WikiMedMap: Expanding the Phenotyping Mapping Toolbox Using Wikipedia

Lina Sulieman^{1,*}, Patrick Wu^{1,2,*}, Joshua C. Denny^{1,3}, Lisa Bastarache¹

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

²Medical Scientist Training Program, Vanderbilt University School of Medicine, Nashville, TN, USA

³Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

* Authors made equal contributions to the paper

Correspondence to:

Lisa Bastarache, MS

Email: lisa.bastarache@vumc.org

Department of Biomedical Informatics

Vanderbilt University Medical Center

2525 West End Ave., Suite 1500

Nashville, TN 37203

Tel: (615) 322-0828

Abstract

Researchers utilizing phenotypic data from diverse sources require matching of phenotypes to standard clinical vocabularies. Mapping phenotypes to vocabulary can be difficult, as existing tools are often incomplete, can be difficult to access, and can be cumbersome to use, especially for non-experts. We created WikiMedMap as a simple tool that leverages Wikipedia and maps phenotype strings to standard clinical vocabularies. We assessed WikiMedMap by mapping phenotype strings from questionnaires in the UK Biobank and from Mendelian diseases in Online Mendelian Inheritance in Man (OMIM) database to eight vocabularies: International Classification of Diseases, Ninth Revision (ICD-9), ICD-10, ICD-O, Medical Subject Headings (MeSH), OMIM, Disease Database, and MedlinePlus. WikiMedMap outperformed conventional mapping tools in finding potential matches for phenotype strings. We envision WikiMedMap as a technique that complements existing and established tools to map strings to clinical vocabularies that usually do not coexist in one source.

Word Count: 2650 words

Keywords: biomedical vocabularies, text mining

Introduction

With the completion of the Human Genome Project in 2000, many genome-wide association studies (GWAS) have been published with the goal of understanding how the genetic variation influences phenotypes, including diseases. In the past decade, researchers have used increasingly large datasets and meta analyses to conduct genotype-phenotype association studies at much larger scales. Large cohorts like the UK Biobank (UKBB) (Sudlow et al., 2015), Million Veteran Program (Gaziano et al., 2016), and *All of Us* Research Program (Collins and Varmus, 2015) include dense phenotypic information from a variety of sources, including healthcare data, participant surveys, and objective measurements from in-person visits. Efforts such as the Electronic Medical Records and Genomics (eMERGE) Network (McCarty et al., 2011) and the Million Veteran Program (Gaziano et al., 2016) have demonstrated the power of electronic health records (EHRs) for genomic research. In addition, resources like the genome-wide association study (GWAS) catalog (Buniello et al., 2019) have curated data from thousands of studies for a diverse set of diseases and traits. Initiatives like the Undiagnosed Disease Network (Splinter et al., 2018) and MatchMaker (Buske et al., 2015) have developed large datasets of dense phenotypic information for rare diseases. Though “big data” has the potential to improve biomedical research, it can also create unique challenges in the biomedical research space.

Despite the availability of these datasets, it is often challenging to use phenotypic data to answer questions in biomedicine. While there are widely accepted standards for referencing genetic variants, (1000 Genomes Project Consortium et al., 2015; Kent et al., 2002) there are no universally accepted standards or guidelines for describing the human phenome (Banda et al., 2018; Wilcox, 2015). To describe human phenotypes, researchers have developed many ontologies and vocabularies to meet the needs of particular domains and use cases (Petersen et al., 1999; Shivade et al., 2014). For example, the World Health Organization developed International Classifications of Diseases (ICD) codes to track mortality and morbidity, and is commonly used for diagnosis and billing in healthcare (Steindel, 2010). A second example is National Library of Medicine’s (NLM) Medical Subject Headings (MeSH) vocabulary that is used to index scientific publications (Lipscomb, 2000). Third, Human Phenotype Ontology (HPO) was developed to describe the signs and symptoms of Mendelian disease (Robinson et al., 2008). Fourth, Online Mendelian Inheritance in Man (OMIM) and Orphanet were developed to specify monogenic conditions (McKusick, 2007). And fifth, Systematized Nomenclature of Medicine (SNOMED) is a standard terminology commonly used to represent diseases (Cornet and de Keizer, 2008).

Research groups have leveraged these disparate resources to create specific disease cohorts using EHR data by creating phenotype algorithms that map ICD diagnosis codes and other elements to phenotypes of interest (Denny et al., 2013, 2010; Hripcsak et al., 2018; Kirby et al., 2016; Newton et al., 2013; Wu et al., 2019). However, variability in the description of human phenotypes can impede researchers from efficiently conducting genotype-phenotype association studies (Sabb et al., 2009). For example, a researcher studying ovarian cancer in the UKBB must first identify the relevant data elements, which might be ICD10-CM codes or text in the survey questionnaires. At this time, this process cannot be fully automated, as there are many synonyms for “ovarian cancer”, including “ovarian carcinoma” and “Malignant neoplasm of ovary” (ICD-10 C56). To find an equivalent code for “ovarian cancer”, a researcher who is inexperienced phenotyping can review the literature or collaborate with clinical colleagues. While a purely manual effort is manageable for a small number of phenotypes, it is hard to scale this process for hundreds to thousands of phenotypes. Further, these linking efforts are continuous, as new phenotype maps are required when clinical vocabularies are updated and

new sources of phenotype information are made available. Hence, tools that partially automate the linking of phenotypes from disparate data sources to standard vocabularies are essential.

Some current tools and databases that facilitate translation between different medical and scientific vocabularies have allowed genotype-phenotype association studies to be conducted using EHR data. However, researchers may not have access to all the vocabularies or have the training necessary to use them appropriately. For instance, researchers commonly use the Unified Medical Language System (UMLS) (Humphreys et al., 1998), a medical terminology thesaurus that binds terms from >100 medical vocabularies to unique concept identifiers and thus can serve as an interlingua between many of these vocabularies. However, the UMLS is not a comprehensive solution to translate across vocabularies. It can be difficult to use, may not include some clinical vocabularies (e.g., consumer health vocabularies or cohort questionnaires), and may lack the necessary synonymy and some relationships (Geller et al., 2009; Zeng and Tse, 2006). The complex data structures and concept relationships in these resources creates an obstacle for scientists interested in mapping their phenotypes to a standard vocabulary, which can impede their research. Even for researchers who have extensive experience working with UMLS, automated mapping often fails to identify equivalent standard medical concepts and require input from domain experts familiar with both the phenotype in question and the standard clinical terminology (Hripcsak et al., 2018).

In this paper, we present a Wikipedia-based tool called WikiMedMap, an open-source tool designed to facilitate the translation of phenotype strings to standard vocabularies. WikiMedMap takes raw strings as inputs and outputs standard codes from common medical vocabularies including ICD-9, ICD-10, MeSH, DiseaseDB, OMIM, and Orphanet. Wikipedia has three features that make it a useful resource for linking medical terms. First, it has a large database of alternative names, common misspellings, and abbreviations to redirect search strings to the correct page (e.g. “regional enteritis” is redirected to the page “Crohn’s disease”), which is a powerful resource for string normalization. Second, pages for medical concepts in Wikipedia are structured using a specially-designed template that includes an “infobox” with classifications using a set of standard terminologies. Third, Wikipedia has an API to execute searches and access results. While it is not comprehensive, WikiMedMap is easy to use for identifying relationships between terms that may not be present in conventional resources like the UMLS.

Results

Comparison of WikiMedMap to UMLS

To assess the potential of WikiMedMap in the phenotyping pipeline, we used 691 UKBB phenotypes strings that UKBB participants were asked in questionnaires, to compare WikiMedMap with UMLS. We defined a match as a phenotype with at least one link between the phenotype and the standard medical vocabulary in question. Overall, we found that WikiMedMap finds more standard concept links than UMLS for UKBB terms for all tested vocabularies (Tables 1 and 2). For example, WikiMedMap found ICD-9 codes for 33 UKBB disease terms compared to 57 found with the UMLS. We also compared the performance of WikiMedMap to UMLS in mapping 347 HPO terms (Table 2). WikiMedMap found more links between the HPO terms for 0.75 (6/8) of the standard clinical vocabularies. UMLS found more concept links in MeSH (WikiMedMap 153, UMLS 170) and OMIM (WikiMedMap 71, UMLS 253).

To evaluate the performance of WikiMedMap in mapping phenotypes to standard clinical vocabulary, we compared the ICD-9 mappings from UMLS to those found using WikiMedMap. Compared to UMLS mapping of UKBB terms, WikiMedMap successfully mapped the same 0.89 (51/57) of terms to ICD-9, 0.96 (54/56) of terms to ICD-10, 0.79 (122/154) of terms to MeSH terms, 0.89 (100/112) of terms to MedlinePlus, and 0.37 (38/104) of terms to OMIM. For HPO terms, WikiMedMap and UMLS mapped 0.88 (59/67), 0.94 (51/54), 0.67 (114/170), 0.91 (67/74), and 0.26 (65/253) of the same terms to ICD-9, ICD-10, MeSH, MedlinePlus, and OMIM respectively.

Table 1. Number of UKBB phenotypes with ≥ 1 mappings for each standard clinical vocabulary for the 691 UKBB phenotypes.

Standard Clinical Vocabulary	WikiMedMap	UMLS	Both
ICD-9	332	57	51
ICD-10	344	56	54
ICD-O	24	0	0
MeSH	292	154	122
MedlinePlus	300	112	100
OMIM	114	104	38
Orphanet	6	0	0
Disease Database	282	0	0
Median (IQR)	287 (91, 308)	56 (0,106)	44 (0,65)

Table 2. Number of HPO phenotypes with ≥ 1 mappings for each standard clinical vocabulary for 347 HPO terms

Standard Clinical Vocabulary	WikiMedMap	UMLS	Both
------------------------------	------------	------	------

ICD-9	210	67	59
ICD-10	206	54	51
ICD-O	16	0	0
MeSH	153	170	114
MedlinePlus	152	74	67
OMIM	71	253	65
Orphanet	8	0	0
Disease Database	173	0	0
Median (IQR)	152 (57,181)	60 (0,98)	55 (0,65)

Discussion

In this study, we developed and applied WikiMedMap, a tool that mines Wikipedia to match phenotypes strings to standard clinical vocabularies. Compared to UMLS, WikiMedMap found matching clinical vocabulary codes for more UKBB phenotypes and HPO terms. We envision WikiMedMap as a complementary tool to existing resources by matching concepts codes for phenotype strings that may not be found with automated mapping strategies using conventional databases. By facilitating the mapping of phenotypes to standard clinical vocabularies, WikiMedMap can resolve a key bottleneck in biomedical research and has the potential to accelerate the translation of biomedical data to improve human health. We do not envision researchers with extensive training in biomedical data science and/or are familiar with UMLS and its derivatives as the primary users of WikiMedMap. On the other hand, biomedical researchers who are unfamiliar with biomedical data science tools, but are interested in conducting genotype-phenotype association studies can benefit from this tool. For example, WikiMedMap may be useful for researchers interested in extracting medical concepts from patient survey questionnaires, consumer health data, and disease community forums.

Since WikiMedMap leverages Wikipedia's web search tool, it has the built-in ability to handle disease acronyms, misspellings, and synonyms. Healthcare providers commonly use acronyms in clinical notes to save time in documenting patient interactions. In some cases, disease acronyms in clinical notes that are unfamiliar to researchers may impede phenotype mapping. For example, "Diastolic heart failure", "diastolic dysfunction", and "Heart Failure with preserved Ejection Fraction" refers to the evolving nomenclature of the same clinical entity and is commonly abbreviated "HFpEF". For this phenotype, WikiMedMap redirects "HFpEF" to the correct Wikipedia article ("Heart failure with preserved ejection fraction," n.d.) and extracts the appropriate codes. Narrative text on the internet, such as those found in disease forums or on Twitter, are other unconventional sources for health data. In those sources, users often describe their symptoms and diseases using lay medical terms (Doing-Harris and Zeng-Treitler, 2011; Zeng and Tse, 2006). WikiMedMap can help researchers map disease names and symptoms in social media posts to standardized medical concept codes. For example, using WikiMedMap to search for "stomach flu" directs to the Wikipedia page for "Gastroenteritis" and extracts the ICD-9 codes 008.8 "Viral enteritis NOS", 009.0 "Infectious colitis, enteritis, and gastroenteritis", 009.1 "Colitis, enteritis, and gastroenteritis of presumed infectious origin", and 558 "Other and unspecified noninfectious gastroenteritis and colitis".

WikiMedMap can be used to extract codes for phenotypes strings in studies that link genotypes to phenotypes, especially for researchers who are unfamiliar with conventional phenotyping approaches. Researchers may use WikiMedMap to map phenotypes to standard medical codes that traditional mapping methods may miss (e.g., linking "Self-harm" to ICD-10 codes X60-X84), and to map HPO and Mendelian diseases terms to codes in disparate terminologies. Moreover, this tool can empower researchers in their first-pass of hypothesis generation step by providing the codes that help them in identifying the cohort and extracting their EHR data.

We envision that WikiMedMap will have other potential valuable uses for biomedical researchers across numerous domains. Researchers can automatically extract the concepts codes for survey questions or answers rather than following a manual approach. In addition, WikiMedMap may be helpful for creating maps for disease or phenotype terms in languages other than English. We present four potential use cases for WikiMedMap (Figure 1): synonym, medical acronyms, phenotypes with difficulty in identifying equivalents in UMLS, and translation of medical concepts in English to other languages.

Limitations

Our tool has some limitations. First, we did not perform a comprehensive analysis to assess the accuracy of WikiMedMap mappings to UMLS. We made this decision, as WikiMedMap is designed to help researchers generate hypotheses, and does not replace standardized vocabularies. We recommend that researchers use more conventional mapping tools or collaborate with individuals with expertise with phenotype mapping for their final studies. Second, though WikiMedMap can parse and analyze disease names from short strings, this tool is not designed to tokenize and analyze clinical notes. For researchers interested in using WikiMedMap to analyze clinical notes, we recommend that researchers pre-process and tokenize clinical notes before feeding the smaller substrings into WikiMedMap. Other more sophisticated natural language processing packages, such as KnowledgeMap (Denny et al., 2003), cTAKES (Savova et al., 2010), and CLAMP (Soysal et al., 2017) can process longer strings and notes. However, they rely on UMLS as the main source for defining the clinical terms which subject them to some of the UMLS limitations discussed earlier. Third, we aimed to simulate real-world scenarios when we compared the performance of WikiMedMap to UMLS. Thus, we used simple string matching instead of concept relationships in the UMLS. We made this decision because we assumed that most of WikiMedMap's intended users would lack the technical expertise to use this tool in UMLS.

Future directions

In the future, we plan on expanding this package that has the ability to capture ICD codes ranges. For example, for the string "colon/cancer", the package currently captures the beginning and end ICD-9 codes, 153.0-154.1. An ideal output for "colon/cancer" would be all the ICD-9 codes between 153.0-154.1. Another potential direction would be to extract more detailed information about phenotypes such as symptoms, complications, and differential diagnosis associated with diseases.

Conclusions

Biomedical data science tools have allowed researchers to conduct studies that would have been impossible just ten years ago (Ezer and Whitaker, 2019). Phenotype mapping is one bottleneck that impedes translation of biomedical knowledge to improve human health. WikiMedMap facilitates the translation of phenotypes to several commonly used vocabularies and may aid in exploring datasets and generating EHR phenotype algorithms.

Materials and methods

WikiMedMap algorithm

WikiMedMap is a standalone package that leverages the Wikipedia API to extract concepts codes from different terminologies documented in Wikipedia. Figure 3 depicts a screenshot of our notebook for the tool. Our tool uses three main components that are easy to install on any computer or cloud service that supports Python. It relies on two python libraries: MediaWiki ("MediaWiki," n.d.), and Pandas ("Python Data Analysis Library — pandas: Python Data Analysis Library," n.d.). Pandas is an open source library that supports easy-to-use data structures and high performance computational analysis in Python . MediaWiki is a Python wrapper and parser for the original MediaWiki API, a web service that combines HTML, JSON, and XML to crawl and retrieve Wikipedia articles ("MediaWiki," n.d.).

We retrieved the Wikipedia page of the source or the search term using MediaWiki. Then, we parsed the page returned by the API to locate the presence of eight standardized vocabularies: ICD-9, ICD-10, MeSH, MedlinePlus, OMIM, Orphanet, DiseaseDB, and ICD-O. For each

terminology, we extracted the codes and the links to the code's page on the terminology's website. All analyses were conducted using Python (version 3.6.7), Pandas (version 0.23.4), and Wikipedia API (version 1.4.0).

The Wikipedia API retrieves the most probable page based on the passed source or search string. A user can provide the source string and WikiMedMap returns a list of concept codes from defined vocabularies in the Wikipedia infobox. We used the MediaWiki Python package and located the Wikipedia page of the given term and extracted the page content. MediaWiki returned the most relevant page that represented term whether by exact matching or submatching (Figure 2). Thus, the API often can return the correct Wikipedia entry even if the input string was different from the title of the Wikipedia page. For instance, for "stomach flu", the API returned the Wikipedia page for "Gastroenteritis".

Comparison of WikiMedMap to UMLS

We compared the effectiveness of the WikiMedMap to UMLS in extracting medical vocabularies for strings from two well-known biomedical research repositories: disease substrings from UKBB questionnaires and Mendelian disease terms from OMIM. We extracted codes for each string using UMLS and using WikiMedMap. In UMLS, we used string matching to retrieve vocabulary codes that matched the input string. We avoided using the substring match since doing so may increase the number of false-positive results. We then used the same strings in WikiMedMap, which returned all the matching codes. For UKBB and OMIM strings, we counted the number of unique strings that we were able to successfully map to at least one code for each of the eight standard vocabularies.

Software and data availability

The WikiMedMap code is available on three platforms: GitHub (<https://github.com/Linasulieman/WikiMedMap/blob/master/WikiMedMapCode.ipynb>), binder (<https://mybinder.org/v2/gh/Linasulieman/WikiMedMap/master>), and Google Colab (<https://colab.research.google.com/drive/1ZbIPYG6yvM6qRb5Yv8uOUYP38PulhL2u>). The files that include the UKBB surveys and OMIM Mendelian diseases are available as well.

Acknowledgements

This project was supported by the National Library of Medicine, of the National Institutes of Health, grant R01 LM010685.

Conflicts of interests

The authors have no conflict of interest to declare.

Figure 1 Use cases and extensions of WikiMedMap. For the string ‘colon/cancer’, WikiMedMap identifies the Wikipedia article for ‘Colorectal cancer’ (Wikipedia contributors, 2019) and extracts the relevant ICD codes: ICD-9 153.0-154.1, ICD-10 C18, C19, C20. With WikiMedMap, “Rheumatoid Arthritis” finds the relevant article in the Chinese version of Wikipedia and displays summary from the article to the user. WikiMedMap can understand medical acronyms, such as “Stomach flu”, which redirects to “Gastroenteritis” and extracts ICD-9. WikiMedMap identifies relevant ICD-10 X60-X84 codes that are relevant for the string “self-harm”.

Figure 2. UMLS and WikiMedMap pipeline to map phenotypes. (A) To use UMLS to automate high-throughput mapping of human phenotypes requires many steps. (B) WikiMedMap is simple and robust for automated high-throughput mapping of human phenotypes. UMLS: Unified Medical Language System. ICD: International Classification of Diseases.

```

In [13]: wikipedia.set_lang("en")
phenotype_string="Rheumatoid arthritis"

In [14]: diseasePage=wikipedia.page(phenotype_string)#Squamous cell skin cancer")

In [6]: page_external_links=diseasePage.references

In [11]: page_external_links=pd.DataFrame(page_external_links)
page_external_links.columns=["links"]
mapped_codes_list=[]

page_external_links=pd.DataFrame(page_external_links)
page_external_links.columns=["links"]
for key in concepts_external_links:
    code_page_link=page_external_links.loc[page_external_links["links"].apply(lambda x: concepts_external_links[key] in x)]
    if code_page_link.shape[0] > 0:
        for link_i in range(code_page_link.shape[0]):
            mapped_code_value=code_page_link.values[link_i].replace( concepts_external_links[key] , "")
            mapped_codes_list.append([key,mapped_code_value.replace(".htm", ""),code_page_link.values[link_i],diseasePage.title])

mapped_codes_df=pd.DataFrame(mapped_codes_list)
mapped_codes_df.columns=["Terminology", "Code", "CodeLink", "title", "page_id"]
print (mapped_codes_df)

```

	Terminology	Code	CodeLink \
0	ICD-10	M05	http://apps.who.int/classifications/icd10/brow...
1	ICD-10	M06	http://apps.who.int/classifications/icd10/brow...
2	ICD-10	M45	http://apps.who.int/classifications/icd10/brow...
3	MedlinePlus	000431	https://www.nlm.nih.gov/medlineplus/ency/artic...
4	ICD-9	714	http://www.icd9data.com/getICD9Code.ashx?icd9=714
5	MESH	D001172	https://www.nlm.nih.gov/cgi/mesh/2015/MB_cgi?f...
6	OMIM	180300	https://omim.org/entry/180300
7	DiseaseDB	ddb11506	http://www.diseasesdatabase.com/ddb11506.htm

Figure 3. A screenshot of WikiMedMap notebook page

References

- 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* **526**:68–74. doi:10.1038/nature15393
- Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. 2018. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* **1**:53–68. doi:10.1146/annurev-biodatasci-080917-013315
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousseau O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**:D1005–D1012. doi:10.1093/nar/gky1120
- Buske OJ, Schietecatte F, Hutton B, Dumitriu S, Misyura A, Huang L, Hartley T, Girdea M, Sobreira N, Mungall C, Brudno M. 2015. The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. *Hum Mutat* **36**:922–927. doi:10.1002/humu.22850
- Collins FS, Varmus H. 2015. A new initiative on precision medicine. *N Engl J Med* **372**:793–795. doi:10.1056/NEJMp1500523
- Cornet R, de Keizer N. 2008. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* **8 Suppl 1**:S2. doi:10.1186/1472-6947-8-S1-S2
- Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**:1102–1110. doi:10.1038/nbt.2749
- Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**:1205–1210. doi:10.1093/bioinformatics/btq126
- Denny JC, Smithers JD, Miller RA, Spickard A 3rd. 2003. “Understanding” medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc* **10**:351–362. doi:10.1197/jamia.M1176
- Doing-Harris KM, Zeng-Treitler Q. 2011. Computer-assisted update of a consumer health vocabulary through mining of social network data. *J Med Internet Res* **13**:e37. doi:10.2196/jmir.1636
- Ezer D, Whitaker K. 2019. Data science for the scientific life cycle. *Elife* **8**. doi:10.7554/eLife.43979
- Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, Guarino P, Aslan M, Anderson D, LaFleur R, Hammond T, Schaa K, Moser J, Huang G, Muralidhar S, Przygodzki R, O’Leary TJ. 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**:214–223. doi:10.1016/j.jclinepi.2015.09.016
- Geller J, Morrey CP, Xu J, Halper M, Elhanan G, Perl Y, Hripcsak G. 2009. Comparing inconsistent relationship configurations indicating UMLS errors. *AMIA Annu Symp Proc* **2009**:193–197.
- Heart failure with preserved ejection fraction. n.d. . *Wikipedia*. https://en.wikipedia.org/wiki/Heart_failure_with_preserved_ejection_fraction

- Hripcsak G, Levine ME, Shang N, Ryan PB. 2018. Effect of vocabulary mapping for conditions on phenotype cohorts. *J Am Med Inform Assoc* **25**:1618–1625. doi:10.1093/jamia/ocy124
- Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. 1998. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc* **5**:1–11.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**:996–1006. doi:10.1101/gr.229102
- Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB, Lingren T, Thompson WK, Savova G, Haines J, Roden DM, Harris PA, Denny JC. 2016. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* **23**:1046–1052. doi:10.1093/jamia/ocv202
- Lipscomb CE. 2000. Medical Subject Headings (MeSH). *Bull Med Libr Assoc* **88**:265–266.
- McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struwing JP, Wolf WA, eMERGE Team. 2011. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* **4**:13. doi:10.1186/1755-8794-4-13
- McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**:588–604. doi:10.1086/514346
- MediaWiki. n.d. https://www.mediawiki.org/wiki/API:Main_page
- Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC. 2013. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* **20**:e147–54. doi:10.1136/amiajnl-2012-000896
- Petersen LA, Wright S, Normand SL, Daley J. 1999. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *J Gen Intern Med* **14**:555–558.
- Python Data Analysis Library — pandas: Python Data Analysis Library. n.d. <https://pandas.pydata.org/>
- Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**:610–615. doi:10.1016/j.ajhg.2008.09.017
- Sabb FW, Burggren AC, Higier RG, Fox J, He J, Parker DS, Poldrack RA, Chu W, Cannon TD, Freimer NB, Bilder RM. 2009. Challenges in phenotype definition in the whole-genome era: multivariate models of memory and intelligence. *Neuroscience* **164**:88–107. doi:10.1016/j.neuroscience.2009.05.013
- Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* **17**:507–513. doi:10.1136/jamia.2009.001560
- Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* **21**:221–230. doi:10.1136/amiajnl-2013-001935
- Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, Xu H. 2017. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc*. doi:10.1093/jamia/ocx132
- Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheattle-Jarvela AM, Eng CM, Esteves C, Gahl WA, Hamid R, Jacob HJ, Kikani B, Koeller DM, Kohane IS, Lee BH, Loscalzo J, Luo X, McCray AT, Metz TO, Mulvihill JJ, Nelson SF, Palmer CGS, Phillips JA 3rd, Pick L, Postlethwait JH, Reuter C, Shashi V, Sweetser DA, Tifft CJ, Walley NM, Wangler MF, Westerfield M, Wheeler MT, Wise AL, Worthey EA, Yamamoto S, Ashley EA, Undiagnosed Diseases Network. 2018. Effect of Genetic Diagnosis on Patients with

- Previously Undiagnosed Disease. *N Engl J Med* **379**:2131–2139. doi:10.1056/NEJMoa1714458
- Steindel SJ. 2010. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc* **17**:274–282. doi:10.1136/jamia.2009.001230
- Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. 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**:e1001779. doi:10.1371/journal.pmed.1001779
- Wikipedia contributors. 2019. Colorectal cancer --- Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Colorectal_cancer&oldid=909387098
- Wilcox AB. 2015. Leveraging Electronic Health Records for Phenotyping In: Payne PRO, Embi PJ, editors. *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*. London: Springer London. pp. 61–74. doi:10.1007/978-1-4471-4646-9_4
- Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, Zhao J, Carroll R, Bastarache L, Denny JC, Theodoratou E, Wei W-Q. 2019. Developing and Evaluating Mappings of ICD-10 and ICD-10-CM Codes to PheCodes. *bioRxiv*. doi:10.1101/462077
- Zeng QT, Tse T. 2006. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc* **13**:24–29. doi:10.1197/jamia.M1761

Input

WikiMedMap

Output

Source search term
e.g., "colorectal cancer"

Step 1

Search for term using
Wikipedia API

Step 2

Extract codes from
Wikipedia page

Codes:
ICD-9
ICD-10
ICDO
MeSH

Input

UMLS preparation

Search

Output

Step 1

Sign up UMLS license
(annual survey
required for active
account)

Step 1

Retrieve UMLS concept
identifier using string
match

Step 2

Get familiar with UMLS
(time):

- UMLS data structure
- Understand
relationships between
tables

Step 2

Get all UMLS concepts
with the same UMLS
identifier (if included)

Step 3

- Load files in database
- Identify relationship
and mapping
between concepts

Step 3

Get target codes

Codes:
ICD-9
ICD-10

Source search term
e.g., "colorectal cancer"

Rewording, synonyms

Translation

Slang

UK BioBank

Colon cancer

Rheumatoid Arthritis in
Chinese

HFpEF

Self-harm

Searching for title

WikiMedMap

Did not find title

Redirect

Found title

Translate if specified

WIKIPEDIA
the free encyclopedia

Article Talk

Colorectal cancer

From Wikipedia, the free encyclopedia

Colorectal cancer (CRC), also known as b

维基百科
自由百科全书

条目 讨论 中文 中英混合

類風濕性關節炎

类风湿病， 慢性关节炎

Article Talk

Heart failure with preserved ejection fraction

From Wikipedia, the free encyclopedia

Heart failure with preserved ejection fraction (HFpEF) is a form of **heart failure** in which the heart is rigid and the left ventricle with each heartbeat ejected by the volume of blood when the greater than 50%.^[1] This may be measured by **echocardiography** or **cardiac catheterization**. HFpEF preserved ejection fraction, while the other half have a reduction in ejection fraction, called heart failure with reduced ejection fraction (HFrEF) include **hypertension**, **diabetes mellitus**, **obesity**, and **chronic kidney disease**.

Self-harm

From Wikipedia, the free encyclopedia

Self-harm, also known as self-injury, is defined as the intentional **suicide**.^{[1][2]} Other terms such as cutting and self-mutilation have been **used**.^{[3][4][5]} The most common form of self-harm is using a sharp object, or biting body parts. While clear definitions included in **diagnostic manuals**, but pulling (**intentional**) and the ingestion of poisons (**intentional**) from the poison self-harm.

replaced series 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900, 5000, 5100, 5200, 5300, 5400, 5500, 5600, 5700, 5800, 5900, 6000, 6100, 6200, 6300, 6400, 6500, 6600, 6700, 6800, 6900, 7000, 7100, 7200, 7300, 7400, 7500, 7600, 7700, 7800, 7900, 8000, 8100, 8200, 8300, 8400, 8500, 8600, 8700, 8800, 8900, 9000, 9100, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, 10000, 10100, 10200, 10300, 10400, 10500, 10600, 10700, 10800, 10900, 11000, 11100, 11200, 11300, 11400, 11500, 11600, 11700, 11800, 11900, 12000, 12100, 12200, 12300, 12400, 12500, 12600, 12700, 12800, 12900, 13000, 13100, 13200, 13300, 13400, 13500, 13600, 13700, 13800, 13900, 14000, 14100, 14200, 14300, 14400, 14500, 14600, 14700, 14800, 14900, 15000, 15100, 15200, 15300, 15400, 15500, 15600, 15700, 15800, 15900, 16000, 16100, 16200, 16300, 16400, 16500, 16600, 16700, 16800, 16900, 17000, 17100, 17200, 17300, 17400, 17500, 17600, 17700, 17800, 17900, 18000, 18100, 18200, 18300, 18400, 18500, 18600, 18700, 18800, 18900, 19000, 19100, 19200, 19300, 19400, 19500, 19600, 19700, 19800, 19900, 20000, 20100, 20200, 20300, 20400, 20500, 20600, 20700, 20800, 20900, 21000, 21100, 21200, 21300, 21400, 21500, 21600, 21700, 21800, 21900, 22000, 22100, 22200, 22300, 22400, 22500, 22600, 22700, 22800, 22900, 23000, 23100, 23200, 23300, 23400, 23500, 23600, 23700, 23800, 23900, 24000, 24100, 24200, 24300, 24400, 24500, 24600, 24700, 24800, 24900, 25000, 25100, 25200, 25300, 25400, 25500, 25600, 25700, 25800, 25900, 26000, 26100, 26200, 26300, 26400, 26500, 26600, 26700, 26800, 26900, 27000, 27100, 27200, 27300, 27400, 27500, 27600, 27700, 27800, 27900, 28000, 28100, 28200, 28300, 28400, 28500, 28600, 28700, 28800, 28900, 29000, 29100, 29200, 29300, 29400, 29500, 29600, 29700, 29800, 29900, 30000, 30100, 30200, 30300, 30400, 30500, 30600, 30700, 30800, 30900, 31000, 31100, 31200, 31300, 31400, 31500, 31600, 31700, 31800, 31900, 32000, 32100, 32200, 32300, 32400, 32500, 32600, 32700, 32800, 32900, 33000, 33100, 33200, 33300, 33400, 33500, 33600, 33700, 33800, 33900, 34000, 34100, 34200, 34300, 34400, 34500, 34600, 34700, 34800, 34900, 35000, 35100, 35200, 35300, 35400, 35500, 35600, 35700, 35800, 35900, 36000, 36100, 36200, 36300, 36400, 36500, 36600, 36700, 36800, 36900, 37000, 37100, 37200, 37300, 37400, 37500, 37600, 37700, 37800, 37900, 38000, 38100, 38200, 38300, 38400, 38500, 38600, 38700, 38800, 38900, 39000, 39100, 39200, 39300, 39400, 39500, 39600, 39700, 39800, 39900, 40000, 40100, 40200, 40300, 40400, 40500, 40600, 40700, 40800, 40900, 41000, 41100, 41200, 41300, 41400, 41500, 41600, 41700, 41800, 41900, 42000, 42100, 42200, 42300, 42400, 42500, 42600, 42700, 42800, 42900, 43000, 43100, 43200, 43300, 43400, 43500, 43600, 43700, 43800, 43900, 44000, 44100, 44200, 44300, 44400, 44500, 44600, 44700, 44800, 44900, 45000, 45100, 45200, 45300, 45400, 45500, 45600, 45700, 45800, 45900, 46000, 46100, 46200, 46300, 46400, 46500, 46600, 46700, 46800, 46900, 47000, 47100, 47200, 47300, 47400, 47500, 47600, 47700, 47800, 47900, 48000, 48100, 48200, 48300, 48400, 48500, 48600, 48700, 48800, 48900, 49000, 49100, 49200, 49300, 49400, 49500, 49600, 49700, 49800, 49900, 50000, 50100, 50200, 50300, 50400, 50500, 50600, 50700, 50800, 50900, 51000, 51100, 51200, 51300, 51400, 51500, 51600, 51700, 51800, 51900, 52000, 52100, 52200, 52300, 52400, 52500, 52600, 52700, 52800, 52900, 53000, 53100, 53200, 53300, 53400, 53500, 53600, 53700, 53800, 53900, 54000, 54100, 54200, 54300, 54400, 54500, 54600, 54700, 54800, 54900, 55000, 55100, 55200, 55300, 55400, 55500, 55600, 55700, 55800, 55900, 56000, 56100, 56200, 56300, 56400, 56500, 56600, 56700, 56800, 56900, 57000, 57100, 57200, 57300, 57400, 57500, 57600, 57700, 57800, 57900, 58000, 58100, 58200, 58300, 58400, 58500, 58600, 58700, 58800, 58900, 59000, 59100, 59200, 59300, 59400, 59500, 59600, 59700, 59800, 59900, 60000, 60100, 60200, 60300, 60400, 60500, 60600, 60700,

[Back to top](#)
[Home](#)
[About Us](#)
[Contact Us](#)
[Privacy Policy](#)
[Terms of Service](#)

	Terminology	Code	Code Link
0	100-10	100	http://apple.com/retail/us/100/100/100/100/100/
1	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/
2	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/
3	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/
4	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/
5	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/
6	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/
7	100-10	100	http://apple.com/retail/us/100/100/100/100/100/100/