Bridging the Granularity Gap in Family History Information Extracted from Clinical Narratives

Sungrim Moon PhD[1], Liwei Wang MD, PhD[1], Xuan Chen BS[2], Nan Wang BS[3], Sheila M Manemann MPH[3], Nicholas B Larson PhD[3], Suzette J Bielinski PhD[3], Hongfang Liu PhD[1]
[1]Department of Artificial Intelligence and Informatics; [2]Center for Clinical and Translational Science; [3]Department of Quantitative Health Sciences
Mayo Clinic, Rochester, Minnesota, USA

Abstract

Family history (FH) is important for disease risk assessment and prevention. However, incorporating FH information derived from electronic health records (EHRs) for downstream analytics is challenging due to the lack of standardization. We aimed to automatically align FH concepts derived from a clinical corpus to disease category resources popularly used, including Clinical Classification System (CCS), Phecode, Comparative Toxicogenomics Database (CTD), Human phenotype ontology, and Human disease ontology (HDO). Leveraging the Unified Medical Language System (UMLS), we achieved high mapping coverages of FH concepts in those resources, using the parent and broader/alike relations available in the UMLS. Among the five resources, CTD has the best coverage (93%) of FH concepts, HDO has the coarsest granularity of FH disease categories, while CCS showed the finest-grained regarding disease categories. The study suggests that we can mitigate the challenge of various degrees of granularity of NLP-derived FH using those ontology or terminological resources.

Introduction

A comprehensive understanding of diseases in close family members can be beneficial for risk estimation, individualized screening plans, and prevention of common diseases of patients such as cardiovascular disease (CVD)[1-5]. Since people from the same family share biological and genetic susceptibility, as well as environmental and behavioral risk factors, family history (FH) information provides a reasonable degree of interpretable clues on the health status of a patient or the need for further clinical tests and referral in the clinical settings[2, 6, 7]. Previous research has identified positive family history as being predictive and actionable in CVD risk calculation[8, 9]. The detailed information of FH is often available in the form of narrative documents in electronic health records (EHRs). Natural language processing (NLP) approaches are powerful to extract FH information from clinical narratives; however, the systematic collection, analysis, and utilization of this rich FH information have been underutilized because of the lack of standardization and various degree of concept granularity.

To interpret FH information correctly, categorizing diseases of FH information into closely relevant disease groups is necessary, on top of identifying family members[6]. In clinical documents, FH information is represented as terms to describe diseases of the patient's family members, which is mainly documented in the FH section. These FH terms need to be converted into clinical concepts, then to process additional alignment to corresponding disease categories utilizing existing resources containing common human disease categories, where curated code sets based on the International Classification of Diseases (ICD) have been frequently used for EHR-based studies[10]. However, the purpose of ICD is for billing/reimbursement, not for capturing FH information. We investigated the mapping of NLP-derived FH concepts from a large corpus to disease category resources defined in commonly used disease ontology or terminological resources leveraging the unified medical language system (UMLS). The mapping of the FH concepts was assessed by the mapping coverage (i.e., the ability to identify the corresponding human disease categories). We also compared the density of FH concepts in those existing resources (i.e., the granularity variation in disease categories) which may fill the granularity gap of NLP-derived FH concepts. The datasets deriving from this study are available for public use at https://github.com/OHNLP/FHOntology.

Background

Resources associated with human disease categorizations

A few standards address how common human disease information should be interpreted and categorized. These state-of-the-art biomedical ontologies or terminological resources contain disease categories and their hierarchical relations

with a multi-level granularity. We choose the UMLS and five existing resources (disease codes systems or disease ontologies) as follows; Clinical Classifications Software (CCS)[1], PheWas[2], Comparative Toxicogenomics Database (CTD)[3], Human phenotype Ontology (HPO)[4], and Human Disease Ontology (DO)[5].

The UMLS Metathesaurus, initiated and governed by the US National Library of Medicine, has integrated information from the diverse terminological sources and linked all identifiers in multiple resources into an appropriate CUI (concept) as a Rich Release Format (RRF), MRCONSO.RFF. Another file, MRREL.RFF contains all pair-wise relationships defined in original terminological sources.

CCS Software was initially developed and maintained by the Healthcare Cost and Utilization Project, sponsored by the Agency for Healthcare Research and Quality (AHRQ), to analyze costs, usage, and outcome associated with patient diagnoses and procedures. CCS was reorganized by aggregating the disparate ICD, 9th Revision, Clinical Modification (ICD-9-CM) codes into clinically meaningful categories to facilitate clinical research[11]. This code system was expanded to ICD-10-CM/PCS code in the form of Clinical Classification Software Refine (CCSR). CCS code system has 285 mutually-exclusive diagnosis categories with associated ICD diagnosis codes.

Phecode aggregated related ICD billing codes customarily into distinct diseases or traits to replicate genetic associations/discovery (i.e., Phenome-wide association studies, PheWAS). Phecode represents clinical phenotypes, which refers to the observable physical properties of an organism, using clinical records for enabling clinical and genetic research[12-14]. Phecodes X (extended version) covered both ICD-9-CM and ICD-10-CM codes[15]. Phecode has the top 22 clinical categories to encode common diseases.

CTD was built on understanding how environmental exposures (chemical, gene, disease, pathway, and exposures) affect human health. It has manually curated information utilizing Medical Subject Headings (MeSH) or Online Mendelian Inheritance in Man (OMIM) based on literature[16]. CTD divides diseases into the top 36 categories.

HPO is a standardized ontology of phenotypic abnormalities in human diseases. It has been used for clinical diagnostics in bioinformatics research, primarily focused on the relationships between human phenotypic abnormalities and human disease[17]. The ontology is based on laboratory or textual EHR data[18]. The 18 different resources (including CUIs in the UMLS) are cross-referenced into HPO ID. HPO ID itself also is available in the UMLS. All HPO categories merge into the "Phenotypic abnormality."

HDO is a standardized representation for common and complex human diseases to provide human disease terms and phenotype characteristics for the biomedical community[19]. An identifier, DO ID, harmonized disease representation in the 13 different resources (including CUIs in the UMLS) through database cross-references to understand the evolution and changes of human diseases. All HDO categories belonged to the "disease" category.

Relevant studies

Similar to this study, Wei et al. demonstrated the comparison among the codes, CCS for ICD-9, ICD-9-CM, and phecode, and presented phecodes has better alignments phenotypes in clinical practices[12]. They found ICD-9-CM were typically too detailed (requiring custom groupings), while CCS were often not granular enough. Also, phecode generally produced more substantial odds ratios and lower p-values for known associations than ICD-9-CM and CCS. However, these studies restricted their mappings to human disease-relevant phenotypes rather than FH concepts.

Method

Overview

Figure 1 presents the overview of this study. We extracted clinical concepts (i.e., CUIs) associated with human diseases from the "Family History" section in clinical notes (CNs) of Mayo EHR utilizing two NLP pipelines (MetaMap and MedTagger) and the UMLS (v2021AA) Meta Rich Release Format (RRF). Three trained experts

[1] https://www.hcup-us.ahrq.gov/tools_software.jsp
[2] https://phewascatalog.org/phecode_x
[3] http://ctdbase.org/downloads/#alldiseases
[4] https://hpo.jax.org/app/download/ontology
[5] https://github.com/DiseaseOntology/HumanDiseaseOntology/tree/main/src/ontology

manually verified the validity of these extracted CUIs given the most frequent clinical sentences to ensure the concepts were extracted correctly. The validated CUIs were then mapped to the five resources via CUIs in the UMLS. We used two mapping processes: (1) directly map to the individual resource via available codes (including CUIs) of resources to identify disease categories (direct mapping); (2) identify extended CUIs sets (the parent or broader or alike relational CUIs) of a given CUI using the MRREL.RFF and then map to those resources (indirect mapping) After mapping, we investigated the coverage of FH CUIs and compared the level of granularity of disease categories of five resources.

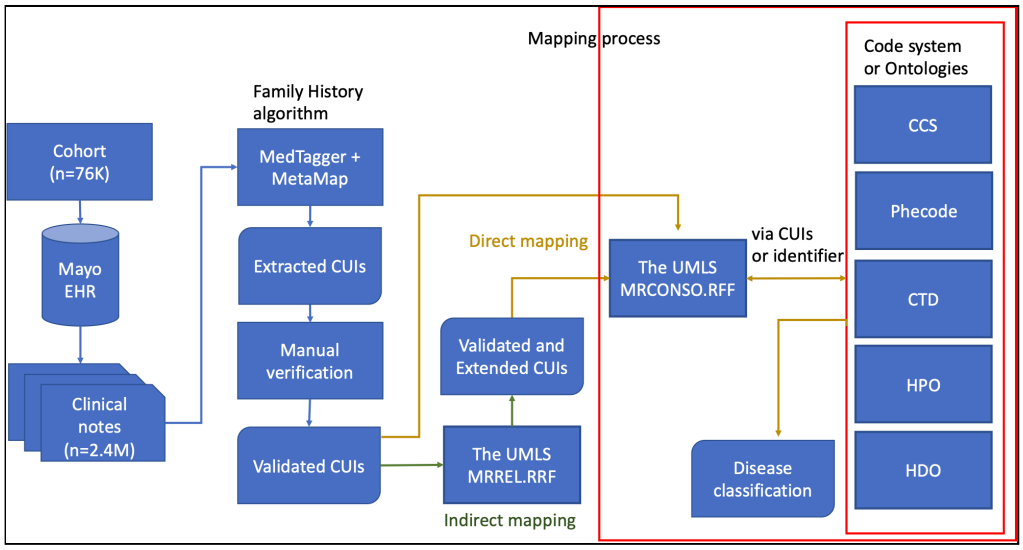

Figure 1. Diagram outlining the overview of this study.

Materials

We utilized a previously ascertained cohort of 76,255 individuals who were 30 years of age or older and resided in Olmsted County, Minnesota, on January 1, 2006[20]. For this effort, we restricted to those who were Mayo Clinic patients. All 2,413,794 clinical documents in Mayo EHRs were generated during the baseline data collection period of January 1, 2001, through December 31, 2005. Sentences potentially associated with the family member of patients were extracted from the "Family History" sections in clinical documents.

Natural Language Processing

Figure 2 describes the detailed NLP-FH algorithm for this study. We use both MedTagger and MetaMap pipelines to generate the concepts from the clinical narratives. MedTagger was utilized to detect sentences of clinical documents. The detected sentences went through tokenization, concept identification, the assertion in both NLP pipelines separately. Heuristics rules and a set of keywords identifying family members were applied to the sentences to extract the concise and unambiguous FH concepts, then encode FH concepts with associated CUIs. The scope of the family members ranges from first-degree relatives (Parents, Siblings, and Children) to second-degree relatives (Aunts, Uncles, Grandparents, and Nieces/Nephews). To collect disease concepts, we constrain concepts that belong to the set of semantic types of the UMLS, such as "Disease or Syndrome," "Neoplastic process," "Mental or Behavior Dysfunction," "Injury or poisoning," and "Sign or Symptom." We directly collected the results of the pipeline as CUIs for MetaMap. For MedTagger, we gathered the normalized terms from clinical narratives then mapped them to the UMLS (MRCONSO.RFF) to find associated CUIs. As a result, we assembled CUIs associated with the given FH sentences from two pipelines. For example, C0027051 (Myocardial Infarction) has a set of N sentences such as (sentence1::'Father, Heart attack', sentence2::'Father, Heart attack', sentence3::'Mom died at age 89 of a MI', sentence4::'Paternal cousin has died at the age of 35 with myocardial infarction', …… , sentenceN::'Brother had a stroke at age 71.')

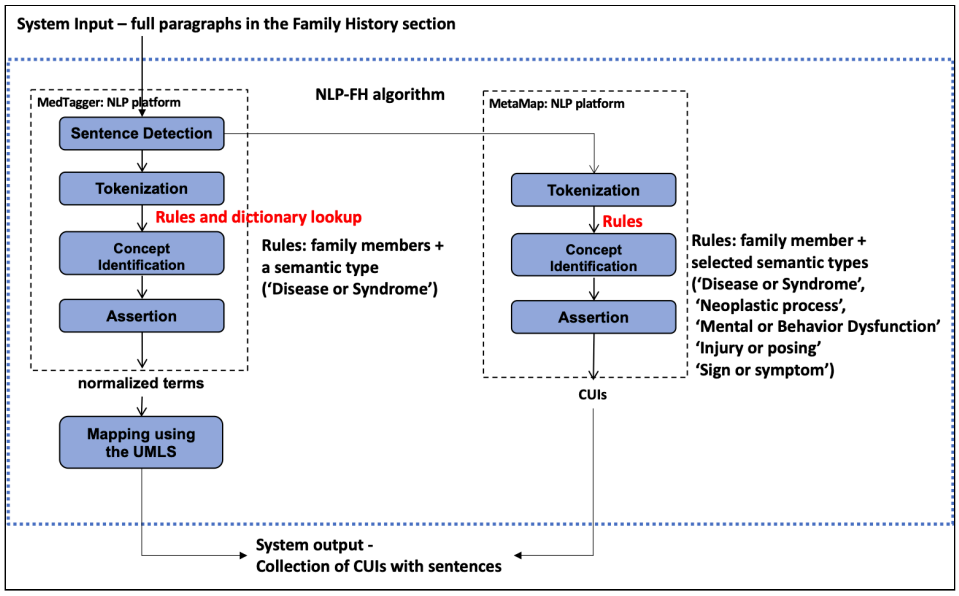

Figure 2. Diagram outlining the natural language processing algorithm for family history.

The validity for CUIs

To collect the correct concepts which are clinically relevant to the FH, we evaluated the validity of the CUIs by manually reviewing the most frequent associated sentences in the entire sentence set per CUI. Two trained annotators (CX and NW) reviewed separately to determine whether the CUI corresponds to the context of given sentences. Annotators referred to the definitions and synonyms of CUIs through UMLS Metathesaurus[6]. The third independent annotator (LW) adjudicated the conflicts when a disagreement arose between annotators. We kept CUIs that correctly correspond to the given sentence as the valid CUIs sets for the subsequent mapping processes. Note that we do not validate false-negative cases of NLP-FH algorithms in this study. The two annotators' inter-annotator agreement (IAA) was calculated using overall percentage agreement and Cohen's Kappa statistic.

Mapping disease code systems and ontologies via CUIs

Table 1. The identifier and vocabulary in the UMLS in this study.

The identifier	The vocabulary of interest in the UMLS	Mapped resources
UMLS CUI	All vocabularies for HPO and HDO. Clinical Classifications Software (CCS) and Clinical Classifications Software Refined for ICD-10-CM (CCSR_ICD10CM) for CCS	CCS, HPO, HDO
ICD diagnosis codes	ICD-9-CM Entry Terms (MTHICD9), International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM), International Classification of Diseases and Related Health Problems, Tenth Revision (ICD10), International Classification of Diseases, Tenth Revision, Clinical Modification (ICD10CM)	CCS, Phecode, HPO, HDO
SNOMED-CT	SNOMED CT, US Edition (SNOMEDCT_US)	HPO, HDO
Other identifiers	Human Phenotype Ontology (HPO), Online Mendelian Inheritance in Man (OMIM), Medical Subject Headings (MSH), SNOMED CT, US Edition (SNOMEDCT_US), NCI Thesaurus (NCI), KEGG Pathway Database (NCI_KEGG), Medical Dictionary for Regulatory Activities Terminology (MDR)	CTD, HPO, HDO
All codes	Above-mentioned vocabularies of interest per identifier	

[6] https://uts.nlm.nih.gov/uts/umls/home

We considered all vocabulary identifiers in the UMLS to be identical if these identifiers shared the same CUI in MRCONSO.RFF. And we assumed the resources for disease categorization are correct without manual validation. As a preprocessing, we filtered non-English CUIs in MRCONSO.RFF and collected existing identifiers (i.e., CCS has ICD diagnosis codes, one of the identifiers of HPO is SNOMED-CT codes, and one of the identifiers of CTD is Medical Subject Headings) per resource. Each validated CUI mapped to the filtered MRCONSO.RFF and a resource containing disease categories via identifiers if the resource shared identifier information (or CUI) with the interesting vocabulary in the UMLS (Table 1).

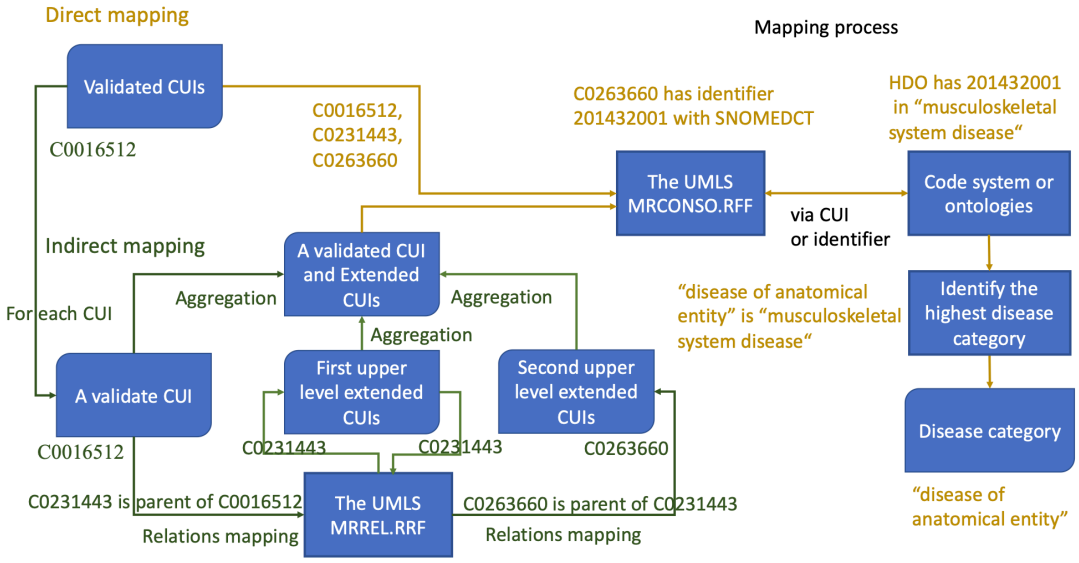

Figure 3. Direct and indirect mapping with an example of CUI C0016512.

Direct mapping

The UMLS MRCONSO.RFF consists of CUIs and identifiers in diverse vocabularies, and each resource containing disease categories have CUIs or identifiers in subset vocabularies of the UMLS. CCS, HPO, and HDO include CUIs as cross-references of the UMLS, which mapped to the disease category. If an identifier in vocabulary belongs to both the UMLS and resource ("via CUI or identifier" in Figure 3), the CUI can be mappable to the disease classification of the resource. ("Disease category" in Figure 3). In Figure 3, both the UMLS and HDO has identifier '201432001' in SNOMED-CT vocabulary for a C0263660, which maps to the "musculoskeletal system disease" category of HDO.

Different additional task for each resource is necessary to get the highest level of disease categories in its resource ("Identify the highest disease category" in Figure 3). The top disease categories of CCS are in "CCS" vocabulary, with the term type as a single-level diagnosis in the UMLS. Suppose a CUI map to identify with a multi-level diagnosis of "CCS" or "CCSR" vocabularies in the UMLS; this CUI converts into a single-level disease category by looking up the code comparison list verified by the AHRQ. In the case of Phecode and CTD, there is no additional process.

If CUI or identifier is mapped for the HPO and HDO, it has the corresponding disease category. In Figure 3, C0263660 maps to "musculoskeletal system disease." We recursively track its parent disease category to reach up to the "Phenotypic abnormality" category for HPO and "disease" category for HDO, and then took the last-child disease category (or categories) as the top disease category ("Identify the highest disease category" in Figure 3). The "musculoskeletal system disease" has a parent "disease of anatomical entity," and the "disease of anatomical entity" has a parent "disease" in order. Therefore, C0263669 maps the "disease of anatomical entity" category ("Disease category" in Figure 3).

Indirect mapping

For each CUI, we generated the extended CUI set, taking into account the parent relational, broad relational, or narrow relational CUIs of the original CUI utilizing MRREL.RFF. We investigated up to two level-up relationships to

generate extended sets in this study. For example, C0016512 (Foot pain) has no mapped category for any resource through direct mapping processes. We generated the first level extended CUI set containing parent and broad/narrow relationships of C0016512 by mapping to MRREL.RFF ("Relations mapping" in Figure 3). C0016512 has several parent CUIs, and one of them is C0231443 (Musculoskeletal symptom), which is "First upper-level extended CUIs" in Figure 3. Collected the first upper-level extended CUIs remapped to MRREL.RFF to get their upper-level relational CUIs, which are the 2nd level of relationship with C0016512 ("Second upper-level extended CUIs" in Figure 3). We aggregated these extended CUI sets and original CUI ("Aggregation" in Figure 3), then went through the same process of direct mapping per extended CUI. One of the parents of C0231443 (Musculoskeletal symptom) is C0263660 (Musculoskeletal and connective tissue disorders), mappable to HDO via '201432001' in SNOMED-CT. As a result, the original C0016512 to map the "disease of anatomical entity" category of HDO.

Evaluation

We are interested in how many NLP-derived FH concepts can be automatically mapped to diseases categories per resource. We calculated the mapping coverage of FH concepts per resource as follow:

$$Coverage = \frac{The\ count\ of\ (validated\ CUIs \cap mapped\ CUIs\ to\ any\ correspond\ diesese\ category\ per\ resource)}{total\ number\ of\ validated\ CUIs} * 100$$

If one CUI is mappable to all five resources, we term it as the commonly mapped CUI. For the commonly mapped CUIs, we compared the distribution of mapped disease categories per resource to understand the various degree of granularity of disease categories within a resource or among resources. We calculated the density of individual disease category per resource as follow:

$$Density = \frac{The\ count\ of\ (commonly\ mapped\ CUIs \cap mapped\ CUIs\ to\ the\ particular\ diesese\ category\ in\ the\ particular\ resource)}{total\ number\ of\ commonly\ mapped\ CUIs\ using\ all\ resources} * 100$$

Results

After filtering the non-family history sections, 213,345 CNs (9% of total CNs) of 41,328 patients (54% of the total patients) contain family history information in this cohort. We extracted 148,048 unique sentences. In total, 4,043 CUIs were extracted by the NLP-FH algorithm. Among them, 3,590 CUIs (89%, 3590/4043) were determined as valid FH concepts. The percentage agreement was 0.88 (3564/4043), and Cohen's kappa score was 0.36 (Fair agreement) between two annotators.

Table 2. The coverages of FH concepts

	The identifier	Clinical Classification Software (CCS)	Phecode	Comparative Toxicogenomics Database (CTD)	Human Phenotype Ontology (HPO)	Disease ontology (HDO)
Direct mapping	UMLS CUI	5%[a]	-	-	26%	32%
	ICD-9-CM ICD-10-CM	41%	28%	-	0%	25%
	SNOMED-CT	-	-	-	27%	34%
	Other codes[b]	-	-	50%	27%	40%
	All codes[c]	42%	28%	50%	34%	46%
Indirect mapping	Extend CUIs	90%	90%	93%	87%	91%

[a] CCS code conversion from multiple to single diagnosis code is included. [b]Other codes are OMIM and MeSH for CTD, MeSH, and MedDRA for HPO, OMIM, MeSH, NCI, KEGG, and MedDRA for HDO, respectively. [c]All codes combined UMLS CUI, ICD 9/10 CM, SNOMED-CT, and other codes.

Table 2 represents the coverages of FH concepts via CUIs and identifiers. Using direct mapping via all codes, 522 CUIs exclusively map to one resource in CCS, Phecode, CTD, HPO, and HDO. In contrast, 1,992 (subtract 522 CUIs from 2,514 all mapped CUIs) maps to more than one resource. For example, C0427055 (Facial Paresis) maps to two disease categories in each CCS and Phecode, three in each HPO and CTD.

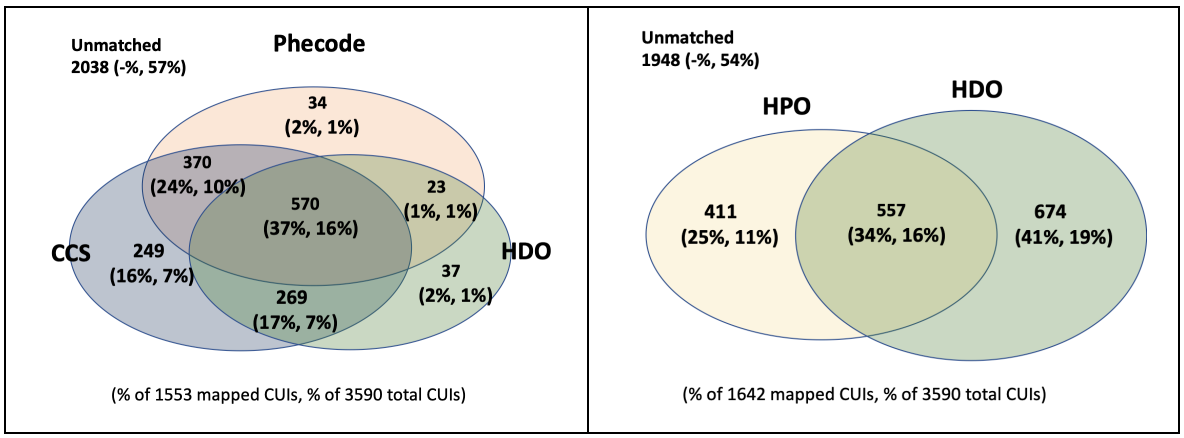

Figure 4. ICD and SNOMED-CT code mapping distribution in the form of Venn diagrams. Left) Direct mapping via ICD codes to CCS, Phecode and HDO. HPO was not presented in this figure because one mapped CUI via ICD codes. Right) Direct mapping via SNOMED-CT codes to HPO and HDO; CCS=Clinical Classification Software; HPO=Human Phenotype Ontology; HDO=Human Disease ontology

Figure 4 shows ICD and SNOMED-CT code mapping distribution in the form of Venn diagrams. A set of 1,552 CUIs via ICD diagnosis codes map in at least one of CCS, Phecode, HPO, or HDO resources (Figure 4, Left, HPO was excluded in Figure), while 1,642 CUIs via the SNOMED-CT map to HPO or HDO (Figure 4, Right). For mapping processing via ICD code, 1,458, 997, 1, and 899 CUIs map in CCS, Phecodes, HPO, and HDO, respectively, which are about 25-41% coverages of the total 3,590 CUIs. CCS has the highest coverage (41%) with respect to the mapping via ICD codes. Meanwhile, 968 CUIs for HPO and 1,231 CUIs for HDO were mapped via SNOMED-CT code. Overall, the SNOMED-CT codes have slightly higher coverage than ICD diagnosis codes.

Table 3. The distribution of CUIs using indirect mapping according to the frequent high order.

Rank	Mapped code systems or Ontologies	Freq	% of 3370 CUIs	% of 3590 CUIs	Rank	Mapped code systems or Ontologies	Freq	% of 3370 CUIs	% of 3590 CUIs
1	CCS, Phecode, CTD, HPO, HDO	3037	90	1	15	CCS, Phecode, CTD	8	0	0
2	Unmatched	220	7	0	16	CCS, Phecode, HPO	7	0	0
3	CCS, Phecode, CTD, HDO	86	3	0	17	Phecode	6	0	0
4	CTD	44	1	0	18	Phecode, CTD, HDO	6	0	0
5	CCS, CTD	40	1	0	19	CCS, Phecode, HDO	4	0	0
6	CCS, Phecode, CTD, HPO	30	1	0	20	CTD, HPO	3	0	0
7	CTD, HDO	18	1	0	21	Phecode, CTD	2	0	0
8	Phecode, CTD, HPO, HDO	17	1	0	22	Phecode, HDO	2	0	0
9	CTD, HPO, HDO	13	0	0	23	CCS, CTD, HPO	1	0	0
10	CCS, CTD, HDO	9	0	0	24	CCS, Phecode, HPO, HDO	1	0	0
11	HPO	9	0	0	25	HDO	1	0	0
12	CCS	8	0	0	26	HPO, HDO	1	0	0
13	CCS, CTD, HPO, HDO	8	0	0	27	Phecode, CTD, HPO	1	0	0
14	CCS, Phecode	8	0	0	28	Sets of (CCS, HDO), (CCS, HPO), (CCS, HPO, HDO), (Phewas, HPO), (Phewas, HPO, HDO)	0	0	0

CCS=Clinical Classification Software; CTD=Comparative Toxicogenomics Database; HPO=Human Phenotype Ontology; HDO=Human Disease ontology

In total, 94% (3,370 out of 3,590) of FH concepts were mapped to at least one disease category of five resources using indirect mapping. 3,037 out of 3,590 CUIs (85%) were the commonly mapped CUIs, meaning a CUI map to a disease category of all five resources. Table 3 presented the distributions of CUIs to map which set of resources using indirect mapping according to the frequent high order. Major CUIs are the commonly mapped CUIs (Table 3, Rank 1), and the following highest frequent CUI set is 220 CUIs that unmapped to any disease category (Table 3, Rank 2). Due to the increased coverage of all resources compared to direct mapping, each resource does not contribute exclusively to map disease categories for CUIs. Only 44, 9, 8, 6, and 1 CUIs were mapped to solely CTD, HPO, CCS, Phecode, and HDO, respectively.

Table 4. The top 10 high-density categories per resource for the common 3,037 CUIs using indirect mapping.

Rank	Clinical Classification Software (CCS)		Phecode		Comparative Toxicogenomics Database (CTD)		Human Phenotype Ontology (HPO)		Disease ontology (HDO)	
	Category	%	Category	%	Category	%	Category	%	Category	%
1	Other nervous system disorders	6	Neoplasms	25	Cancer	25	Abnormality of the nervous system	21	disease of anatomical entity	59
2	Malignant neoplasm without specification of site	5	Neuro	11	Nervous system disease	17	Neoplasm	13	disease of cellular proliferation	25
3	Other congenital anomalies	4	Cardio	9	Cardiovascular disease	12	Abnormality of the digestive system	12	disease of mental health	10
4	Other and unspecified benign neoplasm	3	GI	9	Digestive system disease	10	Abnormality of the cardiovascular system	11	disease of metabolism	3
5	Cancer; other and unspecified primary	2	Musc/Skel	7	Mental disorder	9	Abnormality of the genitourinary system	9	disease by infectious agent	3
6	Other gastrointestinal disorders	2	Cong	7	Signs and symptoms	8	Abnormality of the musculoskeletal system	7	syndrome	3
7	Viral infection	2	Mental	6	Musculoskeletal disease	8	Abnormality of blood and blood-forming tissues	7	physical disorder	1
8	Other nutritional; endocrine; and metabolic disorders	2	Genitourinary	5	Urogenital disease (female)	7	Abnormality of the immune system	7	genetic disease	0
9	Neoplasms of unspecified nature or uncertain behavior	2	Blood	5	Skin disease	7	Abnormality of the integument	6		
10	Other lower respiratory disease	2	Resp	4	Genetic disease (inborn)	6	Abnormality of the respiratory system	6		
Total	249 categories		22 categories		38 categories		22 categories		8 categories	

*Note one concept may map several categories per resource. e.g., the sum of % of HDO is 104%

For the commonly mapped 3,037 CUIs, Table 4 lists the 10 most high-density categories per resource. HDO has 59% of the commonly mapped CUIs into one category, “dieasese of anatomical entity” as a top density. However, the top density of disease category in CCS is “Other nervous system disorders,” with 5% of the same commonly mapped CUIs. It is because HDO contains the lowest numbers (n=8) of top disease categories, while CCS has the highest number (n=249) in Table 4. Because each resource has different numbers of top categories, the density and granularity of mapped FH-relevant human diseases are significantly different per resource.

Discussion

Disease category resources used in this study vary in granularity and structure. For example, CCS has 26 different categories relevant to "Cancer" (e.g., Cancer of stomach, Cancer of colon), indicating CCS codes have fine-grained granularity, which is aligned with results from a previous study[14]. In contrast, "Disease of cellular proliferation" has a child term, "Cancer" for HDO, indicating HDO may contain coarse granular disease categories. As a result, the category "disease of anatomical entity" of HDO has the highest density (59%) in our study. Each existing resource has pros and cons depending on the purpose of usability of disease categories. CCS, Phecode, and CTD have a cohesive set of code variations to map their own categories. Meanwhile, HPO and HDO have associated child/parent disease categories as a hierarchical structure with depth information. Therefore, utilizing multiple disease category resources may be beneficial to obtain comprehensive clinical concepts and enrich the FH information since a single resource may not have sufficient granularity and quality, which can hinder their applicability in real-world applications[21, 22].

ICD codes have widely served as a reference to categorize human diagnoses for clinical practice or research for decades. Despite the long historical and frequent usages, utilizing ICD diagnosis codes is still challenging to represent FH disease fully. In this study, CCS has a large capacity to cover the number of clinically equivalent ICD codes; however, too fine granularity of the disease code system without meaningful organization for human diseases results in a low capability to represent FH knowledge. The overall low coverage (25-41%) using ICD codes in this study reflects that ICD codes were insufficient to encode clinical FH concepts or semantic expressions for diseases in clinical documents. Note that ICD codes were developed for billing rather than to reflect meaningful diseases.

The main challenge underlying automated FH disease categorization is that the FH concepts do not capture the approximate terms and phrases used to refer to the patients' diseases in clinical documents and existing disease categorizations. For example, physicians recorded the diagnosis of patients in the "Problem list" section with detailed descriptions (e.g., Atrial fibrillation) and standardized codes, often following the ICD code, in EHR. But they recorded relevant FH information within a few lines using comparably too general terminologies (e.g., Heart diseases). The low coverage (about 50%) FH concepts via direct code mapping in this study refer to the different granularity of FH concepts compared to patients' diagnoses. We used the parent or broader or alike relational CUIs associated with the FH concept in indirect mapping, significantly improving the coverage (43-62% per resource). We plan to derive an FH resource based on the work conducted here to facilitate the downstream applicability of NLP-derived FH concepts.

Conclusion

FH information in clinical narratives has high utility for clinical practice and research, particularly for diagnosing disorders and preventing conditions. However, it has been underutilized due to the limited standardization. We demonstrated an approach to improve the utility of NLP-derived FH information through aligning those concepts to resources containing disease category information with approximately 94% coverage in mapping disease categories. Although there are no single best or preferred resources for disease categorization, we showed the degree of granularity varied among those resources.

Acknowledgement

This work was supported by National Institutes of Health under awards HL136659 and U01TR002062, by the National Institute on Aging of the National Institutes of Health under awards AG034676 and AG052425, and by the National Center for Advancing Translational Sciences (NCATS) under awards UL1 TR002377.

References

1. Sivapalaratnam S, Boekholdt SM, Trip MD, Sandhu MS, Luben R, Kastelein JJ, et al. Family history of premature coronary heart disease and risk prediction in the EPIC-Norfolk prospective population study. Heart. 2010;96(24):1985-9.
2. Valdez R, Yoon PW, Qureshi N, Green RF, Khoury MJ. Family history in public health practice: a genomic tool for disease prevention and health promotion. Annual review of public health. 2010;31:69-87.

3. Nasir K, Michos ED, Rumberger JA, Braunstein JB, Post WS, Budoff MJ, et al. Coronary artery calcification and family history of premature coronary heart disease: sibling history is more strongly associated than parental history. Circulation. 2004;110(15):2150-6.
4. Murabito JM, Pencina MJ, Nam B-H, D'Agostino RB, Wang TJ, Lloyd-Jones D, et al. Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. Jama. 2005;294(24):3117-23.
5. Mansour-Chemaly M, Haddy N, Siest G, Visvikis S. Family studies: their role in the evaluation of genetic cardiovascular risk factors. 2002.
6. Claassen L, Henneman L, Janssens ACJ, Wijdenes-Pijl M, Qureshi N, Walter FM, et al. Using family history information to promote healthy lifestyles and prevent diseases; a discussion of the evidence. BMC Public Health. 2010;10(1):1-7.
7. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? Genetics in Medicine. 2002;4(4):304-10.
8. Kolber MR, Scrimshaw C. Family history of cardiovascular disease. Canadian Family Physician. 2014;60(11):1016-.
9. Nadeem M, Ahmed SS, Mansoor S, Farooq S. Risk factors for coronary heart disease in patients below 45 years of age. Pakistan journal of medical sciences. 2013;29(1):91.
10. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association. 2016;23(6):1046-52.
11. Elixhauser A, Steiner CA, Whittington C. Hospital inpatient statistics, 1996: Diane Publishing Company; 1999.
12. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine. 2015;7(1):1-14.
13. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;26(9):1205-10.
14. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PloS one. 2017;12(7):e0175508.
15. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Developing and Evaluating Mappings of ICD-10 and ICD-10-CM codes to Phecodes. BioRxiv. 2019:462077.
16. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative toxicogenomics database (CTD): update 2021. Nucleic acids research. 2021;49(D1):D1138-D43.
17. Robinson PN, Mundlos S. The human phenotype ontology. Clinical genetics. 2010;77(6):525-34.
18. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research. 2014;42(D1):D966-D74.
19. Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, et al. The Human Disease Ontology 2022 update. Nucleic acids research. 2022;50(D1):D1255-D61.
20. Manemann SM, St Sauver JL, Liu H, Larson NB, Moon S, Takahashi PY, et al. Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population. BMJ open. 2021;11(6):e044353.
21. Amith M, He Z, Bian J, Lossio-Ventura JA, Tao C. Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. Journal of biomedical informatics. 2018;80:1-13.
22. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in bioinformatics. 2015;16(6):1069-80.