}
Bacterial infection	10374	395058	0.2	0.0	1.2	5.39×10^{-38}
Internal derangement of knee	15615	389817	0.1	0.0	1.1	1.08×10^{-37}
Type 1 diabetes	3520	401912	0.3	0.0	1.3	1.62×10^{-36}

Esophagitis, GERD and related diseases	38281	367151	0.1	0.0	1.1	2.67x10^{-36}
Acute renal failure	11750	393682	0.1	0.0	1.2	2.97x10^{-35}
Chronic ulcer of leg or foot	1473	403959	0.4	0.0	1.5	2.10x10^{-34}
Diseases of esophagus	42030	363402	0.1	0.0	1.1	4.30x10^{-34}
Cholelithiasis with other cholecystitis	5901	399531	0.2	0.0	1.2	2.03x10^{-33}
Umbilical hernia	4336	401096	0.2	0.0	1.2	2.11x10^{-33}
Congestive heart failure; nonhypertensive	5488	399944	0.2	0.0	1.2	6.03x10^{-33}
Shortness of breath	7444	397988	0.2	0.0	1.2	8.17x10^{-33}
Other symptoms of respiratory system	10543	394889	0.1	0.0	1.2	1.74x10^{-32}
Diaphragmatic hernia	31231	374201	0.1	0.0	1.1	5.46x10^{-30}
Type 2 diabetes with neurological manifestations	694	404738	0.5	0.0	1.7	1.87x10^{-28}
Other disorders of intestine	18417	387015	0.1	0.0	1.1	2.06x10^{-28}
Chronic ulcer of skin	2747	402685	0.2	0.0	1.3	1.38x10^{-26}
Chronic liver disease and cirrhosis	10226	395206	0.1	0.0	1.1	1.87x10^{-26}
Respiratory failure	2779	402653	0.2	0.0	1.3	1.05x10^{-25}
Pulmonary heart disease	5231	400201	0.2	0.0	1.2	8.39x10^{-25}
Chronic bronchitis	4005	401427	0.2	0.0	1.2	2.02x10^{-24}
Respiratory failure, insufficiency, arrest	2992	402440	0.2	0.0	1.3	3.09x10^{-24}
Unstable angina (intermediate coronary syndrome)	5407	400025	0.2	0.0	1.2	5.33x10^{-24}
Major depressive disorder	16072	389360	0.1	0.0	1.1	6.70x10^{-24}
Depression	16072	389360	0.1	0.0	1.1	6.70x10^{-24}
Chronic Kidney Disease, Stage III	3605	401827	0.2	0.0	1.2	1.32x10^{-23}
Heart failure NOS	4202	401230	0.2	0.0	1.2	2.02x10^{-23}
Cardiomegaly	3481	401951	0.2	0.0	1.2	3.09x10^{-23}
Obstructive chronic bronchitis	3426	402006	0.2	0.0	1.2	3.18x10^{-23}
Diabetic retinopathy	3991	401441	0.2	0.0	1.2	5.70x10^{-23}
Other chronic nonalcoholic liver disease	3957	401475	0.2	0.0	1.2	8.02x10^{-23}
Abdominal hernia	55987	349445	0.1	0.0	1.1	2.69x10^{-22}
Iron deficiency anemias, unspecified or not due to blood loss	9812	395620	0.1	0.0	1.1	3.27x10^{-22}
Gout	2747	402685	0.2	0.0	1.2	4.81x10^{-22}
Type 2 diabetes with ophthalmic manifestations	3582	401850	0.2	0.0	1.2	6.53x10^{-22}
Gout and other crystal arthropathies	3271	402161	0.2	0.0	1.2	6.95x10^{-22}

GERD	18715	386717	0.1	0.0	1.1	1.37x10^{-21}
Postoperative infection	4888	400544	0.2	0.0	1.2	7.18x10^{-21}
Iron deficiency anemias	10253	395179	0.1	0.0	1.1	7.19x10^{-21}
Staphylococcus infections	3518	401914	0.2	0.0	1.2	1.57x10^{-20}
Spondylosis and allied disorders	12774	392658	0.1	0.0	1.1	3.24x10^{-20}
Ventral hernia	3967	401465	0.2	0.0	1.2	3.46x10^{-20}
Mood disorders	17088	388344	0.1	0.0	1.1	7.8x10^{-20}
Edema	2256	403176	0.2	0.0	1.3	2.39x10^{-19}
Spondylosis without myelopathy	11730	393702	0.1	0.0	1.1	4.97x10^{-19}
Intervertebral disc disorders	7038	398394	0.1	0.0	1.1	1.74x10^{-18}
Congestive heart failure	1967	403465	0.2	0.0	1.3	2.30 x10^{-18}
Pulmonary embolism and infarction, acute	4225	401207	0.2	0.0	1.2	2.62x10^{-18}
Acute pulmonary heart disease	4225	401207	0.2	0.0	1.2	2.62x10^{-18}
Cholecystitis without cholelithiasis	2902	402530	0.2	0.0	1.2	6.13x10^{-18}
Diabetes type 2 with peripheral circulatory disorders	489	404943	0.5	0.1	1.6	1.15x10^{-17}
Psoriasis	2660	402772	0.2	0.0	1.2	8.09x10^{-17}
Other disorders of synovium, tendon, and bursa	18192	387240	0.1	0.0	1.1	1.35x10^{-16}
Spinal stenosis	4723	400709	0.1	0.0	1.2	2.34x10^{-16}
Psoriasis vulgaris	2015	403417	0.2	0.0	1.2	2.39x10^{-16}
Rheumatoid arthritis and other inflammatory polyarthropathies	4052	401380	0.2	0.0	1.2	8.08x10^{-16}
Precordial pain	4435	400997	0.1	0.0	1.2	9.39x10^{-16}
Nonspecific chest pain	4435	400997	0.1	0.0	1.2	9.39x10^{-16}
Back pain	7049	398383	0.1	0.0	1.1	8.69x10^{-15}
Nonrheumatic aortic valve disorders	3277	402155	0.2	0.0	1.2	3.86x10^{-14}
Peripheral enthesopathies and allied syndromes	25714	379718	0.1	0.0	1.1	2.71x10^{-13}
Spinal stenosis of lumbar region	3481	401951	0.1	0.0	1.2	5.36x10^{-13}
Degeneration of intervertebral disc	3433	401999	0.1	0.0	1.2	6.01x10^{-13}
Symptoms and disorders of the joints	4064	401368	0.1	0.0	1.1	6.35x10^{-13}
Septicemia	5245	400187	0.1	0.0	1.1	6.80x10^{-13}
Hypertensive heart and/or renal disease	1801	403631	0.2	0.0	1.2	1.13x10^{-12}
Peritoneal adhesions (postoperative) (postinfection)	3906	401526	0.1	0.0	1.1	1.49x10^{-12}
Other local infections of skin and subcutaneous tissue	2779	402653	0.2	0.0	1.2	6.83x10^{-12}
Superficial cellulitis and abscess	2438	402994	0.2	0.0	1.2	1.21x10^{-11}

Hypertensive chronic kidney disease	1586	403846	0.2	0.0	1.2	1.95x10^{-11}
Varicose veins of lower extremity	11285	394147	0.1	0.0	1.1	3.13x10^{-11}
Phlebitis and thrombophlebitis	1339	404093	0.2	0.0	1.2	4.05x10^{-11}
Varicose veins	11806	393626	0.1	0.0	1.1	4.84x10^{-11}
Urinary incontinence	4369	401063	0.1	0.0	1.1	5.14x10^{-11}
Phlebitis and thrombophlebitis of lower extremities	968	404464	0.2	0.0	1.3	1.10x10^{-10}
Hypoventilation	124	405308	0.7	0.1	2.0	1.28x10^{-10}
Liver abscess and sequelae of chronic liver disease	6660	398772	0.1	0.0	1.1	1.98x10^{-10}
Other forms of chronic heart disease	2809	402623	0.1	0.0	1.2	2.19x10^{-10}
Disorders of sweat glands	719	404713	0.3	0.0	1.3	2.38x10^{-10}
Cardiac conduction disorders	7173	398259	0.1	0.0	1.1	4.38x10^{-10}
Hidradenitis	189	405243	0.5	0.1	1.7	5.20x10^{-10}
Noninfectious disorders of lymphatic channels	918	404514	0.2	0.0	1.3	5.84x10^{-10}
Other deficiency anemia	1347	404085	0.2	0.0	1.2	3.24x10^{-9}
Megaloblastic anemia	1282	404150	0.2	0.0	1.2	6.11x10^{-9}
Varicose veins of lower extremity, symptomatic	823	404609	0.2	0.0	1.3	7.79x10^{-9}
Chronic pulmonary heart disease	1099	404333	0.2	0.0	1.2	1.21x10^{-8}
Calcaneal spur; Exostosis NOS	790	404642	0.2	0.0	1.3	1.68x10^{-8}
Thoracic or lumbosacral neuritis or radiculitis, unspecified	4436	400996	0.1	0.0	1.1	4.06x10^{-8}
Cardiomyopathy	1572	403860	0.2	0.0	1.2	4.60x10^{-8}
Displacement of intervertebral disc	3415	402017	0.1	0.0	1.1	4.63x10^{-8}
Urinary calculus	7934	397498	0.1	0.0	1.1	5.25x10^{-8}
Primary/intrinsic cardiomyopathies	1533	403899	0.2	0.0	1.2	5.94x10^{-8}
Rupture of tendon, nontraumatic	5459	399973	0.1	0.0	1.1	6.52x10^{-8}
Abnormal glucose	852	404580	0.2	0.0	1.3	6.82x10^{-8}
Vitamin deficiency	2934	402498	0.1	0.0	1.1	7.68x10^{-8}
Other disorders of pancreatic internal secretion	1307	404125	0.2	0.0	1.2	8.08x10^{-8}
Other specified erythematous conditions	1254	404178	0.2	0.0	1.2	9.24x10^{-8}
Hypoglycemia	1279	404153	0.2	0.0	1.2	1.26x10^{-7}
Synovitis and tenosynovitis	8584	396848	0.1	0.0	1.1	1.32x10^{-7}
Sciatica	1583	403849	0.2	0.0	1.2	1.95x10^{-7}
Type 2 diabetes with renal manifestations	423	405009	0.3	0.1	1.4	1.96x10^{-7}
Incisional hernia	293	405139	0.4	0.1	1.4	2.42x10^{-7}

Vitamin B-complex deficiencies	1647	403785	0.2	0.0	1.2	2.45x10^{-7}
Fluid overload	750	404682	0.2	0.0	1.3	2.85x10^{-7}
Renal failure NOS	1618	403814	0.2	0.0	1.2	5.28x10^{-7}
Acquired foot deformities	10733	394699	0.1	0.0	1.1	8.24x10^{-7}
Ingrowing nail	848	404584	0.2	0.0	1.2	8.38x10^{-7}
Acquired spondylolisthesis	1733	403699	0.1	0.0	1.2	1.15x10^{-6}
Malignant neoplasm of uterus	1548	403884	0.1	0.0	1.2	1.33x10^{-6}
Other alveolar and parietoalveolar pneumonopathy	440	404992	0.3	0.1	1.3	1.57x10^{-6}
Nephritis; nephrosis; renal sclerosis	7532	397900	0.1	0.0	1.1	2.17x10^{-6}
Bundle branch block	4465	400967	0.1	0.0	1.1	3.05x10^{-6}
Atrioventricular block	2846	402586	0.1	0.0	1.1	4.43x10^{-6}
Otitis externa	959	404473	0.2	0.0	1.2	5.49x10^{-6}
Primary pulmonary hypertension	504	404928	0.2	0.1	1.3	7.96x10^{-6}
Difficulty in walking	216	405216	0.4	0.1	1.4	8.67x10^{-6}
Pulmonary collapse; interstitial and compensatory emphysema	2840	402592	0.1	0.0	1.1	8.70x10^{-6}
Cirrhosis of liver without mention of alcohol	838	404594	0.2	0.0	1.2	1.08x10^{-5}
Decubitus ulcer	1342	404090	0.1	0.0	1.2	1.44x10^{-5}
Cellulitis and abscess of trunk	680	404752	0.2	0.0	1.2	1.45x10^{-5}
Joint effusions	1640	403792	0.1	0.0	1.1	1.52x10^{-5}
Respiratory abnormalities	609	404823	0.2	0.0	1.2	1.63x10^{-5}
Polycystic ovaries	264	405168	0.3	0.1	1.4	2.02x10^{-5}
Chronic Kidney Disease, Stage IV	557	404875	0.2	0.1	1.2	2.53x10^{-5}
Chronic venous insufficiency	271	405161	0.3	0.1	1.3	4.24x10^{-5}
Other venous embolism and thrombosis	592	404840	0.2	0.0	1.2	5.51x10^{-5}
Mixed hyperlipidemia	182	405250	0.4	0.1	1.4	6.28x10^{-5}
Other disorders of the kidney and ureters	4840	400592	0.1	0.0	1.1	6.52x10^{-5}
Barrett's esophagus	3516	401916	0.1	0.0	1.1	9.99x10^{-5}
Other abnormal blood chemistry	558	404874	0.2	0.1	1.2	1.02x10^{-4}
Other arthropathies	1934	403498	0.1	0.0	1.1	1.44x10^{-4}
Other abnormal glucose	496	404936	0.2	0.1	1.2	1.94x10^{-4}
Complication of internal orthopedic device	362	405070	0.2	0.1	1.3	2.51x10^{-4}
Other disorders of soft tissues	1288	404144	0.1	0.0	1.1	3.21x10^{-4}
Pilonidal cyst	511	404921	0.2	0.1	1.2	3.37x10^{-4}
Hereditary and idiopathic peripheral neuropathy	534	404898	0.2	0.1	1.2	3.88x10^{-4}

Localized adiposity	77	405355	0.5	0.1	1.6	4.24x10^{-4}
Benign neoplasm of other endocrine glands and related structures	1100	404332	0.1	0.0	1.1	4.54x10^{-4}
Cardiac arrest and ventricular fibrillation	861	404571	0.1	0.0	1.1	6.90x10^{-4}
Ovarian dysfunction	306	405126	0.2	0.1	1.3	6.93x10^{-4}
Swelling of limb	215	405217	0.3	0.1	1.3	8.94x10^{-4}
Type 1 diabetes with neurological manifestations	178	405254	0.3	0.1	1.3	9.68x10^{-4}
Diabetes type 1 with peripheral circulatory disorders	97	405335	0.4	0.1	1.5	0.001
Benign neoplasm of adrenal gland	237	405195	0.3	0.1	1.3	0.001
Wheezing	326	405106	0.2	0.1	1.2	0.001
Disorders of adrenal glands	3146	402286	0.1	0.0	1.1	0.001
Portal hypertension	708	404724	0.1	0.0	1.2	0.001
Osteomyelitis	324	405108	0.2	0.1	1.2	0.002
Suppurative and unspecified otitis media	822	404610	0.1	0.0	1.1	0.003
Osteomyelitis, periostitis, and other infections involving bone	363	405069	0.2	0.1	1.2	0.003
Cardiac arrest	478	404954	0.2	0.1	1.2	0.004
Endometrial hyperplasia	982	404450	0.1	0.0	1.1	0.004
Left bundle branch block	2343	403089	0.1	0.0	1.1	0.005
Acute and chronic tonsillitis	2063	403369	0.1	0.0	1.1	0.005
Arthropathy associated with neurological disorders	74	405358	0.4	0.1	1.5	0.006
Otitis media	1820	403612	0.1	0.0	1.1	0.007
Atrial fibrillation	1104	404328	0.1	0.0	1.1	0.008
Sinoatrial node dysfunction (Bradycardia)	556	404876	0.1	0.1	1.1	0.009
Cancer of kidney and renal pelvis	1504	403928	0.1	0.0	1.1	0.009
Infective connective tissue disorders	104	405328	0.3	0.1	1.4	0.009
Other symptoms involving abdomen and pelvis	4125	401307	0.0	0.0	1.0	0.010
Bronchitis	731	404701	0.1	0.0	1.1	0.010
Chronic tonsillitis and adenoiditis	1122	404310	0.1	0.0	1.1	0.011
Cellulitis and abscess of fingers/toes	658	404774	0.1	0.0	1.1	0.012
Flat foot	197	405235	0.2	0.1	1.2	0.012
Myopathy	322	405110	0.2	0.1	1.2	0.012
Gouty arthropathy	215	405217	0.2	0.1	1.2	0.014
Bariatric surgery	42	405390	0.4	0.2	1.6	0.016

Coagulation defects	1190	404242	0.1	0.0	1.1	0.017

Abbreviations: PRS, polygenic risk score; SE, standard error; OR, odds ratio

[a] Information shown in the table includes significant associations (Bonferroni corrected p-value of p = 5.6 x 10^{-6} compared to individuals with normal range BMI) with class 3 obese BMI that are replicated with genome-wide PRS for obesity in UK Biobank and eMERGE cohorts (results significant to FDR-adjusted p-values of 0.019 and 0.015, respectively). Logistic regression models are adjusted for age, sex, site, and 10 principal components.

Supplemental Table 9. Effect Sizes of Phenotypes Associated with Obesity Clinically and with PRS for Obesity in eMERGE and UK Biobanks

Phenotype	Clinical Effect Size (per 1-SD BMI) [a]	eMERGE PRS Causal Effect Size [b]	UK Biobank PRS Causal Effect Size [c]
Obesity	3.78	2.06	1.83
Overweight, obesity and other hyperalimentation	3.34	1.85	1.82
Localized adiposity	2.81	1.76	1.62
Bariatric surgery	2.38	2.81	1.56
Sleep apnea	2.30	1.51	1.43
Type 2 diabetes with neurological manifestations	2.06	1.55	1.66
Polycystic ovaries	2.04	1.54	1.37
Chronic venous insufficiency	1.92	1.33	1.35
Diabetes type 2 with peripheral circulatory disorders	1.92	1.53	1.59
Type 2 diabetes with renal manifestations	1.90	1.51	1.36
Endometrial hyperplasia	1.88	1.30	1.12
Type 2 diabetes	1.88	1.52	1.36
Ovarian dysfunction	1.86	1.30	1.26
Diabetes mellitus	1.79	1.50	1.34
Arthropathy associated with neurological disorders	1.79	1.60	1.47
Type 2 diabetes with ophthalmic manifestations	1.75	1.59	1.22
Gouty arthropathy	1.66	1.26	1.22
Essential hypertension	1.63	1.32	1.21
Hypertension	1.63	1.33	1.21
Gout	1.61	1.30	1.25
Malignant neoplasm of uterus	1.60	1.35	1.16
Incisional hernia	1.60	1.39	1.44
Ventral hernia	1.59	1.44	1.19
Noninfectious disorders of lymphatic channels	1.58	1.17	1.28
Gout and other crystal arthropathies	1.58	1.25	1.23
Other disorders of intestine	1.58	1.30	1.11
Osteoarthrosis, localized, primary	1.56	1.19	1.21
Hypertensive heart and/or renal disease	1.56	1.21	1.23
Hypoventilation	1.56	1.33	2.01
Osteoarthritis; localized	1.54	1.16	1.22
Hidradenitis	1.54	1.50	1.73
Disorders of sweat glands	1.53	1.22	1.33
Calcaneal spur; Exostosis	1.52	1.19	1.27

Osteoarthrosis	1.52	1.21	1.23
Diabetes type 1 with peripheral circulatory disorders	1.51	1.35	1.49
Osteoarthrosis	1.49	1.21	1.22
Umbilical hernia	1.48	1.28	1.25
Edema	1.47	1.25	1.26
Cardiomegaly	1.45	1.22	1.23
Diabetic retinopathy	1.44	1.54	1.21
Disorders of lipoid metabolism	1.43	1.15	1.15
Hyperlipidemia	1.43	1.15	1.15
Mixed hyperlipidemia	1.43	1.19	1.44
Chronic ulcer of leg or foot	1.42	1.25	1.47
Hypertensive chronic kidney disease	1.42	1.21	1.23
Other chronic nonalcoholic liver disease	1.42	1.57	1.21
Vitamin B-complex deficiencies	1.40	1.20	1.17
Pilonidal cyst	1.40	1.33	1.21
Angina pectoris	1.40	1.16	1.17
Cellulitis and abscess of trunk	1.39	1.28	1.22
Other specified erythematous conditions	1.39	1.24	1.20
Unstable angina (intermediate coronary syndrome)	1.39	1.17	1.18
Congestive heart failure; nonhypertensive	1.39	1.24	1.22
Cholelithiasis with other cholecystitis	1.39	1.23	1.21
Congestive heart failure	1.38	1.27	1.27
Type 1 diabetes with neurological manifestations	1.38	1.39	1.35
Osteoarthrosis, generalized	1.38	1.19	1.21
Spinal stenosis of lumbar region	1.37	1.18	1.16
Abnormal glucose	1.37	1.06	1.25
Flat foot	1.36	1.17	1.24
Chronic liver disease and cirrhosis	1.35	1.54	1.14
Chronic Kidney Disease, Stage III	1.35	1.16	1.22
Benign neoplasm of adrenal gland	1.34	1.49	1.29
Ingrowing nail	1.34	1.14	1.22
Hypercholesterolemia	1.33	1.10	1.15
Chronic pulmonary heart disease	1.33	1.26	1.23
Varicose veins of lower extremity, symptomatic	1.32	1.15	1.27
Other arthropathies	1.32	1.14	1.11
Benign neoplasm of other endocrine glands and related structures	1.32	1.13	1.14
Type 1 diabetes	1.32	1.43	1.29

Fluid overload	1.31	1.32	1.25
Vitamin deficiency	1.31	1.17	1.13
Pulmonary heart disease	1.31	1.22	1.19
Swelling of limb	1.31	1.14	1.31
Acute pulmonary heart disease	1.31	1.17	1.18
Pulmonary embolism and infarction, acute	1.31	1.15	1.18
Heart failure NOS	1.30	1.23	1.21
Atrial fibrillation	1.30	1.17	1.10
Coronary atherosclerosis	1.30	1.22	1.13
Ischemic Heart Disease	1.29	1.20	1.15
Cholelithiasis	1.29	1.22	1.20
Infective connective tissue disorders	1.29	1.43	1.36
Other chronic ischemic heart disease, unspecified	1.29	1.18	1.20
Acquired spondylolisthesis	1.29	1.10	1.15
Atrial fibrillation and flutter	1.28	1.16	1.14
Spinal stenosis	1.28	1.13	1.16
Other abnormal glucose	1.28	1.07	1.23
Primary pulmonary hypertension	1.28	1.30	1.27
Wheezing	1.27	1.23	1.24
Sleep disorders	1.26	1.09	1.35
Varicose veins of lower extremity	1.26	1.12	1.08
Phlebitis and thrombophlebitis of lower extremities	1.26	1.16	1.28
Psoriasis	1.26	1.11	1.21
Cholelithiasis and cholecystitis	1.25	1.22	1.20
Superficial cellulitis and abscess	1.25	1.16	1.18
Precordial pain	1.25	1.10	1.16
Other forms of chronic heart disease	1.25	1.16	1.16
Chronic Kidney Disease, Stage IV	1.24	1.24	1.24
Bronchitis	1.24	1.09	1.12
Rupture of tendon, nontraumatic	1.24	1.14	1.09
Postoperative infection	1.24	1.27	1.18
Barrett's esophagus	1.23	1.23	1.08
Shortness of breath	1.23	1.17	1.18
GERD	1.22	1.08	1.09
Cirrhosis of liver without mention of alcohol	1.22	1.35	1.20
Left bundle branch block	1.22	1.19	1.07
Psoriasis vulgaris	1.22	1.11	1.25
Other local infections of skin and subcutaneous tissue	1.22	1.13	1.17

Degeneration of intervertebral disc	1.22	1.07	1.16
Other peripheral nerve disorders	1.21	1.15	1.24
Abdominal hernia	1.21	1.08	1.06
Thoracic or lumbosacral neuritis or radiculitis, unspecified	1.21	1.10	1.10
Other disorders of pancreatic internal secretion	1.21	1.20	1.20
Cancer of kidney and renal pelvis	1.20	1.16	1.08
Spondylosis without myelopathy	1.20	1.07	1.11
Other alveolar and parietoalveolar pneumonopathy	1.20	1.16	1.32
Urinary incontinence	1.20	1.07	1.13
Hypothyroidism NOS	1.20	1.11	1.15
Chronic renal failure	1.20	1.19	1.22
Sciatica	1.19	1.09	1.17
Cholecystitis without cholelithiasis	1.19	1.18	1.21
Esophagitis, GERD and related diseases	1.19	1.08	1.08
Chronic tonsillitis and adenoiditis	1.19	1.19	1.10
Hypothyroidism	1.19	1.10	1.15
Spondylosis and allied disorders	1.19	1.07	1.10
Disorders of adrenal glands	1.19	1.16	1.07
Displacement of intervertebral disc	1.19	1.11	1.12
Internal derangement of knee	1.19	1.10	1.13
Myocardial infarction	1.18	1.22	1.16
Intervertebral disc disorders	1.18	1.08	1.14
Sinoatrial node dysfunction (Bradycardia)	1.18	1.18	1.14
Renal failure	1.18	1.21	1.15
Acute renal failure	1.17	1.27	1.15
Other venous embolism and thrombosis	1.17	1.20	1.22
Diseases of esophagus	1.17	1.08	1.08
Other disorders of soft tissues	1.17	1.11	1.13
Diaphragmatic hernia	1.16	1.06	1.08
Primary/intrinsic cardiomyopathies	1.16	1.19	1.18
Cardiomyopathy	1.16	1.17	1.18
Varicose veins	1.16	1.10	1.08
Portal hypertension	1.16	1.41	1.15
Acute and chronic tonsillitis	1.16	1.14	1.08
Otitis externa	1.15	1.09	1.19
Nonspecific chest pain	1.15	1.06	1.16
Bundle branch block	1.15	1.16	1.09

Suppurative and unspecified otitis media	1.15	1.08	1.13
Peripheral enthesopathies and allied syndromes	1.14	1.04	1.06
Complication of internal orthopedic device	1.14	1.17	1.26
Acquired foot deformities	1.14	1.07	1.06
Phlebitis and thrombophlebitis	1.14	1.16	1.24
Cardiac arrest	1.14	1.21	1.17
Other symptoms of respiratory system	1.14	1.08	1.15
Other abnormal blood chemistry	1.13	1.07	1.22
Hypoglycemia	1.13	1.28	1.20
Respiratory abnormalities	1.13	1.21	1.23
Difficulty in walking	1.13	1.13	1.44
Hereditary and idiopathic peripheral neuropathy	1.13	1.12	1.20
Otitis media	1.13	1.06	1.08
Cellulitis and abscess of fingers/toes	1.13	1.09	1.13
Atrioventricular block	1.13	1.08	1.11
Liver abscess and sequelae of chronic liver disease	1.13	1.33	1.10
Synovitis and tenosynovitis	1.13	1.09	1.07
Other disorders of synovium, tendon, and bursa	1.13	1.06	1.08
Rheumatoid arthritis and other inflammatory polyarthropathies	1.12	1.07	1.17
Joint effusions	1.12	1.11	1.14
Nephritis; nephrosis; renal sclerosis	1.11	1.15	1.07
Cardiac arrest and ventricular fibrillation	1.11	1.18	1.15
Renal failure NOS	1.11	1.20	1.16
Osteomyelitis, periostitis, and other infections involving bone	1.11	1.21	1.21
Back pain	1.11	1.05	1.12
Peritoneal adhesions (postoperative) (postinfection)	1.11	1.29	1.15
Urinary calculus	1.11	1.07	1.08
Chronic ulcer of skin	1.11	1.24	1.28
Depression	1.10	1.16	1.10
Major depressive disorder	1.10	1.06	1.10
Osteomyelitis	1.10	1.26	1.23
Megaloblastic anemia	1.10	1.29	1.22
Myopathy	1.09	1.14	1.18
Mood disorders	1.09	1.15	1.09
Nonrheumatic aortic valve disorders	1.09	1.09	1.17
Symptoms and disorders of the joints	1.09	1.06	1.15
Cardiac conduction disorders	1.08	1.09	1.09

Other disorders of the kidney and ureters	1.08	1.16	1.07
Iron deficiency anemias, unspecified or not due to blood loss	1.08	1.13	1.13
Pulmonary collapse; interstitial and compensatory emphysema	1.08	1.17	1.11
Other symptoms involving abdomen and pelvis	1.07	1.06	1.05
Respiratory failure	1.07	1.25	1.27
Iron deficiency anemias	1.07	1.14	1.12
Other deficiency anemia	1.05	1.16	1.21
Coagulation defects	1.04	1.11	1.09
Staphylococcus infections	1.03	1.20	1.21
Respiratory failure, insufficiency, arrest	1.03	1.24	1.25
Cardiac dysrhythmias	1.03	1.06	1.11
Chronic bronchitis	1.02	1.18	1.22
Obstructive chronic bronchitis	1.02	1.23	1.23
Bacterial infection NOS	1.01	1.17	1.17
Septicemia	0.99	1.21	1.13
Decubitus ulcer	0.94	1.28	1.15
Chronic airway obstruction	0.91	1.15	1.18

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); SD, standard deviation; OR, odds ratio; PRS, polygenic risk score

[a] Clinical effect size determined using the association of linear mean BMI with phenotypes. Effect size = exp(beta*population SD). SD of BMI in population is 6.99 kg/m^2. Logistic regression models are adjusted for age, sex, and self-reported race.

[b] Logistic regression models are adjusted for age, sex, site, and 10 principal components. Obesity genome-wide PRS scaled to mean of 0 and SD of 1.

[c] Logistic regression models are adjusted for age, sex, and 10 principal components. Obesity genome-wide PRS scaled to mean of 0 and SD of 1.

Supplemental Table 10. Disease Attributable to Obesity with Normalization of BMI in Obese Individuals [a]

Phenotype	Attributable Disease Events	Attributable Risk Proportion (%)
Obesity	39910	98
Overweight, obesity and other hyperalimentation	43785	96
Bariatric surgery	3829	94
Sleep apnea	17624	87
Ovarian dysfunction	2602	85
Type 2 diabetes with neurological manifestations	4741	84
Diabetes type 2 with peripheral circulatory disorders	745	80
Arthropathy associated with neurological disorders	317	79
Type 2 diabetes with renal manifestations	3116	78
Hidradenitis	338	77
Endometrial hyperplasia	340	75
Type 2 diabetes with ophthalmic manifestations	1635	75
Gouty arthropathy	1185	75
Calcaneal spur; Exostosis	381	72
Type 2 diabetes	25512	72
Disorders of sweat glands	565	70
Gout	3349	70
Incisional hernia	1411	70
Chronic venous insufficiency	736	69
Diabetes mellitus	25916	68
Gout and other crystal arthropathies	3367	68
Osteoarthrosis, localized, primary	7119	66
Cholelithiasis with other cholecystitis	480	64
Osteoarthritis; localized	8147	64
Ventral hernia	931	62
Pilonidal cyst	159	62
Umbilical hernia	720	61
Benign neoplasm of adrenal gland	251	61
Malignant neoplasm of uterus	580	60
Diabetic retinopathy	1349	60
Osteoarthrosis	7360	59
Mixed hyperlipidemia	9874	57
Angina pectoris	2213	57
Noninfectious disorders of lymphatic channels	693	55
Osteoarthrosis	13270	55

Spinal stenosis of lumbar region	2523	55
Benign neoplasm of other endocrine glands and related structures	865	55
Unstable angina (intermediate coronary syndrome)	2018	54
Hypertensive heart and/or renal disease	5126	54
Flat foot	217	53
Other chronic nonalcoholic liver disease	3281	53
Diabetes type 1 with peripheral circulatory disorders	71	53
Other disorders of intestine	2856	52
Acquired spondylolisthesis	904	50
Edema	5420	50
Cellulitis and abscess of trunk	947	48
Osteoarthrosis, generalized	1568	48
Hyperlipidemia	22612	48
Disorders of lipoid metabolism	22847	48
Essential hypertension	33929	48
Type 1 diabetes with neurological manifestations	464	48
Cholelithiasis	1746	48
Hypercholesterolemia	6118	47
Hypertension	34763	47
Other specified erythematous conditions	176	47
Spinal stenosis	2782	47
Cardiomegaly	3554	47
Chronic liver disease and cirrhosis	3291	46
Other arthropathies	2051	46
Chronic Kidney Disease, Stage III	1661	45
Abnormal glucose	4234	45
Psoriasis	859	45
Hypertensive chronic kidney disease	2341	45
Chronic ulcer of leg or foot	786	45
Ingrowing nail	408	44
Rupture of tendon, nontraumatic	584	44
Vitamin B-complex deficiencies	1031	43
Type 1 diabetes	2683	42
Precordial pain	461	42
Other alveolar and parietoalveolar pneumonopathy	363	41
Cholelithiasis and cholecystitis	1914	41
Other chronic ischemic heart disease, unspecified	2153	41
Barrett's esophagus	212	40

Thoracic or lumbosacral neuritis or radiculitis, unspecified	1621	40
Sciatica	756	40
Vitamin deficiency	4265	39
Psoriasis vulgaris	529	39
Hypoventilation	217	39
Pulmonary embolism and infarction, acute	775	39
Displacement of intervertebral disc	1318	39
Wheezing	674	39
Degeneration of intervertebral disc	3139	38
Acute pulmonary heart disease	860	38
Other disorders of pancreatic internal secretion	197	38
Other peripheral nerve disorders	2788	38
Internal derangement of knee	1598	38
Phlebitis and thrombophlebitis of lower extremities	228	38
Congestive heart failure; nonhypertensive	5115	38
Coronary atherosclerosis	7995	37
Fluid overload	1412	37
Swelling of limb	1210	37
Spondylosis without myelopathy	2522	37
Bronchitis	377	37
Varicose veins of lower extremity, symptomatic	301	37
Ischemic Heart Disease	9207	36
Other abnormal glucose	2549	36
Sleep disorders	3892	36
Congestive heart failure	3357	36
Spondylosis and allied disorders	2803	35
Pulmonary heart disease	2050	35
Chronic pulmonary heart disease	1179	35
Cancer of kidney and renal pelvis	630	35
GERD	6945	35
Primary pulmonary hypertension	236	34
Intervertebral disc disorders	4148	34
Other forms of chronic heart disease	1526	34
Chronic tonsillitis and adenoiditis	434	33
Heart failure	1045	33
Left bundle branch block	267	33
Cholecystitis without cholelithiasis	322	33
Diaphragmatic hernia	849	33

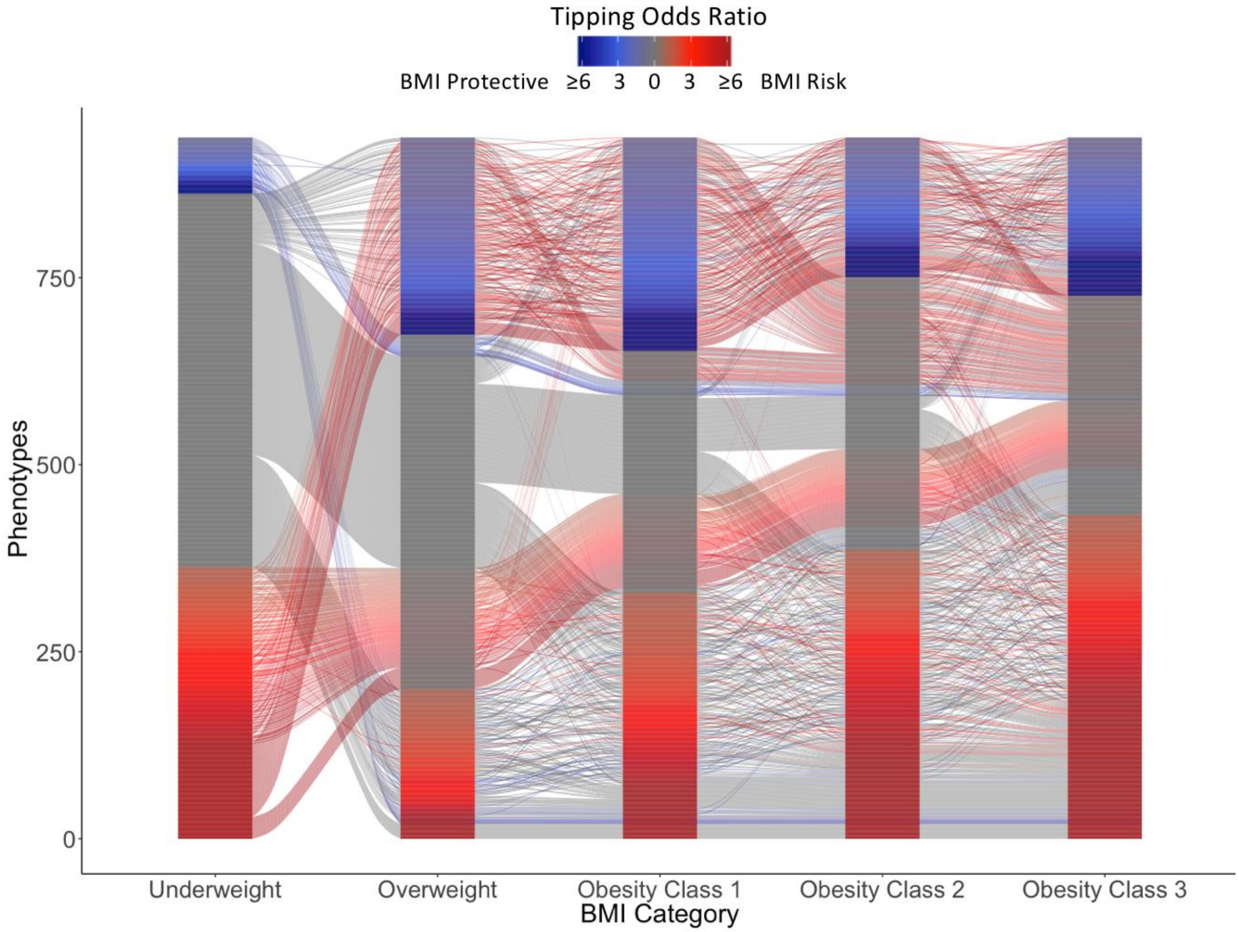
Urinary incontinence	1456	32
Cirrhosis of liver without mention of alcohol	1103	32
Atrial fibrillation	3628	31
Esophagitis, GERD and related diseases	6761	31
Chronic Kidney Disease, Stage IV	504	31
Other local infections of skin and subcutaneous tissue	548	31
Disorders of adrenal glands	662	31
Atrial fibrillation and flutter	3633	30
Superficial cellulitis and abscess	3106	30
Hypothyroidism	4139	30
Portal hypertension	410	30
Hypothyroidism	4743	30
Myocardial infarction	1795	30
Peripheral enthesopathies and allied syndromes	4001	29
Varicose veins of lower extremity	307	29
Joint effusions	639	29
Shortness of breath	6060	28
Acquired foot deformities	1186	28
Abdominal hernia	2234	28
Diseases of esophagus	6642	28
Suppurative and unspecified otitis media	604	28
Acute and chronic tonsillitis	440	27
Otitis externa	354	27
Hereditary and idiopathic peripheral neuropathy	1018	26
Infective connective tissue disorders	73	26
Other disorders of synovium, tendon, and bursa	1960	26
Synovitis and tenosynovitis	985	26
Complication of internal orthopedic device	490	26
Chronic renal failure	2373	26
Other abnormal blood chemistry	1069	25
Rheumatoid arthritis and other inflammatory polyarthropathies	1130	25
Other venous embolism and thrombosis	1265	24
Otitis media	695	23
Nonspecific chest pain	8202	23
Other disorders of soft tissues	78	23
Sinoatrial node dysfunction (Bradycardia)	546	23
Bundle branch block	301	23
Primary/intrinsic cardiomyopathies	998	23

Cardiomyopathy	1156	22
Liver abscess and sequelae of chronic liver disease	487	22
Cardiac arrest	152	21
Renal failure	3332	21
Phlebitis and thrombophlebitis	192	20
Nephritis; nephrosis; renal sclerosis	383	19
Difficulty in walking	314	19
Urinary calculus	838	19
Atrioventricular block	309	19
Back pain	4508	18
Acute renal failure	1634	18
Major depressive disorder	1140	18
Other symptoms of respiratory system	7856	17
Cardiac arrest and ventricular fibrillation	171	17
Symptoms and disorders of the joints	1492	17
Depression	2912	16
Varicose veins	223	16
Cellulitis and abscess of fingers/toes	203	16
Mood disorders	3212	15
Peritoneal adhesions (postoperative) (postinfection)	73	14
Nonrheumatic aortic valve disorders	487	13
Other symptoms involving abdomen and pelvis	575	12
Megaloblastic anemia	129	12
Respiratory abnormalities	309	12
Cardiac conduction disorders	1021	11
Osteomyelitis, periostitis, and other infections involving bone	246	11
Myopathy	154	11
Osteomyelitis	194	11
Other disorders of the kidney and ureters	734	10
Renal failure NOS	142	10
Pulmonary collapse; interstitial and compensatory emphysema	1177	9
Iron deficiency anemias, unspecified or not due to blood loss	314	8
Hypoglycemia	61	7
Iron deficiency anemias	290	6
Chronic ulcer of skin	206	6
Other deficiency anemia	46	3
Coagulation defects	142	3
Chronic bronchitis	49	3

Respiratory failure	154	2
Cardiac dysrhythmias	551	2
Obstructive chronic bronchitis	18	1

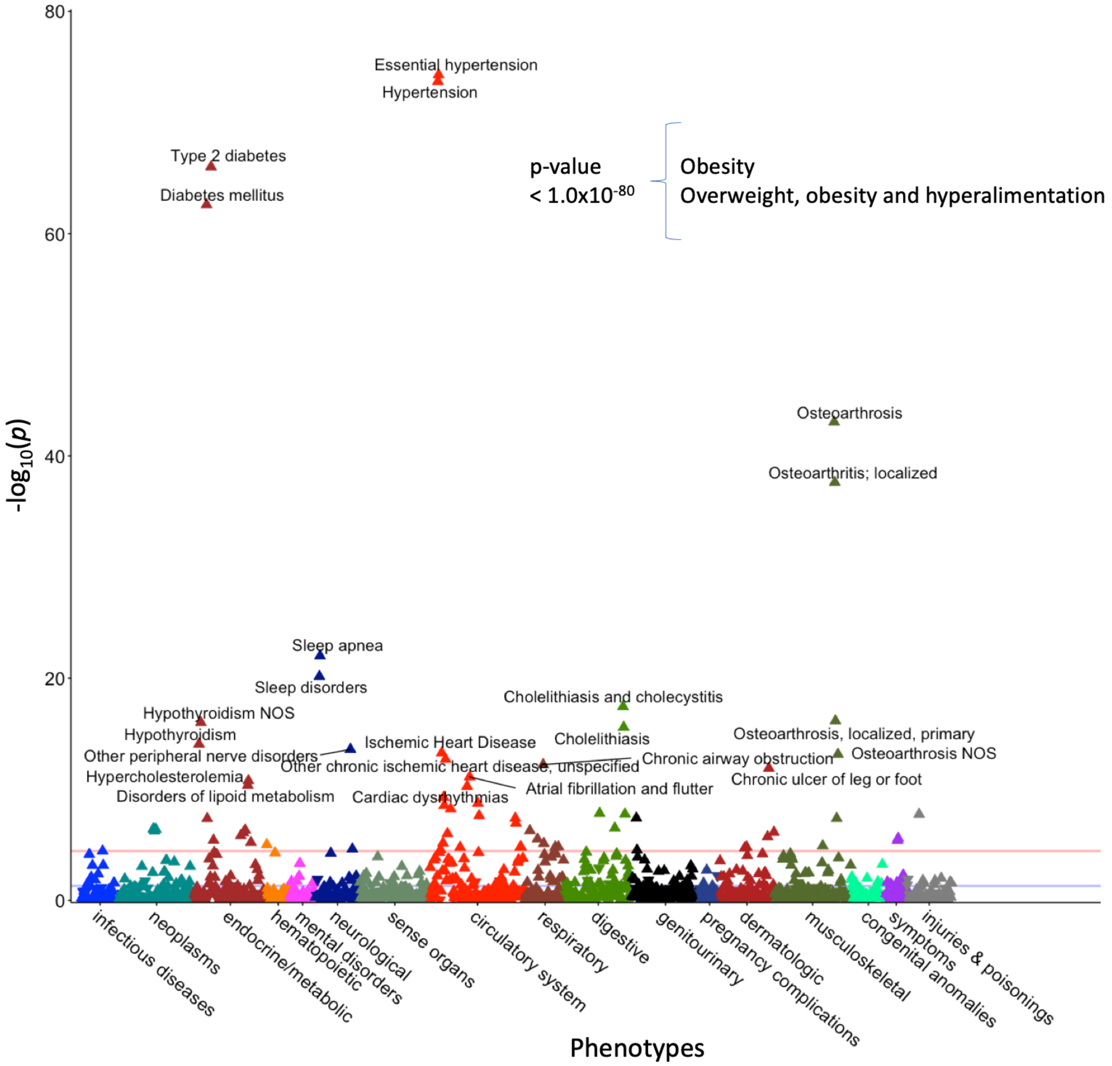
[a] Information shown in the table includes the number of predicted attributable disease events if individuals with class 1-3 obesity have BMI normalized for phenotypes showing association in all 3 cohorts (class 3 obesity in clinical cohort and obesity PRS in eMERGE and UK biobanks). Logistic regression models are adjusted for age, sex, and self-reported race.

Supplemental Figure 1. Tipping Point Analysis for Possible Unmeasured Confounders



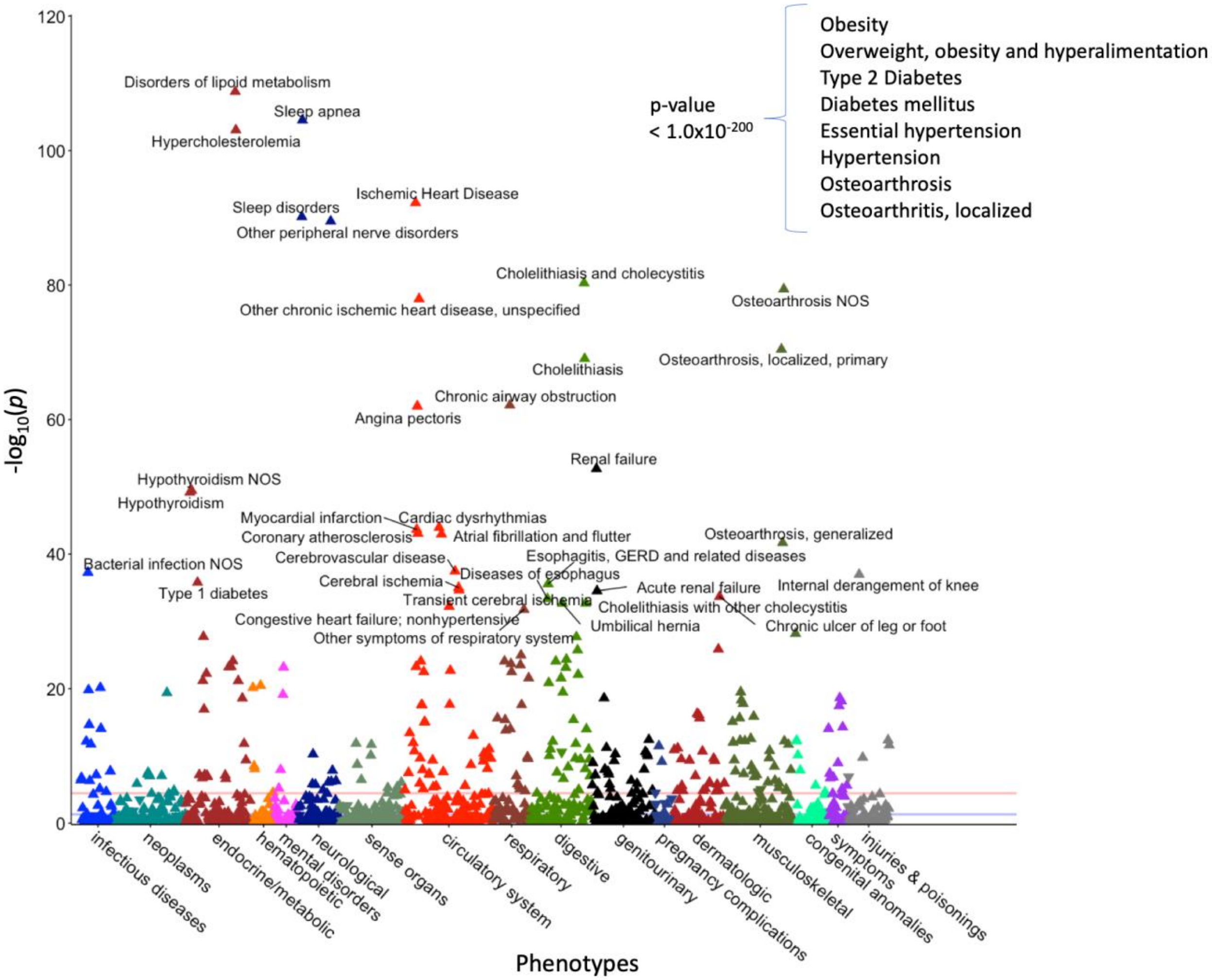
Trends of estimated odds ratios of confounders needed to change significant phenotype associations. Control is normal BMI category. The prevalence rate difference between the specific BMI category and normal BMI is 10%. Only phenotypes meeting Bonferroni significance threshold (p = 5.6 x 10^{-6}) in at least one of the BMI categories are shown. Gray represents non-significant findings. Most phenotypes with association with class 3 obesity would need an unmeasured confounder with OR greater than 2 to bias association findings. Among the 644 phenotypes associated positively or negatively with class 3 obesity, 528 (82.0%) required the hypothetical unmeasured binary confounder to have an OR>2 to change the significance conclusion when its prevalence rate difference between class 3 obesity and normal BMI was 10%. The percentages were 77.2%, 67.4%, 57.2%, and 79.0% for class 2 obesity, class 1 obesity, overweight, and underweight, respectively.

Supplemental Figure 2. Association of Obesity 97-SNP PRS with Diseases in PheWAS in UK Biobank Cohort



Blue horizontal line represents *p* = 0.05. Red horizontal line represents Bonferroni significance threshold *p* = 3.1 x 10^{-5}. Point direction relates to directionality of odds ratio: upward triangles are associated with increased risk for patients while downward triangles are associated with decreased risk.

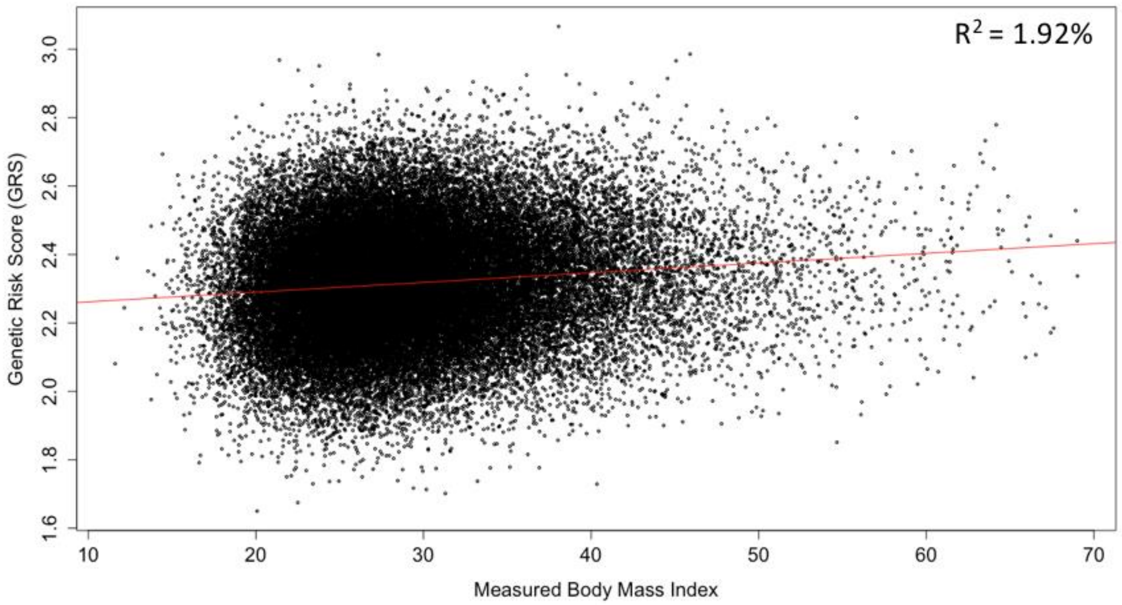
Supplemental Figure 3. Association of Genome-wide PRS with Diseases in PheWAS in UK Biobank Cohort



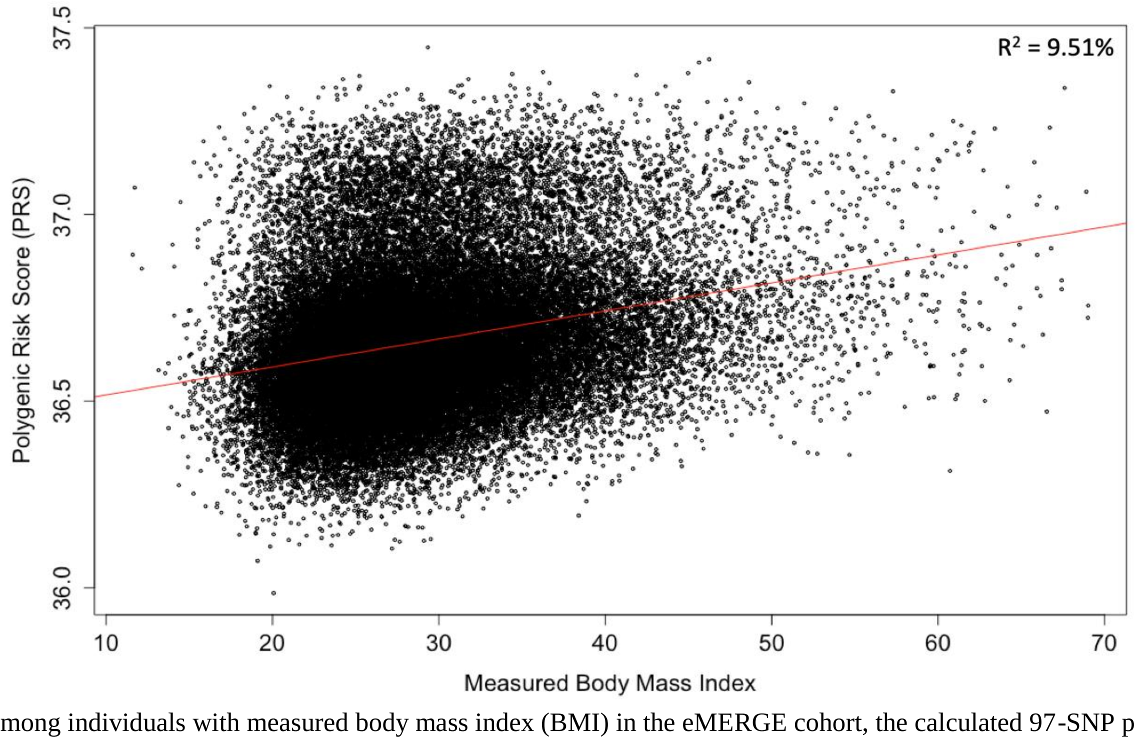
Blue horizontal line represents $p = 0.05$. Red horizontal line represents Bonferroni significance threshold $p = 3.1 \times 10^{-5}$. Point direction relates to directionality of odds ratio: upward triangles are associated with increased risk for patients while downward triangles are associated with decreased risk.

Supplemental Figure 4. A. 97-SNP Obesity PRS Compared to Measured Body Mass Index B. Genome-wide Obesity PRS Compared to Measured Body Mass Index

A.

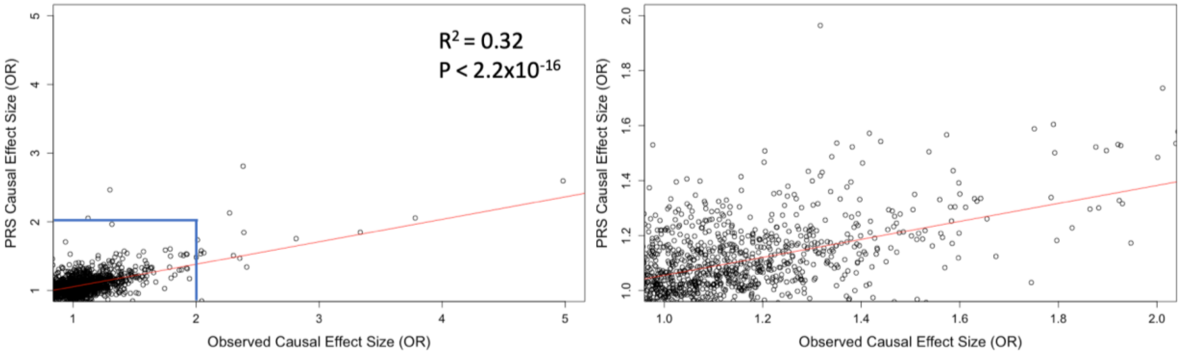


B.



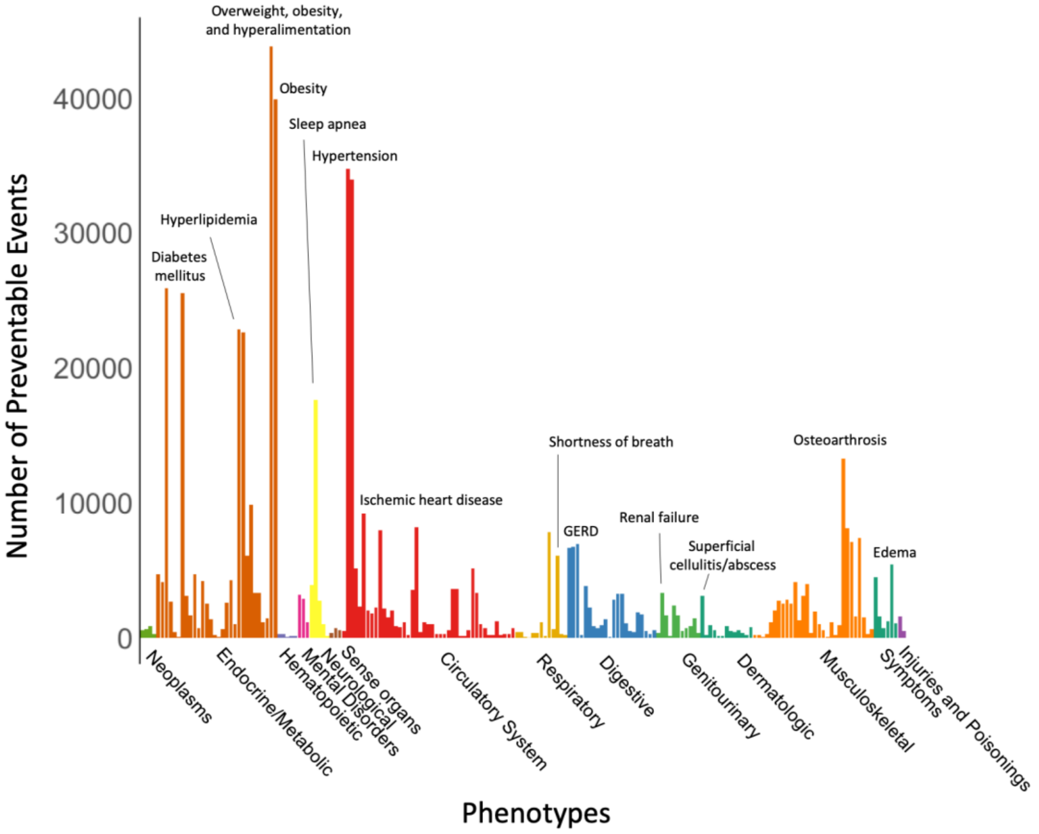
Among individuals with measured body mass index (BMI) in the eMERGE cohort, the calculated 97-SNP polygenic risk score for obesity explained 1.92% of the variance in mean BMI and the calculated genome-wide polygenic risk score for obesity explained 9.51% of the variance in mean BMI.

Supplemental Figure 5. Clinically Observed versus Genome-wide Obesity PRS PheWAS Causal Effect Sizes for All Phenotypes



1816 total phenotypes with each dot representing the observational versus genomic effect size (in the eMERGE cohort). Red line represents linear regression. Adjusted $R^2 = 0.32$.

Supplemental Figure 6. Number of Disease Events Attributable to Obesity with Normalization of Obesity Classes 1-3 to Normal BMI



Phenotypes shown are 199 phenotypes that were associated with class 3 obesity in the clinical cohort and obesity polygenic risk score in both eMERGE and UK biobank cohorts. Example phenotypes are annotated.

CHAPTER VII

Summary

The body of research in this dissertation demonstrates that secondary use of clinical data within the EHR along with linkage of genetic data provides an efficient method for aggregating otherwise disparate information for research. Use of EHR data has proven a powerful method for elucidation of genomic influences on diseases, traits, and drug-response phenotypes and will continue to have increasing applications in large cohort studies with the goal of driving personalized medicine. Through the research and findings in this dissertation, I have made significant advancements to the multidisciplinary field of biomedical informatics by elucidating patterns of clinically enigmatic disease processes using novel phenotyping methods and genome-wide risk association analyses.

Contributions to EHR Phenotyping

The first main contribution this body of work makes to the field of biomedical informatics is development of phenotyping methods to demonstrate patterns of disease. Through the use of different phenotyping methods and dissimilar disease processes, we were able to illustrate the ability to use data that was curated for clinical medicine to improve the current knowledge on both common and rare diseases.

Systemic loxoscelism is extremely rare. In this research, we aimed to determine the incidence and refine the phenotype of this uncommon disease. The first phenotyping method utilized in this study used a programmed search of the medical record, by which text data of “loxoscelism” was extracted from clinical documents. Over a 25-year period at Vanderbilt Medical Center in Nashville, TN, a location endemic to the brown recluse spider, there were only 373 text occurrences of “loxoscelism” in the clinical records. Through the use of regular expressions, negation was identified in 168 of these individuals and, after clinical review, true systemic moderate or severe loxoscelism classified only 57 individuals. Thus, this rare disease only occurred 2-3 cases/year at a large academic medical center in an endemic region, yet this study represents the largest cohort of individuals with systemic loxoscelism to date. It would be virtually impossible to, without the use of informatics techniques and computerized text mining, to retrospectively identify these individuals. Denny et al. previously demonstrated the ability to show that the phenotype of hippus, or repetitive oscillation of the pupils, is very rare in the literature, but a comprehensive search of the EHR can identify much larger numbers.(1) Our evaluation for individuals with loxoscelism again demonstrates that retrospectively repurposing the EHR can identify rare diseases in numbers previously unachievable.

Unfortunately, while the addition of the requirement for “brown recluse”, specific ICD codes, or other clinical markers could have improved sensitivity in our search for individuals with loxoscelism, it did not improve the phenotyping accuracy. Thus, manual review of the deidentified patient records was chosen for the accuracy desired in this study. Due to the sample size, this was feasible. While a combination of text mining and clinical review resulted in a highly accurate phenotype for systemic loxoscelism, this method would be challenging if there were thousands of individuals with “loxoscelism” in the patient records.

Larger cohorts often require more high-throughput methods, such as applying Boolean phenotyping algorithms that can be applied across institutions and EHRs. Clinical expertise nevertheless is required for highly accurate phenotyping during the process of curating a precise phenotyping algorithm. This method performs well for applying a single phenotyping algorithm to a large sample size, but not for the capture of a multitude of phenotypes. In our search for individuals with loxoscelism, use of negation and other markers included in Boolean phenotype algorithm creation and validation would have the ability to reach comparative precision.

The second phenotyping method in this study was to evaluate for phenotypes seen in individuals diagnosed with loxoscelism using a phenome-wide phenotyping approach, PheWAS. Analogous to a

GWAS, PheWAS leverages the breadth of phenotypes in the EHR to perform systematic interrogation for associations with an independent variable, typically a genotype. In addition to being applied to large cohorts, PheWAS can utilize thousands of phenotypes determined by billing codes at a scale for which manual curation and validation of individual phenotypes is not practical.(2–4) This use of a phenome-wide approach to characterization of this rare disease allowed for an agnostic approach to disentangling the presentation of this severe, but challenging to diagnose, illness.

The final two aims of this dissertation expanded on the use of phenome-wide approaches by application to large sample sizes to identify phenotypes and phenotypic associations with PheWAS. In the third aim of this dissertation, grouping of ICD codes into phecodes was used for phenotyping of three separate common postoperative complications, demonstrating that associations with phecodes representing postoperative complications could be replicated across EHRs and using both clinical and genomic predictors. In this study, both postoperative infection and incisional hernia demonstrated an association with clinical BMI and genetically-determined BMI in different cohorts. Further, this method showed translation of the phecode mapping and association findings across EHRs and medical systems. On an even larger scale, the fourth aim demonstrated replicability of 199 phenotype associations with BMI and genetic risk for obesity across 3 separate cohorts. It is worth noting that billing codes derived from ICD-9 were used in 2 of the cohorts (Vanderbilt clinical cohort and eMERGE cohorts) and ICD10 codes were used in the UK Biobank cohort. Regardless of the ICD billing code system used, mapping of the ICD code to a phecode was performed to delineate phenotypes. This is one of the first studies showing replication of association findings with phecodes derived from different ICD coding structures, demonstrating the validity of the phecode mapping systems. This is important because as ICD codes change over time, phecodes are more constant and the only update required is additional mapping. While the creation of mapping systems for billing codes to a single phecode system can be laborious, once completed it allows for the use of billing codes from different sources in research.

It has been previously well demonstrated that billing codes in the medical record have variable accuracy in representing the diagnoses of patients, with one study showing a range of positive predictive values (PPVs) for ICD codes from 0.12 to 0.56 across ten diseases.(5) In a recent study by Cohen et al., 629 of 780 patients with congenital heart disease were correctly categorized using ICD-10 codes (sensitivity 0.81, 95% confidence interval 0.78-0.83), with a high degree of specificity of 0.99 (95% confidence interval 0.99-1).(6) Another study showed fairly high accuracy (sensitivity and specificity of 82.3% and 78.3%, respectively) of ICD-9 codes for the identification of surgical patients with sepsis.(7) In contrast, among 4,400 individuals in a community cohort, the use of ICD-9 codes for peripheral arterial disease demonstrated poor sensitivity of 38.7% but high specificity (92.0%).(8) Further, in other populations, such as trauma, the sensitivity of diagnosis codes is even lower. Evaluation of all patients with proximal tibia fractures in the 2011 and 2012 American College of Surgeons' National Trauma Data Bank showed that ICD-9 codes compared to manual chart review for 12 comorbidities ranged in sensitivities from 18.8% for previous myocardial infarction to 2.4% for alcoholism.(9)

It is well accepted that ICD codes generally have low specificity but are highly sensitive for diseases, as a clinician may bill an ICD code for a diagnosis based upon clinical suspicion rather than confirmation of disease.(10, 11) However, as discussed above, sensitivity for some phenotypes may be lower, especially for phenotypes without a specifically representative ICD code, diseases that are challenging to clinically diagnose, or phenotypes with clinical similarities to others. For example, it is not uncommon for a patient to be diagnosed with ulcerative colitis and later be found to have inflammatory bowel disease outside the colon or rectum, representative of a Crohn’s disease diagnosis instead.(10) Thus, the phenotyping using billing codes alone is highly variable, with much of the dependency being the population studied and the phenotype of interest.

ICD codes are generated for billing purposes and similar to other data within the EHR, must be captured retrospectively and repurposed for research. While EHRs contain a wealth of extractable information for phenotype classification, their interface and the data generated within them are used primarily for clinical care and reimbursement, typically with little consideration towards research impact. The secondary use of EHRs for clinical, genomic, and pharmacogenomics discovery can be challenged by variable accuracy resulting from clinical uncertainty, omissions, or billing errors as discussed above, along with lack of standardization, irregular follow-up, incompleteness of patient records, and significant amounts of unstructured information. Due to the abscense of EHR centralization, the length and depth of a patient's record can vary greatly due to where a patient receives his or her care, with patients often seeing multiple disconnected providers within a region. A study to evaluate the effect of potential data fragmentation on the accuracy of a phenotyping algorithm for type 2 diabetes found that almost one-third of cases were missed if EHR data from only a single site was used.(12) While completeness of the EHR is difficult to define, it is important for researchers to understand the likely limitations of the data and how it may affect study findings.(13) Billing codes are often used secondarily for research; thus, quality assessments to understand their strengths and weaknesses along with methods to validate research findings are critical to the advancement of clinical research informatics.

Due to the variability in billing code accuracy in studies in which billing codes are used for phenotyping, each individual study must incorporate methods to minimize both the phenotyping inaccuracy and its effects on results. This body of work has validated that a successful method of reducing inaccuracy in billing codes is grouping of ICD codes into clinically relevant categories.

In the two final aims of this work, we demonstrated the ability of phecodes to overcome the limitations of individual ICD codes and validated their ability to provide enhanced statistical power by identifying genomic associations with disease as well as to reflect actual clinical disease patterns seen in practice. Billing codes were captured from the deidentified clinical record of individuals and grouped into phecode categories which do not have the granularity of the original billing codes.(2, 14) However, the granularity of the original code may not be necessary for the analysis to be performed. For example, in the final aim of this dissertation, we found a strong association between genetic risk for obesity and major depressive disorder. This single phecode for major depressive disorder incorporates 14 ICD codes. In this analysis, we did not need the granularity of the individual ICD codes, as it was not relevant to the research study if the individual had “Major depressive disorder, single episode”, “Major depressive disorder, recurrent episode”, or “Major depressive disorder, single episode, moderate degree”, etc. This grouping allows for a single test to be performed, rather than 14 separate association tests, which would result in significantly reduced statistical power. This example also demonstrates why phecode groupings have been shown to better align with clinical diseases in practice.(15) It is much more likely for a clinician to accurately diagnose “Major depressive disorder” than to accurately determine the number of episodes or degree of severity of the episode. Grouping of billing codes into clinically relevant categories to limit the number of association testings performed in these large analyses gives significantly greater power to find associations. While these phecodes do not have the detail provided by clinical expertise or well-curated phenotypes developed using a combination of concepts from multiple locations in the clinical record, phecodes do provide the ability to evaluate the entire phenome of an individual with ease.

In addition to grouping of ICD codes, another method demonstrated by this research to improve the accuracy of ICD codes is the requirement of 2 or more ICD codes on separate dates for mapping to a single phecode in the PheWAS analyses. We employed this approach in our studies to improve specificity at a minimum to incorporate only codes that are present within the clinical chart at least twice on separate days.(5) In an analysis of ten diseases, using two or more ICD codes (billed on different days) improved the average positive predictive value from 0.71 to 0.84; however, with increasing requirement for ICD code instances, there was a reduction in sensitivity.(5) This suggested the requirement for two or three codes, depending on the phenotype, will maximize precision. We demonstrated the validity of these methods by showing replicability of associations findings across cohorts and using both clinical and genomic predictors.

Lastly, large sample sizes and a phenome-wide approach allows for accrual of adequate case and control numbers to overcome some degree of inaccurate phenotype labeling for identification of associations in research. Validation of the association results in multiple cohorts, spanning providers and institutions, can reduce the likelihood that findings are secondary to biases of practice patterns or EHRs at a single institution. Further demonstrated in the final two aims, the use of both a trait and an instrumental variable such as genetics as the predictor in a PheWAS analysis can substantiate results as well.

Contributions to EHR-linked Genomic Analyses

The second contribution that this dissertation makes to the field of biomedical informatics is that it is the first application of genome-wide risk scores in a phenome-wide approach, demonstrating its ability to greater define disease risk and associations. These methods represent a novel approach to combining genome and phenome-wide data, and the described research makes contributions the component disciplines and application areas of genomics, phenotyping, clinical research informatics, and clinical medicine.

This dissertation uses genome-wide polygenic risk scores for BMI to identify associations with phenotypes. We showed that a genome-wide PRS correlated much more strongly with clinically observed BMI than a PRS composed of only 97 SNPs. The genome-wide PRS explained 9.51% of the variance in BMI compared to 1.92% for the 97-SNP PRS. Pearson correlation coefficient also showed a much stronger correlation between observed BMI and the genome-wide PRS (0.26 [95% CI 0.25-0.27]) compared to the 97-SNP PRS (0.11 [95% CI 0.10-0.12]). Prior studies have demonstrated that genome-wide PRSs have greater correlation with the trait in coronary artery disease, atrial fibrillation, type 2 diabetes, and inflammatory bowel disease.(16) However, for these 4 diseases, only 2-4% of the variance was explained by the trait-specific genome-wide polygenic risk score. This suggests the PRS for BMI utilized in this study was very strongly correlated with BMI.

Not only does broader inclusion of the genome for obesity explain almost 5x the variance in BMI compared to a 97-SNP PRS, it has the benefit of actually performing better in genetic association analyses. Genome-wide PRSs have been applied to obesity in prior studies; however, a genome-wide PRS for BMI has not been assessed for association with phenotypes in a phenome-wide approach. We were able to demonstrate disease associations with BMI across all disease categories and encompassing almost 200 phenotypes, including associations not previously described in the literature, which represent contributions to clinical medicine. It is important to note that phenotypes can have some degree of redundancy, especially when utilizing a hierarchical coding system such as the phecode mapping system. In our study, we reviewed each of the disease associations and confirmed that 95 of the 199 phenotypes associated with clinical BMI and genetically-determined BMI are unique “parent” phenotypes. This is unprecedented compared to other analyses of comorbidities associated with obesity. The genome-wide PRS found 2.2x more associations than the 97-SNP PRS (296 compared to 135), suggesting increased power of the genomic instrument with greater inclusion of the genome. Future studies to assess the association of phenotypes with a genetic risk score for a trait should not focus only on those SNPs known to be associated with that trait, but instead consider the use of genome-wide PRSs.

PRSs for diseases have been applied in a phenome-wide approach in prior studies. The first of these explored the association of PRS calculations for various cancers with the phenome.(17, 18) In one study, different thresholds of p-value significance in GWAS studies and LDpred models were used in construction of the PRS for various cutaneous cancers.(18) In contrast to our findings, they concluded that as the p-value threshold incorporated less significant GWAS SNPs, the predictive ability of the PRS to find associations was not improved.

Another remarkable finding of this body of work was the strong concordance between observational effect sizes for the association of BMI with phenotypes and genomic effect sizes for the association of the genome-wide PRS with phenotypes ($R^2 = 0.54$). This finding further substantiates the strong association between these disease phenotypes and obesity. In review of the literature, we cannot identify a study performing a similar correlation test of the clinical PheWAS to the genetically-predicted PheWAS between two populations as performed in this study to compare the observational and PRS effect sizes. However, a couple of reference studies have shown that the predicted phenomic heritability from family history has an $R^2 \sim 0.3$ (19) and a comparison of effect sizes for the PheWAS of systolic blood pressure genetic risk score between white individuals in the Million Veterans Program and the UK Biobank was R^2 ~ 0.5.(20)

In this study, we did not address the non-linearity of some genetic associations (e.g., phenotypes that may be increased in both underweight and overweight populations). The observational cohort showed several phenotypes with evidence of nonlinear associations with BMI, including asthma and gastroesophageal reflux disease which were increased in all BMI categories compared to individuals with a normal BMI. Nonlinearity would generally bias towards the null hypothesis, thus some associations with obesity may not be demonstrated. Further, all BMI values were included in the determination of each individual's median BMI in the PheWAS analysis. As we included all recorded BMI values, we were unable to assess the temporal relationship between observations and BMI. Many conditions may be entered into the EHR after they actually occurred. An analysis of diagnosis timing with respect to BMI would need to be considered on an individual phenotype basis, evaluating each as chronic or acute, and considering ways to differentiate between new-onset and potentially newly entered but extant diseases, instead of the more phenome-wide approach deployed in this research. For this dissertation, my aim was to focus on the increased phenotypic risk associated with obesity and genetic risk for increased BMI over the course of a lifetime, but further research should focus on phenotypes increased in underweight populations and the temporality of BMI changes with diseases.

As this research demonstrates a significant association of genetic risk for obesity with phenotypes, it suggests a possible causal role for obesity in the occurrence of these diseases. Mendelian randomization is the process of using genomic variants associated with a trait as an instrumental variable in association studies. The main advantage of using Mendelian randomization is that genetic variants are not subject to the same biases as traditional observational studies due to their random assortment in the population and determination at conception, thus allowing for causal inferences.(21, 22) There are three main assumptions that underlie the MR approach and must hold true for a causal role to be concluded. The first is that the SNPs selected for the genetic instrument must be associated with the trait, in this circumstance BMI. We found a very strong association of the genome-wide PRS with BMI, thus the first assumption is met in this study. The second assumption is that the genetic instrument is not associated with confounders, and the third assumption is that the instrument is associated with disease exclusively through their effect on the trait (i.e., obesity).(23) These assumptions are included to ensure that pleiotropic effects of the SNPs are not leading to confounding or apparent associations. We were able to demonstrate that adjustment for observed BMI diminished the statistical association between the genetic instrument and many of the phenotypes, which supports BMI as the causal mediator of the association between the PRS and phenotypes, arguing against a major contribution from pleiotropic effects. However, as evaluation for pleiotropy on a genome-wide and phenome-wide scale is not practical, this study does not claim a causal association.

There are many prior studies that have suggested that genetic risk for obesity, and thus elevated BMI, has a causal role in development of comorbidities through the use of other techniques such as MR-Egger analysis.(24, 25) MR-Egger is a linear regression of estimated SNP effects for the risk allele on exposure against the corresponding estimates of SNPs on the outcome weighted by the inverse variance of the SNP on outcome effect estimates.(24) This approach can demonstrate the presence of pleiotropic SNPs in the instrument that can bias causal estimates, increase the rate of false positives (i.e., type I errors), or introduce bias to the null. MR-Egger, however, provides a causal estimate (making certain assumptions) that is robust to such pleiotropy. In this study, we further explored the potential impact of directional pleiotropy by conducting MR-Egger regression to estimate the average directional pleiotropic effect of the 97 SNPs for a strongly associated phenotype, coronary artery disease, to evaluate for the causal effect of BMI. We found that the intercept from MR-Egger regression was not significantly different from zero (0.00, 95% CI −0.01, 0.01; P = 0.96) suggesting no directional pleiotropy. Further studies should focus on the use of MR-Egger methods in a genome and phenome-wide approach, requiring a methodical evaluation using effect sizes of each SNP for the instrumental variable trait and effect sizes of each SNP on all phenotypes.

Contributions to Translational Medicine

EHR-based biobanks have the potential to integrate genomic data with billing codes, medication receipt, laboratory results, and textual data, thus allowing for greater coverage of the phenome in genomic association studies. This depth of data can be used to drive personalized medicine approaches for more targeted disease treatment, earlier disease detection, identification of risk factors, and prevention strategies. The goal of personalized medicine, a term often used interchangeably with precision or individualized medicine, is to tailor medical decisions and practices to the individual based on each patient's unique subset of factors, including genetic, phenotypic, biomarker, or psychosocial variables that distinguish a given patient from other patients with similar clinical presentations.(26)

In regards to medical care, the application of personalized medicine could minimize harmful side effects, create a more successful result, and potentially be more cost effective by reducing the use of less successful and less direct treatment pathways. The application of this to pharmacology has the ultimate goal to provide 'the right drug, with the right dose at the right time to the right patient'.(27) While genomics is only a portion of the realm of personalized medicine, translation of genetic risk profiles to clinical medicine have the potential to elevate benefits and reduce risks to patients by targeting both prevention and treatment more effectively.

Genetic risk information can prevent disease only if it improves the use of behavioral or medical interventions. The strongest case can be made when a unique intervention is needed for individuals with a particular genotype. For example, testing for *BRCA* risk variants represents a case in which individuals with increased risk for breast cancer may be both screened and treated differently than the rest of the population. Magnetic resonance imaging studies for screening as well as prophylactic mastectomies or oophorectomies for prevention are pursued in many of these patients. However, many screening protocols are provided for individuals regardless of genetic risk based upon known clinical risk factors, for example screening for cervical or prostate cancer. Many disease prevention strategies involve recommendations for healthy lifestyles including no smoking, healthy diets, and exercise, which are best applied to the entire population.(28) Further, current research suggests that genetic risk counseling does not significantly alter self-reported motivation or prevention program adherence when applied to overweight individuals at risk for diabetes.(29) However, there is evidence that genomic risk profiling can increase physician follow-up and does not result in adverse changes in psychological health or follow-up related distress for the patients.(30) Thus, the optimism is that the knowledge of genetic risk for a disease would lead to changes in recommendations for prevention or treatment provided by physicians and choices made by patients.

When these concepts are applied to obesity, one can imagine the opportunities for identifying individuals at risk for obesity based upon risk factors early in life. While some of those risk factors may include environmental exposures (e.g., socioeconomics, exercise, diet), this research confirms the strong genetic risk component to increased BMI. We demonstrate also that it is not a few SNPs that contribute solely to that risk, but rather summation of genetic risk across the genome elucidates further aspects of this complex disease. The use of genome-wide risk profiling could be applied to identification of individuals who would benefit from environmental modifications or heightened medical awareness prior to the onset of obesity and its significant associated morbidities.

A potential opportunity for change in treatment strategies of individuals with genetic risk for obesity is undergoing more aggressive weight loss interventions (i.e. bariatric surgery) earlier in life than otherwise would have been recommended and prior to the development of comorbidities. In the absence of severe class 3 obesity, comorbidities such as type 2 diabetes or severe sleep apnea are typically required for insurance approval for weight loss surgery in individuals with less extreme classes of obesity. Prior research has shown that individuals who have undergone bariatric surgery have a higher genome-wide PRS.(31) A combined analysis of 922 bariatric surgery participants in the UK Biobank and Partners HealthCare System found a high genome-wide PRS in 319 (34.6%) of the patients. Compared with the remainder of the individuals analyzed, a high genome-wide PRS was associated with a 5.0-fold increased risk of severe obesity treated with bariatric surgery. We similarly identified a significantly higher genome-wide PRS in individuals who had undergone bariatric surgery ($p < 2.2x10^{-16}$). It certainly would be plausible that individuals with obesity and a high PRS, in the absence of comorbidities, may benefit from consideration for bariatric surgery prior to the progression of the many comorbidities that develop over a lifetime with obesity. Further studies may even be able to elucidate if certain individuals, at a given BMI, are at greater genomic risk for obesity-related comorbidities.

Conclusion

The body of research presented in this dissertation has made significant contributions to the multidisciplinary field of biomedical informatics, including its component disciplines of genomics, phenotyping, and clinical research informatics, along with important discoveries in clinical medicine that could impact the care provided to patients. This dissertation describes new approaches to defining the clinical impact of common and rare conditions using the electronic health record, phenotyping methods, and genomic associations. The research here advanced phenomic and genomic informatics methods to characterize a common disease, obesity, which has significant previously uncharacterized associations with a broad range of diseases. This analysis is the first application of genome-wide risk scores for obesity in a phenome-wide approach. The novel methodologies developed with the use of genetic risk scores to analyze all risk genes, including those with small and large effect size in a population, have greater accuracy and improved ability to find associations. This body of work has established the framework methodology and validated the use of PheWAS techniques across multiple cohorts and using both clinical and genomic predictors. I am also providing here the first evidence that genome-wide polygenic risk scores show strong concordance with observational effect sizes in phenome-wide association studies. The methods described here demonstrate improved ability for PheWAS techniques to identify novel associations by increasing power and improving predictive capacity. Investigations using methods such as these will help provide the linkages between disease-gene associations, cellular mechanisms, and therapeutic approaches, making critical advances to informatics and treatment of patients.

And lastly, this dissertation also has significant impact to clinical medicine, as it defines the phenotype of a poorly characterized rare disease and also highlights the burden of disease in society attributable to the very common phenotype, obesity. This in itself has significant impacts on clinical medicine, but also further applications of the methodology in this dissertation could make additional discoveries regarding other phenotypes and improve the understanding of polygenic risk for disease. The overall translation of genetic risk profiles to clinical medicine has the potential to increase benefits and reduce risks to patients by targeting both prevention and treatment more effectively. These methods and opportunities do not apply only to obesity, but can impact the way we as informaticians and clinicians think about and treat all polygenic diseases.

Some have advocated for the tempering of the claims for precision or personalized medicine to improve overall medical care due to the lack of generalizability to the majority of disease that a personalized approach brings.(32) In contrast, I advocate that we continue to think outside of the traditional scopes of medical practice to identify opportunities to use existing data within the healthcare record to identify disease patterns, promote research, and translate findings into clinical care. A crucial component to this advancement will be the comprehensible incorporation and visualization of genomic and pharmacogenomic information inside and outside of the EHR with the development of clinical decision support to guide clinicians on treatment strategies and drug dosing. As further information is gained about the polygenic risk for complex diseases, its inclusion in the EHR and medical care will need to be constructed in a pointed manner that also provides for evaluation of its effect on both patients and providers. With the continued accrual of new data comes possibilities for translation with new technologies to improve medical care, and this confluence of research with clinical care will continue to drive evolution of the field of biomedical informatics.

References

1. Denny JC, Arndt FV, Dupont WD, Neilson EG. 2008. Increased hospital mortality in patients with bedside hippus. *Am. J. Med.* 121(3):239–45

2. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* 26(9):1205–10

3. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. 2013. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 14(3):187–91

4. Hebbring SJ, Rastegar-Mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. 2015. Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinforma. Oxf. Engl.* 31(12):1981–87

5. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. 2016. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc. JAMIA.* 23(e1):e20-27

6. Cohen S, Jannot A-S, Iserin L, Bonnet D, Burgun A, Escudié J-B. 2019. Accuracy of claim data in the identification and classification of adults with congenital heart diseases in electronic medical records. *Arch. Cardiovasc. Dis.* 112(1):31–43

7. Ramanathan R, Leavell P, Stockslager G, Mays C, Harvey D, Duane TM. 2014. Validity of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) screening for sepsis in surgical mortalities. *Surg. Infect.* 15(5):513–16

8. Fan J, Arruda-Olson AM, Leibson CL, Smith C, Liu G, et al. 2013. Billing code algorithms to identify cases of peripheral artery disease from administrative data. *J. Am. Med. Inform. Assoc. JAMIA.* 20(e2):e349-354

9. Samuel AM, Lukasiewicz AM, Webb ML, Bohl DD, Basques BA, et al. 2015. ICD-9 diagnosis codes have poor sensitivity for identification of preexisting comorbidities in traumatic fracture patients: A study of the National Trauma Data Bank. *J. Trauma Acute Care Surg.* 79(4):622–30

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

11. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, et al. 2010. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res.* 62(8):1120–27

12. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, et al. 2012. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J. Am. Med. Inform. Assoc. JAMIA*. 19(2):219–24

13. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. 2013. Defining and measuring completeness of electronic health records for secondary use. *J. Biomed. Inform.* 46(5):830–36

14. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31(12):1102–10

15. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS One*. 12(7):e0175508

16. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50(9):1219–24

17. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, et al. 2018. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* 102(6):1048–61

18. Fritsche LG, Beesley LJ, VandeHaar P, Peng RB, Salvatore M, et al. 2019. Exploring various polygenic risk scores for skin cancer in the phenomes of the Michigan genomics initiative and the UK Biobank with a visual catalog: PRSWeb. *PLoS Genet.* 15(6):e1008202

19. Polubriaginof FCG, Vanguri R, Quinnies K, Belbin GM, Yahi A, et al. 2018. Disease Heritability Inferred from Familial Relationships Reported in Medical Records. *Cell*. 173(7):1692-1704.e11

20. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, et al. 2019. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* 51(1):51–62

21. Smith GD, Ebrahim S. 2003. “Mendelian randomization”: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32(1):1–22

22. Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. 2006. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am. J. Epidemiol.* 163(5):397–403

23. Emdin CA, Khera AV, Natarajan P, Klarin D, Zekavat SM, et al. 2017. Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA*. 317(6):626–34

24. Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, et al. 2017. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. *JAMA Cardiol.* 2(8):882–89

25. Lindström S, Germain M, Crous-Bou M, Smith EN, Morange P-E, et al. 2017. Assessing the causal relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum. Genet.* 136(7):897–902

26. Jameson JL, Longo DL. 2015. Precision medicine--personalized, problematic, and promising. *N. Engl. J. Med.* 372(23):2229–34

27. Sadée W, Dai Z. 2005. Pharmacogenetics/genomics and personalized medicine. *Hum. Mol. Genet.* 14 Spec No. 2:R207-214

28. Burke W, Psaty BM. 2007. Personalized medicine in the era of genomics. *JAMA*. 298(14):1682–84

29. Grant RW, O'Brien KE, Waxler JL, Vassy JL, Delahanty LM, et al. 2013. Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. *Diabetes Care*. 36(1):13–19

30. Bloss CS, Schork NJ, Topol EJ. 2014. Direct-to-consumer pharmacogenomic testing is associated with increased physician utilisation. *J. Med. Genet.* 51(2):83–89

31. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, et al. 2019. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*. 177(3):587-596.e9

32. Joyner MJ, Paneth N. 2015. Seven Questions for Personalized Medicine. *JAMA*. 314(10):999–1000