BIOMEDICAL LITERATURE MINING AND KNOWLEDGE DISCOVERY OF PHENOTYPING DEFINITIONS

Samar Hussein Binkheder

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

July 2019

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

____
Josette Jones, RN, Ph.D, Chair

____
Lang Li, Ph.D.

April 8, 2019

____
Sara Kay Quinney, Ph.D.

____
Huanmei Wu, Ph.D.

____
Chi Zhang, Ph.D.

DEDICATION

I dedicate this dissertation to my parents, my husband, my children Ahmad and Basma, my family, and my friends. I would like to thank all of them for their love, support, and encouragement during my Ph.D. journey.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my committee members: Dr. Josette Jones, Dr. Lang Li, Dr. Sara Kay Quinney, Dr. Huanmei Wu, and Dr. Chi Zhang. Besides my committee members, I would like to thank Dr. Heng-Yi Wu for his mentorship and guidance during this research. My committee members and Dr. Heng-Yi Wu, your support, guidance, and motivation have helped during my Ph.D. Journey. I also would like to thank you all for supporting this dissertation work with constructive suggestions and feedback. Without this, it would not be possible to accomplish this work. Finally, I would like to express my appreciation to all members of Dr. Lang Li lab for all of their support for the past three years.

Samar Hussein Binkheder

BIOMEDICAL LITERATURE MINING AND KNOWLEDGE DISCOVERY OF PHENOTYPING DEFINITIONS

Phenotyping definitions are essential in cohort identification when conducting clinical research, but they become an obstacle when they are not readily available. Developing new definitions manually requires expert involvement that is labor-intensive, time-consuming, and unscalable. Moreover, automated approaches rely mostly on electronic health records' data that suffer from bias, confounding, and incompleteness. Limited efforts established in utilizing text-mining and data-driven approaches to automate extraction and literature-based knowledge discovery of phenotyping definitions and to support their scalability. In this dissertation, we proposed a text-mining pipeline combining rule-based and machine-learning methods to automate retrieval, classification, and extraction of phenotyping definitions' information from literature. To achieve this, we first developed an annotation guideline with ten dimensions to annotate sentences with evidence of phenotyping definitions' modalities, such as phenotypes and laboratories. Two annotators manually annotated a corpus of sentences (n=3,971) extracted from full-text observational studies' methods sections (n=86). Percent and Kappa statistics showed high inter-annotator agreement on sentence-level annotations. Second, we constructed two validated text classifiers using our annotated corpora: abstract-level and full-text sentence-level. We applied the abstract-level classifier on a large-scale biomedical literature of over 20 million abstracts published between 1975 and 2018 to classify positive abstracts (n=459,406). After retrieving their full-texts (n=120,868), we extracted sentences from their methods sections and used the full-text sentence-level classifier to extract positive sentences (n=2,745,416). Third, we performed a literature-based discovery utilizing the positively classified sentences. Lexica-based methods were used to recognize medical concepts in these sentences (n=19,423). Co-occurrence and association methods were used to identify and rank phenotype candidates that are associated with a phenotype of interest. We derived 12,616,465 associations from our large-scale corpus. Our literature-based associations and large-scale corpus contribute in building new data-driven phenotyping definitions and expanding existing definitions with minimal expert involvement.

Josette Jones, RN, Ph.D, Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADE	Adverse Drug Event
ANSI	American National Standards Institute
CDC	Centers of Disease Control
CHF	Chronic Heart Failure
CRS	Chronic Rhinosinusitis
CU	Columbia University
CDM	Common Data Model
CTD	Comparative Toxicogenomics Database
CAD	Coronary Artery Disease
CPT	Current Procedural Terminology
CYPs	Cytochrome P450
DTM	Document-Term Matrix
DDI	Drug-Drug Interaction
DILI	Drug Induced Liver Injury
DILIN	Drug Induced Liver Injury Network
EHR	Electronic Health Record
eMERGE	Electronic Medical Records and Genomics
EXC	Exclusion Conclusion
XML	Extensible Markup Language
FN	False Negative
FP	False Positive
FAERS	FDA Adverse Event Reporting System
FDA	Food and Drug Administration
HOI	Health Outcome of Interest
HLGT	High Level Group Term
HLT	High Level Term
INC	Inclusion Conclusion
IBD	Inflammatory Bowel Disease
I2b2	Informatics for Integrating Biology at the Bedside
IE	Information Extraction
IR	Information Retrieval
IRB	Institutional Review Board
IAA	Inter-Annotator Agreement
ITC	Intermediate Conclusion
ICD	International Classification of Diseases
IHTSDO	International Health Terminology Standards Organization
ICH	International Conference on Harmonisation
IMRAD	Introduction, Methods, Results And Discussion
IDF	Inverse Document Frequency
LOINC	Logical Observation Identifiers Names and Codes
LR	Logistic Regression
LLT	Lowest Level Term
MedDRA	Medical Dictionary for Regulatory Activities
MeSH	Medical Subject Headings

MEDIC	Merged Disease Vocabulary
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OMIM	Online Mendelian Inheritance in Man®
PD	Pharmacodynamics
PK	Pharmacokinetics
PheKB	Phenotype Knowledgebase
PT	Preferred Terms
PMR	Proportional Mortality Ratio
SMO	Sequential Minimal Optimization
SIDER	Side Effect Resource
SRS	Spontaneous Reporting Systems
SHARP	Strategic Health IT Advanced Research Projects
SOC	System Organ Class
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
TN	True Negative
TP	True Positive
TF	Term Frequency
TPR	True Positive Rate
T2DM	Type 2 Diabetes Mellitus
UMLS	Unified Medical Language System
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

CHAPTER ONE: INTRODUCTION

1.1 Background

Adverse drug events (ADE) are a big concern in health care, leading to significant costs, morbidity, and mortality (Eriksson, Werge, Jensen, & Brunak, 2014; "U.S. Food and drug administration. Preventable Adverse Drug Reactions: A Focus on Drug Interactions," 2018). It has been estimated that over 2 million serious ADE are reported yearly ("U.S. Food and drug administration. Preventable Adverse Drug Reactions: A Focus on Drug Interactions," 2018). In the first quarter of 2017, the U.S. Food and Drug Administration (FDA) reported around 297,010 serious outcomes and around 44,693 deaths due to ADEs (Somnath Pal, 2017). The annually estimated cost to manage ADEs in the United States is up to 30.1 billion dollars due to hospitalizations, prolonged hospital stays, and prescriptions to treat ADEs (Sultana, Cutroneo, & Trifiro, 2013). The Institute of Medicine has defined an ADE as an unintended drug-related "injury caused by medical management", and they note that most ADEs can be prevented (Homsted, 2000). Furthermore, polypharmacy, where a patient is taking more than one drug, increases the risk of drug reactions ("U.S. Food and drug administration. Preventable Adverse Drug Reactions: A Focus on Drug Interactions," 2018). Serious adverse events should be reported to the FDA, such as death, life-threatening, hospitalization, disability, permanent damage, or birth defects ("U.S. Food and drug administration. What is a Serious Adverse Event?,").

Pharmacovigilance is "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems" (Organization, 2002). One of the main objectives of pharmacovigilance is the early detection of novel and underreported adverse events and drug correlations (Rave Harpaz, Haerian, Chase, & Friedman, 2010) as a part of post-marketing drug discovery. A sub-domain of pharmacovigilance is pharmacoepidemiology that aims in quantifying ADEs in a large population (de Vries & de Jong-van den Berg, 2001). Aims of pharmacovigilance and pharmacoepidemiological research includes: early detection of ADEs as well as identification of contributing risk factors and its quantitative aspects (de Vries & de Jong-van den Berg, 2001). Post marketing surveillance is important to detect both anticipated and unanticipated adverse effects (Czaja et al.). Studies show that early discovery of post-marketing ADEs (R. Harpaz et al., 2013) and identification of causes is necessary to decrease the occurrence harmful events (Tache, Sonnichsen, & Ashcroft, 2011). Post-marketing research provides the opportunity to study factors that contribute to the risk of ADE in general population, such as pharmacokinetics in patients with organ impairment, drug's dose and frequency, and genotype (Sultana et al., 2013).

There are two major sources for mining ADE-drug associations: spontaneous voluntary reporting and electronic health records (EHRs). FDA Adverse Event Reporting System (FAERS) is an example of a spontaneous voluntary reporting system where health professionals, consumers, and manufacturers send reports of adverse events. EHRs have become an emerging source for pharmacovigilance, which is similar to FAERS, support hypothesis generation in areas like drug-adverse effect associations (Castro et al., 2014). Unlike the challenges of voluntary reporting of suspected ADEs (e.g. bias and underreporting), EHRs longitudinal data is capable of providing measurements of drug's harm in actual patients (Eriksson et al., 2014) using medications in real-world settings (Castro et al., 2014) contributing to advancement of the medical knowledge ("U.S. Food and drug administration. Preventable Adverse Drug Reactions: A Focus on Drug Interactions," 2018). However, it also can add some challenges when conducting EHR-based studies for ADEs or any phenotype.

1.2 Problem Statement

Even though EHR secondary research helped in advancement of the overall population health, it accompanied with several challenges. To conduct an EHR-based study, one of the earliest stages in EHR mining is the identification of a cohort of specific cases (Q. Li et al., 2014) which needs a phenotyping case definition. These phenotyping case definitions might not be readily available for all conditions, especially when dealing with large-scale phenotypes. Furthermore, conventional methods to create new case definitions require experts' knowledge or to use existing case definitions require literature evidence and reviews. Either of these methods can be time-consuming and labor-intensive. Within the EHR setting, incorrectly identified phenotypes cases can result in unreliable and biased results (Macdonald, Kilty, & van Walraven, 2016). Therefore, we have identified the lack of availability of phenotyping definitions for many cases of interest or their inconsistencies (Hansen et al., 2013; R. L. Richesson, Hammond, et al., 2013) as a knowledge gap, which creates a barrier when a researcher needs to identify cases for EHR-based research.

With this, there is a need to develop informatics approaches and data-driven approaches to define cases (Lasko, Denny, & Levy, 2013) on large-scale settings (Rubbo et al., 2015). Utilizing biomedical literature, there is a need to discover and understand the repeatable patterns and relationships of phenotypes that help in building new phenotyping case definitions and support existing definitions (Overby et al., 2013; Rasmussen et al., 2014). Such health informatics tools help to utilize literature-based phenotype definitions information for future applications (Kirby et al., 2016; Rasmussen et al., 2014) and to support the knowledge-discovery of unknown relationships across these phenotypes.

Based on our research interest, the cases of interest for this work are primarily derived as phenotypes with an evidence to be an ADE (Duke et al., 2012; H. Y. Wu, Zhang, Desta, Quinney, & Li, 2017). However, several other phenotypes were included because some are considered confounding, risk factors, or other clinical concepts. In other words, our proposed approach is generalizable to not only ADE cases, but also to other phenotypes and diseases.

1.3 Overview of the Dissertation

This dissertation presents an informatics approach for mining phenotyping definitions in the biomedical literature. We developed a text-mining pipeline combining rule-based and machine-learning methods to automate retrieval, classification, and extraction of phenotyping definitions' information from literature. To our knowledge, there is no existing work for mining literature-based phenotyping definitions. To achieve this goal, we proposed three Aims.

Aim 1. Develop a corpus for annotating phenotyping case definitions in published literature. An annotated corpus is needed for building and evaluating text-mining tools. As a starting point, we created a list of phenotype of interest to collect abstracts and full texts. These phenotypes were ADEs that were selected from an observed ADE evidence in previously published in-vitro Pharmacokinetics (PK), in-vivo PK, and clinical Pharmacodynamics (PD) studies. Moreover, we analyzed the presentation of phenotyping case definitions in the biomedical literature to identify sections where this information can be located. A new annotation schema is developed to manually annotate a corpus on a sentence level. One of the major goals for developing the annotated domain-specific corpus is to serve as a gold standard for developing text-mining tools (J.-D. Kim, Ohta, & Tsujii, 2008) that is accomplished in Aim 2.

Aim 2. Automated extraction of sentences with phenotyping case definitions from biomedical literature. These ADE terms served as the building block for developing our dictionary and lexica, corpus, and text-mining pipeline. For building the lexica, a comprehensive dictionary was developed using standard terminologies reflecting major entities, including clinical diagnoses, procedures, and drugs. This dictionary can assist in a number of text-mining tasks, such as named entity recognition (NER), information retrieval, and information extraction. Moreover, we developed a text-mining pipeline that will be composed of two levels of classification. First, Abstract-level classifier to retrieve abstracts with relevant studies describing phenotyping definitions. Second, Full-text sentence-level classifier to classify sentences within methods sections of the full text with that show evidence of phenotyping case definitions. These classifiers will utilize informatics approaches of text-mining, machine learning, and rule-based. The validated classifiers are applied on a large-scale literature and further information extraction and knowledge discovery is performed in Aim 3.

Aim 3. Perform a discovery-based study to evaluate and validate literature-based phenotyping case definitions of selected phenotypes. In this Aim, we will utilize sentences with evidence of phenotyping case definitions from the large-scale screening of literature as well as the lexica and dictionary (from Aim 2). We aimed to use a data-driven approach to prioritize the co-occurrence of terms for a phenotype of interest in literature. Moreover, an approach was proposed to rank the sentences for each ADE of interest based on its significant associated terms. Lastly, we will compare, validate, and evaluate the results of literature-based phenotypes with existing sources. Figure 1 shows the theoretical model for this dissertation.

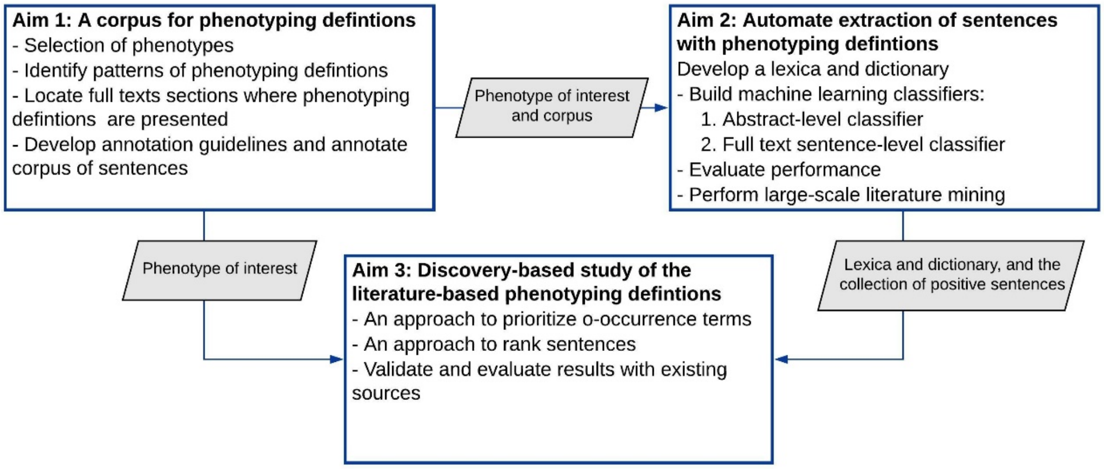

Aim 1: A corpus for phenotyping defintions
- Selection of phenotypes
- Identify patterns of phenotyping defintions
- Locate full texts sections where phenotyping defintions are presented
- Develop annotation guidelines and annotate corpus of sentences
Phenotype of interest and corpus
Aim 2: Automate extraction of sentences with phenotyping defintions
Develop a lexica and dictionary
- Build machine learning classifiers:
1. Abstract-level classifier
2. Full text sentence-level classifier
- Evaluate performance
- Perform large-scale literature mining
Phenotype of interest
Aim 3: Discovery-based study of the literature-based phenotyping defintions
- An approach to prioritize o-occurrence terms
- An approach to rank sentences
- Validate and evaluate results with existing sources
Lexica and dictionary, and the collection of positive sentences

Figure 1 Theoretical model for the dissertation

1.4 Significance

First, to facilitate the utilization of EHRs for clinical research. Here we consider ADEs as an example of that we use in this dissertation. Studies showed that experimentally validating large numbers of drugs-ADEs associations is not feasible and the use of multiple resources that together would be able to derive true supportive evidence of these associations (Banda, Callahan, et al., 2016). Spontaneous voluntary reporting, such as FAERS, have been widely used for signal detection of ADEs. However, the spontaneous voluntary reporting suffers from some limitations, such as bias in reporting, lack of causality ADE-drug relationship, incomplete data, and duplicated reports. For example, ADEs signal scores from FAERS data by themselves do not provide causal ADE-drug relationship when used by itself. Instead, FAERS data provide advantage for mining ADE-drug associations in initial stages of ADE-drug discovery as a guiding resource rather than hypothesis generation. Therefore, it is necessary to utilize other resources, such as biomedical literature and EHRs, to generate ADE-drug causal relationships and hypotheses (R. Harpaz et al., 2012). EHRs are a potential resource to support translational research and hypothesis generation in areas of drug safety (Yao, Zhang, Li, Sanseau, & Agarwal, 2011), and to better understand health outcomes. However, there are several requirements, which majorly is the availability of phenotyping case definitions for building the appropriate cohort of cases, which we are addressing in this work.

Second, to support the biomedical research towards high-dimensional drug interactions in EHRs. Here we extends our example of ADEs research on the need to support high-dimensional discovery, which is necessary as a part of pharmacovigilance research in which the identification of ADEs during the post-marketing stage (Banda, Evans, et al., 2016). However, there are still limitations where existing studies mostly work on a small scale of associations (Duke et al., 2012), and tend to use traditional methods to identify associations of single drugs and single ADE phenotypes (L. Li, 2015). On the other hand, clinical trials are capable of capturing multiple phenotypes, but they suffer from low sample size and lack "real world" factors that contribute to the efficacy of drugs. Some examples of mining high-dimensional ADE-drug associations (L. Li, 2015) are the identification of six novel DDIs that increased the risk of myopathy (Duke et al., 2012). Another example is the identification of 171 novel drug interactions associated with eight (L. Li, 2015; Tatonetti, Fernald, & Altman, 2012). Therefore data mining approaches are capable of expanding the dimensions of associations in health records (L. Li, 2015). This expansion requires also scaling up the phenotyping process and their definitions by using data-driven and data mining approaches.

Identifying health outcome of interest (HOI) is still a concern in observational studies (Fox et al., 2013). For example, Zhang et al. have developed a statistical model to identify high-dimensional myopathy-drug associations in EHR. Myopathy definition was adopted from literature and were mapped to the concept IDs of Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) (Zhang et al., 2015). However, when dealing with large number of ADEs, there is a need to more scalable methods for defining ADEs in EHRs. Accurate definitions are critical in building cohort of patients experiencing ADEs and should be matched with study's goals. De Bie et al. in their drug safety surveillance study defined several ADE phenotypes of interest, such as acute liver injury and upper gastrointestinal bleeding, using iterative process that was initiated by using clinical criteria published in the literature. Diagnosis codes, laboratories, and clinical notes were used to define these ADE phenotypes (de Bie et al., 2015). De Bie et al. stated that their cohort selection has affected the estimation of statistical power of their study, where different definitions can provide different results.

> “The extrapolation of relevant safety outcomes from adults to children does not always work and that it is very important to choose age appropriate events and definitions when setting up EHR-based pediatric surveillance systems” (de Bie et al., 2015).

In the following sections, we identified the challenges associated with developing or using standard phenotyping definitions.

Third, developing new phenotyping definitions is complex. Such development process generates several potential challenges that elaborate in the complexity and inconsistency of the phenotyping definitions. The first challenge is that the development of a new case definition is a long, time-consuming, and labor-intensive process (Lasko et al., 2013; Park & Choi, 2014). A multidisciplinary team works on developing and designing phenotyping definitions for mining EHR data where manual review, multiple iterations, and validation can be also needed (Carroll, Eyler, & Denny, 2011). Further, such a process requires an extensive manual review of EHR charts. For example, Hsu et al. developed an algorithm to identify chronic rhinosinusitis (CRS) cases and controls using ICD-9 International Classification of Diseases, Ninth Revision (ICD-9) and Current Procedural Terminology (CPT) codes. They validated the algorithm using manual chart review as the reference standard (Hsu, Pacheco, Stevens, Smith, & Avila, 2014). As a part of the manual chart review, the authors identified the need of reviewing the encounter notes and CT sinus results, which was held by two reviewers and took a length of 40 hours for reviewing only 200 cases. Therefore, such algorithm development process cannot be scalable to a bigger set of phenotypes or ADEs (Hsu et al., 2014). Even though several Natural Language Processing (NLP) tools have been developed in the medical domain, clinical narrative is the most challenging part of the phenotype definition development because it requires extensive human involvement (Park & Choi, 2014). Therefore, using common terminologies and expanding dictionaries can support newly developed NLP tools for mining ADE phenotypes in EHR.

The second challenge, the multiple cycles of communication during the development process of case definitions can be time-consuming, inflexible, and error-prone. One traditional way of developing new definition is expert-driven, an iterative process that requires multiple cycles of communication between the clinical researcher and the data analyst. The clinical researcher usually uses the phenotype definition in a human-readable format. On the other hand, the data analyst's role is to convert the phenotype definition from the human-readable format into the computable format. However, mismatches between the desired definition and the computable definition are highly possible (Xu et al., 2015). Communication challenges can also arise due to the multidisciplinary nature needed for developing the phenotype definition where it requires input from different medical professionals, such as geneticists, clinicians, informaticians, and epidemiologists (Mo et al., 2015). Therefore, creating a source that supports scientific collaboration for these definitions will be imperative.

The third challenge, after the development process, many case definitions lack portability across different institutions, which can affect phenotyping definitions' generalizability. One of the important aspects for any phenotyping definition is portability across different EHR systems and/or institutions, especially when applying complex selection criteria. The nature of phenotyping definitions across different studies can be highly diverse and inconsistent, which is a recognized problem between different institutions (Christley, Duffy, & Martin, 2012; Rubbo et al., 2015). Definition portability facilitates comparing, sharing, validating, assessing, and disseminating of phenotyping definitions (Simonett et al., 2015). The validation step across multiple institutions is important to ensure that the definition is performing well across different populations (Liao, Ananthakrishnan, et al., 2015). A major player to the lack of portability is the lack in consistency between different EHR systems in recording clinical data (Malinowski, Farber-Eger, & Crawford, 2014). This can be also true in clinical data that are presented in the biomedical literature, as clinical researchers tend to report the same clinical terms differently. In addition, Hsu et al. stated that definition generalizability can be affected by variations in clinical use of standardized codes between different individuals, centers, or specialties (Hsu et al., 2014). Therefore, it is recommended to share phenotyping definitions to improve portability and generalizability (Overby et al., 2013). Properly documenting published phenotyping definitions in the literature can have an important role in discovering the patterns that can cause lack of portability across different phenotyping definitions.

Fourth, literature-based phenotyping definitions are critical, but still not scalable. Reporting phenotyping case definitions is very inconsistent in studies where it can even lack some broad or basic description (Yao et al., 2011). The variation of these definitions and lack of their availability can inhibit the research of ADEs (Christley et al., 2012; Rubbo et al., 2015). The variation in textual descriptions of the phenotyping defining is one of the biggest challenges of implementing existing phenotyping definitions. Inconsistency can be due to the fact that there is no internationally agreed standard that assists in conducting and reporting phenotyping definitions as well as its validation studies (Rubbo et al., 2015). This can result in an inconsistent implementation and interpretation of the definitions because of variation in concept granularities and ambiguities. Several factors contributed to the variations and inconsistencies of definitions that were published in observational studies including the increase use of EHR data and other data sources, and the use of standard codes (e.g. ICD- codes) (Shankar-Hari et al., 2016). These definitions can differ across different studies depending on study purpose and design. For example, Shankar-Hari et al. conducted three studies to develop a new definition for identifying septic shock, which was called to review by Society of Critical Care Medicine (SCCM) and the European Society of Intensive Care Med (ESICM) in January 2014. In their systematic review, 44 studies were identified. They stated that septic shock definitions in the literature reported different cutoffs and combinations for the following phenotypes "blood pressure, fluid resuscitation, vasopressors, serum lactate level, and base deficit" (Shankar-Hari et al., 2016). Their final consensus definition for septic shock was:

> "Adult patients with septic shock can be identified using the clinical criteria of hypotension requiring vasopressor therapy to maintain mean BP 65 mmHg or greater and having a serum lactate level greater than 2 mmol/L after adequate fluid resuscitation" (Shankar-Hari et al., 2016).

Therefore, harmonization of phenotyping definitions that were already published in the literature can be very effective in generating stronger phenotype definitions.

The process of searching literature for evidence-based EHR-phenotyping definitions lack scalability, and can be difficult, slow, and time-consuming. However, there is evidence that "repeatable patterns within phenotyping algorithms exist" (Rasmussen et al., 2014). Using a systematic approach to learn the repeatable patterns in phenotyping definitions can be a strong starting point for advancing the process of their development and our understanding of these definitions (Rasmussen et al., 2014). Moreover, using systematic approaches can improve the consistency and validity of the phenotyping definitions that are generated from different institutions (Overby et al., 2013). A rare condition called drug-induced liver injury (DILI) is an example of a case definition with such semantic challenge (Overby et al., 2013). Overby et al. addressed the challenge of harmonizing the DILI definition between two institutions, Columbia University (CU) and Mayo clinic. They reported that there was significant differences in DILI phenotyping definitions between CU and Mayo clinic. These differences are reflected by the final scope or goal of the study. For example, unlike DILI definition used in CU that used "DILI caused by any medications", the DILI definition that Mayo clinic used was narrowed by medications selection as "DILI caused by a medication preparation of interest to Drug Induced Liver Injury Network (DILIN)". The final harmonized definition selected the narrowed criteria used by Mayo clinic. They reported that factors influencing these differences could be due to the baseline population size, data access characteristics, and multiple interpretations of EHR phenotyping definitions (Overby et al., 2013).

Literature-derived evidence can be established through extraction of evidence from unstructured text using combination of text-mining and data mining approaches, for example, extraction pairs of biological entities (Ananiadou, Kell, & Tsujii, 2006). Don R. Swanson describes knowledge discovery and hypothesis generation from literature as

> "assembling pieces of a puzzle to reveal an unnoticed, unintended, but not unintelligible pattern. The fragmentation of science into specialties makes it likely that there exist innumerable pairs of logically related, mutually isolated literatures" (Swanson, 1988).

In 1988, Swanson discovered a relationship between migraines and magnesium deficiency by identifying 11 factors common between the two conditions, which are

> "type A personality, vascular tone and reactivity, calcium channel blockers, spreading cortical depression, epilepsy, serotonin, platelet activity, inflammation, prostaglandins, substance P, and brain hypoxia" (Swanson, 1988).

Consequently, Swanson generated a hypothesis that supplementing food with magnesium can improve migraines. Unlike traditional manual literature, text-mining supports scalable and high-throughput research, and is capable of discovering unknown associations and patterns hidden in unstructured text. Biomedicine is a “data-rich but hypothesis-poor science”. Accelerating knowledge discovery and hypotheses generation by using data-driven methods followed by experimental data validation is needed (Ananiadou et al., 2006). With this, our goal in this work is to use large-scale evidence extracted from literature to support knowledge discovery of patterns that can assist in defining phenotypes in the EHRs. This can assist in the future by combining biomedical literature-derived knowledge with EHR to advance scientific research as well as novel discoveries and hypotheses generation (Ananiadou et al., 2006; Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012; Spasic, Ananiadou, McNaught, & Kumar, 2005) in areas of EHR-based research.

1.5 Innovation

This work is the first, based on our knowledge, to use text-mining approaches to mine phenotyping definition published in the biomedical literature. Therefore, it is innovative in several ways.

Development of a novel foundational informatics approach for annotating and mining phenotyping definitions in the literature. In this work, we analyzed the major patterns of phenotyping case definitions’ modalities, such as data sources, standardized codes, clinical, and laboratory information. This was supported by using existing terminologies and ontologies as well as proposing new keywords that characterize these phenotyping definitions. Based on the analysis of features and ontologies, we developed a new annotating schema to manually annotate these definitions. The gold standard corpora can assist in training and testing classifiers to automate the extraction of the definitions’ information and to eliminate the barriers of collecting these definitions. The schema, the corpora, and the classifier will contribute in the field of text mining of the biomedical information.

A literature-based large-scale screening of evidence-based phenotyping definitions is capable of performing an advanced information retrieval and extraction. Consequently, it will introduce a new resource with a large collection of phenotyping definitions. Due to the fact that there is limited research in the area of validation of phenotyping algorithms in the literature (C. Barber, D. Lacaille, & P. R. Fortin, 2013), researchers need to conduct a series of literature reviews to validate them (C. Barber et al., 2013). Therefore, the collection of definitions will enable research to improve, use, and validate these definitions as well as discover variability patterns in different definitions.

An approach to prioritize phenotype concepts derived from large-scale corpus that assist in defining phenotypes and identify novel associations. Unlike previous work that relied on abstracts rather than full texts (Botsis & Ball, 2013), in this work we proposed using full texts for more comprehensive retrieval of information. According to our knowledge, this is the first work that uses full texts for mining published phenotyping definitions.

With this, we believe that this work will lay as the foundation of literature-based mining phenotyping case definitions information in the field of health informatics.

1.6 Description of the Chapters

In section 1.2, we proposed three Aims to perform our literature mining and discovery-based study of phenotyping definitions. This dissertation is composed of five chapters, including this chapter (Chapter 1). Each of following chapters (two, three, and four) has introduction, background, methods, discussion, and results. Chapter five is the conclusion chapter. The description of each chapter is as the following:

Chapter 2—this chapter is titled as "A corpus for annotating phenotyping definitions sentences in biomedical literature". To address Aim 1 of this dissertation, the following tasks are performed in this chapter: selection of phenotypes, abstracts and full texts collection and selection, construction process of corpus, and annotation process as well as inter-annotator agreement evaluation.

Chapter 3—this chapter is titled as "An automated text-mining approach of phenotyping definitions in the biomedical literature". This chapter covers Aim 2, with the following tasks: building lexica and dictionary, corpus used for training and building the model, information retrieval and extraction, classifiers performance evaluation, and a literature large-scale screening pipeline.

Chapter 4—this chapter is titled as "Discovery study to represent and validate literature-based phenotyping definitions". This chapter covers Aim 3 that further extend the information extraction. In this chapter, we performed co-occurrence analysis on the positively classified sentences, DICE coefficients for ranking phenotypes and sentences and for building network graphs, and validation of literature-based co-occurrence across three sources.

Chapter 5—this is the discussion and conclusion chapter. It connects the results chapters of this dissertation and provides a discussion of the implication of this work.

CHAPTER TWO: A CORPUS FOR ANNOTATING SENTENCES WITH INFORMATION OF PHENOTYPING DEFINITIONS IN BIOMEDICAL LITERATURE

In Chapter 1, we introduced the problem that phenotyping case definitions are not available for all phenotypes of interest. There are several efforts for generating phenotyping definitions, but the efforts in both literature-based mining and knowledge discovery of phenotyping definitions are still limited. Our aim is to develop an automated approach to mine these definitions in the literature. However, the state-of-art text-mining methods are based on a labeled corpus (Dogan, Leaman, & Lu, 2014; Shatkay & Craven, 2012). Therefore, the goal of this Aim is to build corpora and guidelines to annotate phenotyping definitions in the biomedical literature. These corpora are used in the following chapters.

2.1 Introduction

The current direction is moving towards the utilization of electronic health records (EHRs) for clinical research, including ADE discovery (Chiang et al., 2018; Czaja et al.; Yeleswarapu, Rao, Joseph, Saipradeep, & Srinivasan, 2014; J. Zhao, Henriksson, Asker, & Bostrom, 2015). EHR-based studies, in general, rely on defining a phenotype in a population in order to advance the knowledge of a disease or an adverse event (Glicksberg et al., 2018; R. L. Richesson, Hammond, et al., 2013). In terms of EHR-based research, the term “phenotype” can refer to observable patient characteristics inferred from clinical data, such as biomarkers and diseases (Hripcsak & Albers, 2013, 2017; R. L. Richesson, Hammond, et al., 2013; Shivade et al., 2014). An accurate phenotyping definition is critical to establish a cohort of patients for EHR-based research (Glicksberg et al., 2018; Gurwitz & Pirmohamed, 2010; Kirby et al., 2016; R. L. Richesson, Hammond, et al., 2013), including cross-sectional, and association studies (Banda, Seneviratne, Hernandez-Boussard, & Shah, 2018). Utilizing either structured or unstructured data (Banda et al., 2018; Hripcsak & Albers, 2017; W. Q. Wei & Denny, 2015), there are several methods that can be used for EHR phenotyping, including natural language processing (NLP), rule-based systems, statistical analysis, data mining, machine learning, and hybrid systems (Banda et al., 2018; Shivade et al., 2014). Depending on the phenotype of interest as well as study’s research question, standard queries for defining a phenotype can consist of any of the following: logical operators, standardized codes, data fields, and values sets (concepts derived from vocabularies or data standards) (R. L. Richesson, Hammond, et al., 2013). With this, the goal is to develop the annotation guidelines that are able to capture such information about phenotyping case definitions.

There are two types of methods for developing a phenotyping definition either developing new case definitions, or utilizing existing case definitions' information that are already available in different sources. Traditional phenotyping relies on expert knowledge and these definitions might change overtime (Hripcsak & Albers, 2017). This task is challenging due to complexity of EHRs and heterogeneity of patient’s records (Banda et al., 2018). Furthermore, it is also a labor-intensive process where a multidisciplinary team is needed with team members includes biostatistician, clinical researcher, EHR informatician, and NLP expert (Liao, Cai, et al., 2015). One example of expert-driven definitions is a study that identified patients with chronic rhinosinusitis (CRS) for a better understanding of the “prevalence, pathophysiology, morbidity, and management” using EHR data. Their team developed an algorithm to define CRS cases using ICD-9 diagnosis codes and Current Procedural Terminology (CPT) codes. The process took several iterations until they achieved predictive positive value of 91%. Further, they stated that manual review of notes and sinus CT results took two reviewers 40 hours, which is not scalable to larger number of patients or notes. Not to mention, their CRS definition has only been tested on one site and its performance is not known in other centers (Hsu et al., 2014). This creates further difficulties when creating new definitions. Lessons learned from The Electronic MEdical Records and GEnomics (eMERGE) Network (Gottesman et al., 2013) showed that the process of developing, creating, and validating a phenotyping definition for a single disease is time consuming and can take around 6-8 months. Consequently, the eMERGE network developed Phenotype KnowledgeBase (PheKB) (Kirby et al., 2016), which is a phenotype knowledgebase collaborative environment that allows collaborating and commenting between groups of researchers who were invited by a primary author. PheKB (Kirby et al., 2016) uses an expert-driven approach where new phenotyping definitions are generated by multi-institutional input and available publicly for use. PheKB provides a library of definitions for several phenotypes and incorporates several data modalities, majorly including standard codes, laboratories, medications, and NLP.

Another method relies on deriving phenotyping definitions from existing data sources, such as EHR and biomedical literature. Some of these were addressed manually using systematic reviews (Claire Barber, Diane Lacaille, & Paul R. Fortin, 2013; Fiest et al., 2014; Leong et al., 2013; Lui & Rudmik, 2015; Macdonald et al., 2016; Pace, Peters, Rahme, & Dasgupta, 2017; Souri et al., 2017) or automatically using computational approaches. Systematic reviews have a big role in the medical knowledge. However, with the massive amount of information, there is still a need to use automated approaches to extract medical knowledge; for example, the rate of published clinical trials articles is over 20,000 per year while around 3000 systematic reviews were indexed in MEDLINE yearly. Overall, systematic reviews are time-consuming and labor-intensive (Cohen et al., 2010). On the other hand, the automated approaches for mining phenotypes in the literature were mostly focused on extracting phenotype terminologies (Collier et al., 2015; Henderson, Bridges, Ho, Wallace, & Ghosh, 2017; D. Zhao & Weng, 2011) in studies without the defined scope of EHR-based studies. Some of these studies (Botsis & Ball, 2013; D. Zhao & Weng, 2011) have addressed only one phenotype at a time and utilized abstracts rather than full text. Unlike full texts that are richer in information, abstracts are not sufficient for the granularity of phenotyping definitions information. Furthermore, such approaches might not be generalizable especially when working on a large-scale set of phenotypes. In "Automating case definitions using literature-based reasoning" (Botsis & Ball, 2013), Botsis and Ball (Botsis & Ball, 2013) have developed a corpus and a classifier to automate extraction of "anaphylaxis" definitions from literature. However, Botsis and Ball (Botsis & Ball, 2013) only relied on abstracts rather than full text that provides more rich information. In addition, the classifier was developed for only one condition "anaphylaxis". Even though they focused on some cues of phenotyping definitions e.g. signs and symptoms, they did not consider other cues of phenotyping definitions (e.g. standardized codes and laboratory measures) (Botsis & Ball, 2013). Therefore, this effort did not address our information needs that reflecting modalities of phenotyping definitions such these used in PheKB.

With the goal of minimizing human involvement, we realized that there is a lack of phenotyping tools (Shivade et al., 2014) addressing or automating the extraction of existing definitions from scientific literature. There is a strong motivation for this research, to our knowledge; there is no existing corpora that address our information needs. An example of developing corpus for phenotypes is PhenoCHF (Noha Alnazzawi, Thompson, & Ananiadou, 2014; N. Alnazzawi, Thompson, Batista-Navarro, & Ananiadou, 2015), an annotated corpus by domain experts for phenotypic information relevant to Congestive Heart Failure from literature and EHR. The PhenoCHF corpus data was derived from i2b2 (the Informatics for Integrating Biology at the Bedside) discharge summaries dataset (Uzuner, 2009) and five full text articles retrieved from PubMed that covered the characteristics of Chronic Heart Failure (CHF) and renal failure. However, PhenoCHF focused only on one condition, CHF, and it was built on a small set of only five full text articles. Furthermore, they did not annotate contextual cues for phenotyping case definitions.

2.2 Background

2.2.1 Phenotyping definitions

Different institutions view a phenotyping case definition differently. For example, Strategic Health IT Advanced Research Projects (SHARP), which is a collaboration effort (academic and industries partners) to advance the secondary use of clinical data. It views a phenotyping definition as the

> “inclusion and exclusion criteria for clinical trials, numerator and denominator criteria for clinical quality metrics, epidemiologic criteria for outcomes research or observational studies, and trigger criteria for clinical decision support rules, among others” (Chute et al., 2011).

On the other hand, eMERGE phenotyping definitions extends to include practices as the

> “algorithmic recognition of any cohort within EHR for a defined purpose. These purposes were inspired by the algorithmic identification of research phenotypes” (Chute et al., 2011).

Further practices that eMERGE used in developing phenotyping definitions include other data modalities, such as diagnostics fields, laboratory values, medication use, and NLP-based (Chute et al., 2011). Here, we summarize definitions from different perspectives for defining a phenotyping case definition, which are:

> "The identification of patients' cohort in the EHR by defining an inclusion and exclusion criteria performed for structured data and unstructured clinical text" (Pathak, Kho, & Denny, 2013).
>
> "An EHR-based cohort that only select subset of patients who fulfill the pre-defined phenotype" (Yu et al., 2015).
>
> "EHR-based research is concerned about cohort selection that is the identifying cases and controls for a phenotype of interest. A phenotype definition is developed from combining multiple EHR data, such as billing codes, medications, narrative notes, and laboratory data" (Carroll et al., 2011; Liao, Cai, et al., 2015; Roden & Denny, 2016).
>
> "The process of deriving a cohort of a phenotype of interest using either low-throughput or high-throughput approaches" (R. L. Richesson, Sun, Pathak, Kho, & Denny, 2016).
>
> "The identification of cohort utilizing risk factors, clinical or medical characteristics and complications" (R. Richesson et al.; Yadav, Steinbach, Kumar, & Simon, 2018).

2.2.2 Applications of phenotyping definitions

The aim of this section is only to provide some overview about study designs in the biomedical research where phenotyping case definitions can be used. A phenotyping case definition can be applied to several types of studies, such as cross-sectional, association, and experimental (Banda et al., 2018). For example, pharmacovigilance, predictive modeling, clinical effectiveness research, and risk factors studies are considered use cases for the association case-control or cohort studies. More examples are shown in (Banda et al., 2018). Different study designs require different cohort designs as well as definitions where one phenotype can be defined in different ways depending on the study needs and research question. For instance, type 2 diabetes mellitus, which can be defined as

> “patients with type 2 diabetes or far more nuanced, such as T2DM patients with stage II prostate cancer and urinary urgency without evidence of urinary tract infection” (Banda et al., 2018).

There are two major types of studies in the biomedical domain: primary research that directly collect data and secondary research that relies published information or sources of data. The focus of this section is on the primary research since it is the used research for EHR-phenotyping. Primary research has observational, also called epidemiological studies, and interventional studies, also called experimental studies (Thiese, 2014). Study designs for observational study designs are ecological, proportional mortality, case-crossover, cross-sectional, retrospective and prospective cohort. Each of these has its own strengths and weaknesses. Examples of some of the primary studies that we cover in this work (Thiese, 2014):

- Ecological study design: Generally, called retrospective, and it is used to estimate a prevalence of a disease or an ADE in a population. The grouping is based on geographical locations or temporal associations.
- Proportional mortality ratio study (PMR): Identify relationships between exposure and outcomes. E.g. cardiovascular deaths among different ethnic groups.
- Cross-sectional studies, also prevalence studies: Samples are selected based on exposure without knowing their outcome.
- Case-control study design: Samples are identified based on the case status. This is the optimal study design for rare diseases.
- Retrospective and prospective cohort study design: Cohort studies is to identify patients based on the exposure and observe the development of the outcome of interest either for the future or for the past. Prospective is the gold standard for observational studies.

New research, such as pharmacovigilance, is moving towards the emergence of electronic health information, machine learning, and NLP (Sarker & Gonzalez, 2015). EHRs provide complementary data with some flexibility with extended periods tracking, large sample size, and data heterogeneity (Yadav et al., 2018). The availability of a cohort can create several opportunities for data mining and modeling such as risk models, adverse event detection, measuring the effect of intervention, and building evidence-based guidelines (Yadav et al., 2018). Cohort identification can be accomplished by using phenotyping definitions, which classify patients with specific disease based on EHR data, can be manually developed by experts or machine learned. A phenotyping definition shared some major features, such as logic, temporality, and the use of standard codes (Newton et al., 2013). Furthermore, examples of data categories that are commonly used across institutions are age, sex, race, ethnicity, height, weight, blood pressure, inpatient and outpatient diagnosis codes, laboratory tests, and medications (Newton et al., 2013). On the other hand, there are some challenges with cohort identification process that vary depending on the study type. The phenotyping process is more sophisticated than creating simple code (Banda et al., 2018). Several factors can contribute to their complexity, including the used methods and confounding. For example, when defining acute or less-defined phenotypes, one critical step is addressing confounding using matching of gender and age. These confounders are relatively easy to address, but others, such as co-diseases, might be more difficult. In some of the studies, Castro et al. were not able to identify methods for matching controls in EHR data. Case-control studies may inherent some limitations in detecting comorbidity such as insufficient controls, identification of correct confounders, and matching process. Literature-based comorbidity associations derived by clinical-expert is considered as a reference standard to compare the performance of the matched controls. However, the study reported that those clinical-expert driven associations between a list of PheWAS disease groupings and inflammatory bowel disease (IBD) generated some disagreement among gastroenterologists. Instead, Castro et al stated that their goal is to compare matching algorithms methods in order to identify clinically meaningful comorbidity associations (Castro et al., 2014).

2.2.3 Medical corpora for text mining

Text mining application mostly relies on supervised learning requires a corpora that is a collection of text annotated by experts. This is due to the challenges of recognizing terms as the example provided by Rodriguez-Esteban R for: “the text “early progressive multifocal leukoencephalopathy” could possibly refer to any, or all, of these disease terms: “early progressive multifocal leukoencephalopathy,” “progressive multifocal leukoencephalopathy,” “multifocal leukoencephalopathy,” and “leukoencephalopathy””.

Such annotations based on expert knowledge can be used to train machines, for example, on recognizing biomedical terms in text (Rodriguez-Esteban, 2009). An annotated high-quality corpus requires experienced annotators and comprehensive guidelines (Dogan et al., 2014). The manually annotated corpus can serve as a gold standard for building automated systems including statistical, machine learning, or rule-based (H. Gurulingappa et al., 2012). Examples of annotated biological corpora, are GENIA for annotating biological terms (J. D. Kim, Ohta, Tateisi, & Tsujii, 2003), BioCreative[1] for annotating biological entities in literature e.g. genes and proteins (Krallinger et al., 2015), and BioNLP[2] that is a collection of corpora, such as Colorado Richly Annotated Full-Text Corpus (CRADF)[3] and Protein Residue Corpora[4], for annotating biological entities. Other usages of an annotated corpus is as curated data to create literature-based knowledgebase, such as MetaCore[5] and BRENDA8[6] for enzyme functional data (H. Gurulingappa et al., 2012). However, these are mostly restricted to specific domains such as biological domain which annotates information, such as gene names, protein names, cellular location or events (e.g. protein-protein interaction) (H. Gurulingappa et al., 2012). Availability of corpora in the medical domain is even more limited than biological domain. One of the major reasons is that medical domain confronted with data availability and ethical issues of using electronic medical records (H. Gurulingappa et al., 2012) , including privacy and confidentiality and Health Insurance Portability and Accountability Act (HIPAA) regulations (Ozair, Jamshed, Sharma, & Aggarwal, 2015). Examples of biomedical corpora are Text Corpus for Disease Names and Adverse Effects for annotating diseases and adverse effects entities (Harsha Gurulingappa, Klinger, Hofmann-Apitius, & Fluck, 2010), CLinical E-Science Framework (CLEF) for annotating medical entities and relations (e.g. drugs, indications, findings) in free text of 20,000 cancer patient records (Roberts et al., 2009), and Adverse Drug Effects (ADE) corpus[7] for annotating ADEs entities (H.

[1] http://www.biocreative.org/news/corpora/biocreative-iii-corpus/
[2] http://bionlp-corpora.sourceforge.net/
[3] http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml
[4] http://bionlp-corpora.sourceforge.net/proteinresidue/index.shtml
[5] http://www.genego.com/metacore.php
[6] http://www.brenda-enzymes.org/
[7] https://sites.google.com/site/adecorpus/home/document

Gurulingappa et al., 2012). None of the available corpora serves our needs for this task to annotate contextual cues of defining a phenotype in observational studies on sentence-level annotations from full texts, such as the presence of codes, laboratory tests, and type of data used.

With this, our aim is annotating a corpus that capture sentences with not only phenotype concepts, but also contextual cues of a phenotyping definition that are presented in the literature. We believe that EHR-based studies will provide relevant information for defining phenotypes. An annotation schema is developed, and it serves as a foundational approach for annotating phenotyping definitions-related information in the literature. Both the corpus and the guidelines are designed based on extensive textual analysis of sentences to reflect phenotyping definitions information and cues. Ten dimensions are proposed to annotate the corpus at the sentence-level. Furthermore, after identifying the presence or absence of the ten dimensions, the level of evidence for each sentence was generated automatically using rule-based approach to ensure consistency and accuracy of annotations. All sentences in the methodology section were extracted from full text research papers. To the best of our knowledge, there is no existing corpus that is publicly available for annotating sentences with contextual cues of phenotyping definitions from biomedical full texts.

2.3 Methods

The procedure of a corpus construction consists of documents selection and sentence-level annotation (Verspoor et al., 2013). The documents selection starts with selection of phenotypes of interest that can assist in searching for abstracts. After that, abstracts collection prepared and full texts of selected abstracts were downloaded for the sentence-level annotation. For the sentence-level annotation, ten dimensions are proposed to annotate sentences with cues of a phenotyping case definition e.g. biomedical terms, and standard codes. Finally, their conclusions derived to an overall level of evidence for each sentence.

2.3.1 Selection of Phenotypes

Our group is primarily interested in ADEs (Duke et al., 2012; H. Y. Wu et al., 2017). Therefore, we identified our phenotypes of interest based on our previous work of literature-based discovery (Duke et al., 2012; H. Y. Wu et al., 2017) that have identified drug-drug interactions (DDIs) due to interactions among five Cytochrome P450 (CYPs) enzymes, including CYP2C8, CYP2C9, CYP2C19, CYP2D6, and CYP3A. These CYPs have a significant role in drug metabolism leading to several DDIs (Ogu & Maxa, 2000; J.-F. Wang & Chou, 2010). Furthermore, text-mining technology were used to extract DDI evidence and their corresponding ADEs from biomedical literature. DDIs were identified with evidence in all types of DDI studies, including clinical pharmacodynamics (PD), clinical pharmacokinetics (PK), and in vitro PK studies (H. Y. Wu et al., 2017). Among those clinical PD abstracts with 986 drugs pairs, we explored ADEs from those abstracts containing substrates of five major metabolizing enzymes above mentioned. The drug-enzyme relationships were collected from Flockhart table[1] and FDA. As a result, a list of ADEs (n = 673) was used as the primary list of phenotypes. All the ADE terms for those substrates were Medical Dictionary for Regulatory Activities terminology (MedDRA) (Brown, Wood, & Wood, 1999) preferred terms (PT).

To narrow down our phenotypes of interest, we identified ADEs that showed evidence of drugs-ADEs linkage in Side Effect Resource (SIDER) database (Kuhn, Letunic, Jensen, & Bork, 2016) and found that 398 ADEs were successfully linked to the side effects in SIDER database. At the end, expert reviews were performed by two experts who are Lang Li, Ph.D., and Sara Quinney, Pharm.D., Ph.D. to finalize the list of phenotypes of interest. They excluded ADE terms that are did not meet our lab research interests, such as terms related to infections and cancer. The final list of phenotypes of interest is 279 ADEs (Appendix 1). Figure 2 shows the process of the selection of phenotypes.

[1] https://drug-interactions.medicine.iu.edu/Main-Table.aspx

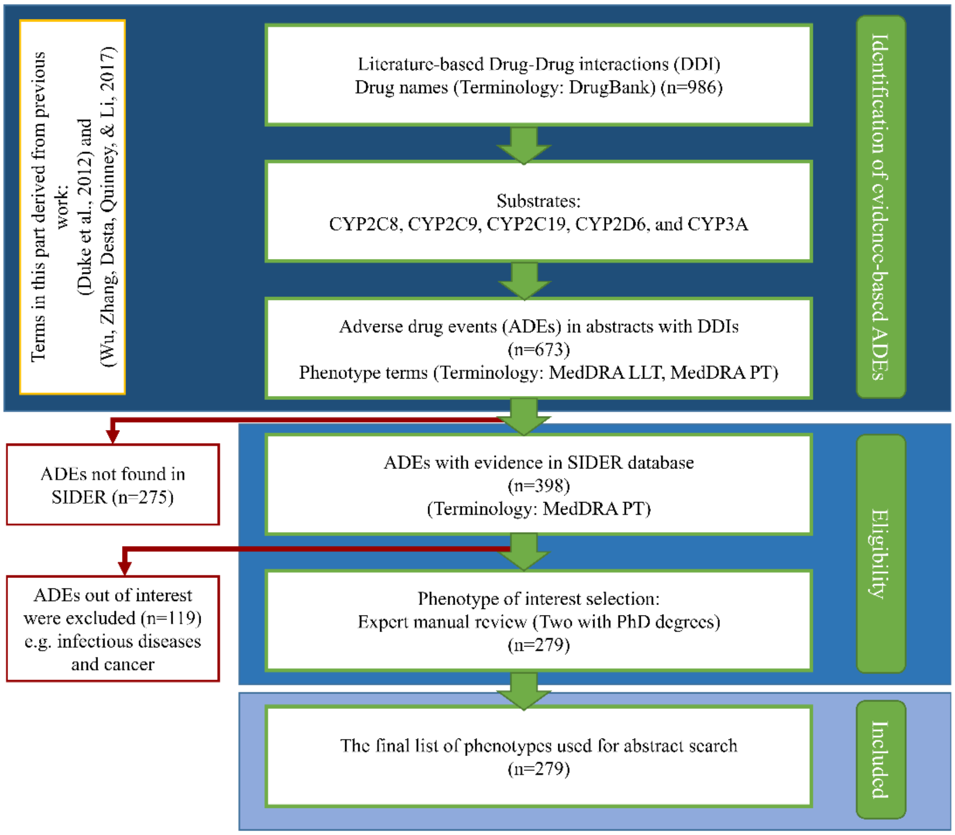

Terms in this part derived from previous work: (Duke et al., 2012) and (Wu, Zhang, Desta, Quinney, & Li, 2017)
Literature-based Drug-Drug interactions (DDI)
Drug names (Terminology: DrugBank) (n=986)
Substrates:
CYP2C8, CYP2C9, CYP2C19, CYP2D6, and CYP3A
Adverse drug events (ADEs) in abstracts with DDIs
(n=673)
Phenotype terms (Terminology: MedDRA LLT, MedDRA PT)
Identification of evidence-based ADEs
ADEs not found in SIDER (n=275)
ADEs with evidence in SIDER database
(n=398)
(Terminology: MedDRA PT)
ADEs out of interest were excluded (n=119) e.g. infectious diseases and cancer
Phenotype of interest selection:
Expert manual review (Two with PhD degrees)
(n=279)
Eligibility
The final list of phenotypes used for abstract search
(n=279)
Included

Figure 2 Flowchart of selection of phenotypes

2.3.2 Abstracts and full texts collection and selection

To search the literature for observational studies, we consulted a medical librarian to assist in building search queries to ensure the highest coverage. A review study reported that due to the broad nature of phenotyping studies, it can be difficult to perform one search that is capable of capturing all EHR phenotyping studies (Banda et al., 2018). Therefore, we collected our abstracts based on two search criterions:

First, we searched PubMed database to identify observational studies of our phenotypes of interest. The searching query consist of [an ADE phenotype of interest term] combined with a set of keywords that were tested to retrieve relevant observational studies (see Table 1). We did not put restrictions on the year of publication and the searched was performed on November 2017. The total number of retrieved abstracts without duplications was 1323 abstracts. One reviewer manually reviewed each abstract to select articles that met the inclusion criteria described in Table 1. It also shows the exclusion criteria that was applied to exclude abstracts. A total of 800 abstracts met our inclusion criteria. From the 800 abstracts, a subset of 57 abstracts were randomly selected for full-text sentence-level annotation task (PMIDs in Appendix 2).

Table 1 Abstract Inclusion-Exclusion criteria

Searching Query	**[A phenotype of interest term]** AND electronic health record (code OR codes OR algorithm* or "case definition" OR "phenotyping" OR "case identification" OR claim OR administrative)
Inclusion Criteria	1. Abstracts should satisfy each of the following: English, full text available and original research. 2. The primary source of data is EHR or EMR. Some accepted terms: Registry, administrative data. 3. The article should use observational data (population-based, surveillance, or cohort/cases) either retrospectively or prospectively. 4. Clearly describe a case definition or algorithm according to any of the following criteria: coding algorithms (SNOMED, ICD9/10, CPT, LOINC, RxNorm, UMLS, READ), laboratory, natural language processing (NLP), or inclusion and exclusion criteria.
Exclusion Criteria	1. Review articles 2. Non-human studies 3. Nurses/practitioners as primary population of the study 4. Not real-world data: e.g. simulation data 5. Tools, systems, or reporting systems that do not address phenotyping or descripting phenotyping definition.

Second, we used abstracts from a previous search that was performed by two reviewers. The used search queries were more generalized such as "electronic health record AND myopathy" (all queries are presented in Appendix 3). The downside of these queries is that it generates large number of abstracts that can be time-consuming and labor-intensive to review all of them. The reviewers collected some relevant abstracts from these search queries. From these collected abstracts, we randomly selected 29 abstracts. The query searches with PMIDs are showed in Appendix 3.

With this, the total number of abstracts derived from the two search criterions is 86 abstracts. We achieved this number based on our goal to reach around 4000 sentences from methods sections. We downloaded their full texts and we tokenized them into sentence tokens using a package called Perl::Tokenizer as preparation for the annotation process. In addition, we manually fixed sentences that were tokenized improperly. After that, we extracted sentences within methods sections.

2.3.3 Corpus construction

The annotation guidelines were developed based on textual analysis of the cues in sentences with a phenotyping definition information that were inspired by major data modalities of phenotyping definitions used in PheKB (Kirby et al., 2016). We performed sentence-level annotations with three major categories for each sentence, which are: inclusion, intermediate, and exclusion. The sentence-level annotations' categories were derived based on the availability of ten dimensions that are shown in Table 2 with their descriptions and examples. Furthermore, some of these dimensions have sub-dimensions. The detailed annotation guidelines is available in Appendix 4. The annotation construction is as the following:

First, inclusion category includes sentences that show evidence of at least one of the dimensions that characterize a phenotyping definition (Table 2). We identified five dimensions for the inclusion category, which are "Biomedical & Procedure", "Standard codes", "Medications", "Laboratories", and "Use of NLP". The proposed dimensions are represented as either keywords or more complex as events where co-occurrence of more than one keyword occurs. For example, "Standard Codes" dimension is represented by the presence of any keyword relative to "Standard Codes", such as ICD9, SNOMED, or a diagnostic code. On the other hand, "Biomedical & Procedure", "Medications", "Laboratories", and "Use of NLP" require an event presence such as co-occurrence of two keywords that were identified for each dimension. A sentence is categorized as positive for the inclusion category if it shows evidence of any of these five dimensions (Table 2), which we called inclusion conclusion (INC) is true (Table 3).

Second, intermediate category includes sentences that do not show direct evidence of a phenotyping definition, but it can assist by providing supporting evidence for phenotyping. We identified two dimensions for the intermediate category, which are "data Entities" and "Study Design or Institutional Review Board (IRB)". Since different studies have different research questions and designs, intermediate category can assist in capturing data types information that matches the study's goals (Yadav et al., 2018). A sentence is categorized as positive for the intermediate category if it shows evidence of any of the two dimensions (Table 2), which we called intermediate conclusion (ITC) is true (Table 3).

Third, exclusion category includes sentences that are out of the scope of a phenotyping definition or phenotyping. A sentence is categorized as positive for the exclusion category if it shows evidence of any of the three dimensions (Table 2), which we called exclusion conclusion (EXC) is true (Table 3).

Finally, the final decision is the overall sentence-level of evidence derived from INC, ITC, and EXC. We note that some sentences can have evidence of more than one dimension which determines final sentence-level conclusions (INC, INT, EXC) in Table 3. We used rule-based approach to produce four final sentence-level decisions, which are Positive, INTERMEDIATE_I, INTERMEDIATE_II, and Negative. The goal is to create an accumulative evidence in each sentence based on the presence of any of the three conclusions (INC, ITC, EXC). This help to ensure consistency, accuracy, and quality of the annotations. Table 3 shows the criteria of the seven rules (R1, R2, R3, R4, R5, R6, and R7). R8 final decision where all the three conclusions (INC, ITC, EXC) are false was combined with R7 indicating negative evidence.

Table 2 Sentence-level annotation’s categories, dimensions, and sub-dimensions

1. Inclusions category (n = 5)	**Description**	**Examples**
1.1. Biomedical & Procedure	Evidence of defining a phenotype when biomedical and procedure entities co-occur with phenotyping definition cues.	“*dyslipidemia* was *defined* as total cholesterol greater than 220 mg/dl…” (PMID:20819866). This sentence provides an evidence of defining a disorder called dyslipidemia. The association of the disorder term with the word “defined” satisfies this dimension.
1.2. Standard Codes	Evidence of using standard terminologies that are commonly used in clinical setting. Examples of these standard coding classifications and/or terminologies are ICD-9/10, SNOMED CT, and CPT codes.	“a primary or any secondary discharge diagnosis (International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] code) of myoglobinuria (791.3)” (PMID:15572716) provides an evidence of the use of ICD-9-CM code.
1.3. Medications	Evidence of the use of medication for defining a phenotype.	“the use of a lipid-lowering medication” (PMID 20819866).
1.4. Laboratories	Evidence of using quantitative values reflecting clinical measurable values (i.e. laboratory tests values, vital values, procedures, clinical).	“Dyslipidemia was defined as total cholesterol greater than 220 mg/dl”. (PMID:20819866) reported the use of “total cholesterol” test, and the value that the study used to define Dyslipidemia.
1.5. Use of Natural Language Processing (NLP)	Evidence of NLP use accompanied with any of the following entities: biomedical, procedure, and/or medications.	“Example of a Clinical Note Represented as a “Bag of Words” Note ID 45893484-02 34695234-01 HF status positive negative Covariate #1 "heart" 3 1 Covariate #2 "pulmonary"“ (PMID:17567225)
2. Intermediate category (n = 2)		
2.1. Data entities	Evidence of information relevant to data entities used in study or phenotyping definitions. Some examples when describing a database used, clinical data, and/o electronic health records (EHR).	“Computerized medical and pharmacy records were reviewed” (PMID:11388131).

2.2. Study design or IRB	Evidence of information about study design or the IRB. For example, an evidence of the method used as “Gold standard”.	“STUDY DESIGN: Retrospective chart review.” (PMID: 11388131).
3. Exclusion category (n = 3)		
3.1. Exclusion 1– Irrelative evidence: 3.1.1. Location 3.1.2. Ethical 3.1.3. Financial 3.1.4. Patient direct contact 3.1.5. Provider or researchers (excluding patients) 3.1.6. Performance 3.1.7. Quality of Care	Evidence of information about other study methodological details that is not supportive for defining a phenotype directly.	“All patients were members of the managed care system and incurred a significant financial advantage from having their prescriptions filled within the system.” (PMID16765240) – (Sub-dimension: Financial) Note: additional examples in the annotation guidelines in the appendix
3.2. Exclusion 2- Computational and statistical evidence: 3.2.1. Alerts 3.2.2. Software 3.2.3. Statistics	Evidence of computational or statistical information that is not supportive for phenotyping definitions.	“We used logistic regression models with generalized estimating equations to adjust for race, year, race x year interactions, age, and sex.” (PMID16567608)) – (Sub-dimension: Statistics)
3.3. Exclusion 3- Insufficient evidence: 3.3.1. Insufficient evidence	Sentences that do not show any evidence in any of the nine dimensions.	“As reported previously, administratively-assigned race/ethnicity is highly concordant with genetic ancestry for European and African Americans” (PMID:28222112)

Table 3 Level of evidence of a sentence with a phenotyping definition (Rule-based final decisions)

Rule	Rule description	Level of evidence	Final Decision	Number of Sentences
R1	If INC= True and ITC= False and EXC = False	The sentence shows **strong evidence** of a phenotyping definition.	Positive	1222 (30.77%)
R2	If INC= True and ITC= True and EXC = False	The sentence shows **strong evidence** of a phenotyping definition.		
R3	If INC= True and ITC= True and EXC = True	The sentence shows **strong intermediate evidence** of a phenotyping definition due to the presence of any of the Exclusion criteria.	INTERMEDIATE_I	701 (17.65%)
R4	If INC= True and ITC= False and EXC = True	The sentence shows **strong intermediate evidence** of a phenotyping definition due to the presence of any of the Exclusion criteria.		
R5	If INC= False and ITC= True and EXC = False	The sentence shows **weak intermediate evidence** of a phenotyping definition due to the absence of any of the Inclusion criteria, but presence of any of the intermediate criteria.	INTERMEDIATE_II	914 (23.01%)
R6	If INC= False and ITC= True and EXC = True	The sentence shows **weak intermediate evidence** of a phenotyping definition due to the absence of any of the Inclusion criteria, but the presence of any of the intermediate criteria.		
R7	If INC= False and ITC= False and EXC = True	The sentence shows **no evidence** of a phenotyping definition.	Negative	1134 (28.55%)

2.3.4 Annotation process

In order to produce a high-quality corpus, it is recommended that the corpus is annotated by more than one annotator (Artstein, 2017). Here, two annotators with a biomedical informatics background (Samar Binkheder, M.S., Heng-Yi Wu, PhD) carried out the annotation process. Both annotators have degrees in biomedical informatics, are familiar with the medical standard terminologies, and are familiar with text-mining. They designed the annotation guidelines iteratively through several meetings and manual analysis of textual patterns of a phenotype definition. When both annotators were satisfied with the final version of the annotation guidelines, they started the annotation of the corpus. For each dimension of the ten dimensions (Table 2), if the dimension is present, the annotator annotates it as 1, otherwise, it is 0. The development of an annotation guidelines is critical to ensure the consistency and quality of the annotations. The process can start by a draft, and then refined iteratively until final draft is satisfied (H. Gurulingappa et al., 2012). During the guideline's development process, subsets of the corpus were annotated until the annotators were satisfied with the guidelines. After that, the full corpus was annotated. The process is shown in Figure 3 which was inspired by (H. Gurulingappa et al., 2012).

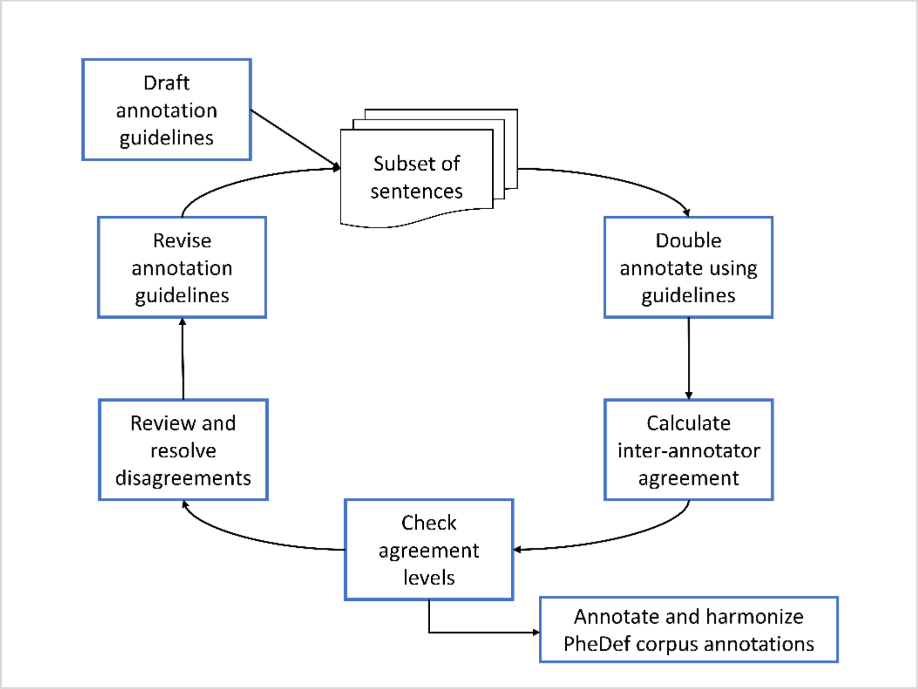

Draft annotation guidelines
Subset of sentences
Double annotate using guidelines
Calculate inter-annotator agreement
Check agreement levels
Annotate and harmonize PheDef corpus annotations
Review and resolve disagreements
Revise annotation guidelines

Figure 3 Iteration process of developing the annotation guidelines and the final annotation

After finalizing the guidelines, both annotators annotated all sentences of the corpus following the final proposed annotation guidelines. The annotation process was divided into several rounds starting from annotation of subset of sentences 400 (first round). After that, the number of sentences for each round were 1000, 1300, and 2700. After each annotation round, there were “consensus sessions” that each took around 1-4 hours where annotators discussed and resolved any disagreements. Moreover, a third Ph.D. annotator addressed disagreements in annotations between annotators if they did not achieve a consensus. The goal was to identify areas of disagreements as well as areas to achieve our 100% gold standard.

2.3.5 Inter-annotator agreement (IAA)

The inter-annotator agreement is to assess the reliability of the annotations. There are several benefits for the manual annotation by multiple people, such as to generate correct annotations, validate and improve the scheme guidelines, resolve ambiguities in data, and evaluate valid interpretations (Artstein, 2017). Further, the written annotations guidelines scheme help in generating consistent and reproducible annotations (Artstein, 2017). Therefore, to measure the agreement between annotators, we used three measures of agreement: percent agreement, overall percent agreement (Wilbur, Rzhetsky, & Shatkay, 2006), and Cohen's kappa (McHugh, 2012). These measures vary in their approaches, but they all aim at producing the best possible reliable and correct annotations as there is no reference for the annotation of some of the sources (Artstein, 2017). The percent agreement and Cohen's kappa (McHugh, 2012) were calculated for each dimension using R packages ('irr'[1] for percent agreement and 'fmsb'[2] for kappa). For example, if the two annotators annotate a dimension as 1, it means an agreement. On the other hand, if one annotator annotates a dimension as 1 and the other as 0, it means disagreement. The overall percent agreement (Wilbur et al., 2006) was calculated over the ten (10) dimension on a sentence-level (Table 2) as the following:

$$\text{Overall sentence level agreement} = \frac{(\#Sentences \times 10) - \#\ disgreement}{(\#\ Sentences\ \times 10)} \times 100$$

2.4 Results

2.4.1 Corpus description

PubTator[3], a web-based tool for annotating biomedical entities, including diseases, genes, mutations, and chemical (C.-H. Wei, Kao, & Lu, 2012). We uploaded our PMID list (n = 86) and run the annotation analysis. Table 4 presents the results from PubTator for the disease terms that were found in more than one abstracts. Disease terms that appeared in single abstracts and terms for other entities (genes, mutations, and chemical) are shown in Appendix 5.

[1] https://cran.r-project.org/web/packages/irr/irr.pdf

[2] https://cran.r-project.org/web/packages/fmsb/fmsb.pdf

[3] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

Table 4 Phenotypes appeared in more than one abstract in our corpus

Term	Number of abstracts
Diabetes	16
Hypertension	11
Diabetes mellitus	8
Heart failure	7
Asthma	3
Bleeding	3
Cancer	3
Coronary heart disease	3
Diabetic	3
Hypertensive	3
Obesity	3
Osteoarthritis	3
Pneumonia	3
Type 2 diabetes	3
Acute renal failure	2
Allergies	2
Death	2
Dementia	2
Gout	2
Myocardial infarction	2
Pulmonary embolism	2
Rhabdomyolysis	2
Rheumatoid arthritis	2
Right bundle branch block	2
Sepsis	2
Stroke	2

We annotated the corpus using our annotation guidelines with ten dimensions (Table 2). The total number of sentences in this corpus was 3971 sentences that were extracted from 86 full texts methods sections. Table 5 shows the number of annotated sentences in for each category and dimension. “Biomedical & Procedure” dimension showed the highest number of annotated sentences with around 1449 (36.5%). “Data entities” and “EXC2 – Computational and statistical evidence” were both over thousand annotated sentences with 1370 (34.5%) and 1314 (33.1%), respectively. The number of annotated sentences for “Medications”, “Standard codes”, and “Laboratories” dimensions from inclusion category were 593 (14.9%), 385 (9.7%), and 246 (6.2%). The number of annotated sentences for "Use of NLP" dimension were the lowest with 49 (1.2%).

Table 5 Corpus description and inter-annotator agreement

Category	# of sentences (%) per category	Dimension	# of sentences (%) per dimension	Percent	Kappa	Kappa 95% CI
Inclusion	1923 out of 3971 (48.4%)	Biomedical & Procedure	1449 (36.5%)	95.00%	88.96%	0.87 - 0.90
		Standard codes	385 (9.7%)	99.47%	97.01%	0.95 - 0.98
		Medications	593 (14.9%)	99.09%	96.44%	0.95 - 0.97
		Laboratories	246 (6.2%)	99.70%	97.42%	0.95 - 0.98
		Use of NLP	49 (1.2%)	99.65%	83.54%	0.74 - 0.92
Intermediate	1851 out of 3971 (46.6%)	Data entities	1370 (34.5%)	96.71%	92.59%	0.91 - 0.93
		Study design and/or IRB	780 (19.6%)	98.00%	93.56%	0.92 - 0.94
Exclusion	2273 out of 3971 (57.3%)	EXC1 – Irrelative evidence	733 (18.4%)	97.27%	91.05%	0.89 - 0.92
		EXC2 – Computational and statistical evidence	1314 (33.1%)	96.84%	92.83%	0.91 - 0.94
		EXC 3 – Insufficient evidence	359 (9.0%)	95.96%	78.72%	0.75 - 0.82

Table 3 shows the rule-based final decisions which are "Positive", "Intermediate I", "Intermediate II", and "Negative". The positive indicated the highest level of evidence of defining a phenotype while the negative indicated no evidence of defining a phenotype. The number of sentences with "Positive" are 1222 (30.77%). "Intermediate I" is the sentences that showed strong intermediate evidence were 701 (17.65%) sentences of the corpus. "Intermediate II" are the sentences that showed weak intermediate evidence were 914 (23.01%) sentences of the corpus. Finally, the number of negative sentences represented in our corpus was 1134 (28.55%) sentences.

2.4.2 Inter-annotator agreement

For inter-annotator agreement, the calculations were based on annotation of each dimension (Table 2 & Table 5). We used the overall sentence-level percent agreement (inspired by Wilbur et al. (Wilbur et al., 2006)), percent agreement, and Kappa. The overall sentence-level percent agreement was high with 97.8%. The percent agreement and kappa measures results are shown in Table 5. Generally, all dimensions showed high agreement on both percent agreement and kappa. For the dimensions of the inclusion category, the "Biomedical & Procedure" showed around 95% percent agreement, and almost perfect kappa with 88.96%. For the "Standard codes", "Medications", and "Laboratories" dimensions, they all showed over 99% percent agreement and over 96% kappa. For the "Use of NLP" dimension, it showed over 99% percent agreement and 83.54% kappa. For the dimensions of the intermediate category, they showed high agreement on percent agreement with over 96%, and kappa with over 92%. Finally, for the dimensions of the exclusion category, both "EXC1 – Irrelative evidence" and "EXC2 – Computational and statistical evidence" showed high agreement on percent agreement with 97.27% and 96.84%, and kappa with 91.05% and 92.83%, respectively. The "EXC 3 – Insufficient evidence" dimension showed high percent agreement (95.96%) and substantial kappa (78.72%).

2.5 Discussion

In this work, our goal was to develop an annotation approach and an annotated corpus that is capable of supporting future text-mining tasks such a literature-based discovery of phenotyping case definitions. In terms of selection of phenotypes, we chose to select a set of phenotypes based on our group research interests, which were mostly ADEs (n = 279). We utilized these phenotypes to search the literature for abstracts and we included 86 abstracts to build the sentence-level corpus from their full texts' methods sections. Annotation approaches were based on evaluating the presence of our proposed ten dimensions in a sentence (Table 2) and the final decisions were derived based on a set of seven rules (Table 3). Our focus in annotating the corpus is to develop a generalized approach to capture contextual features of phenotyping rather than focusing on specific entities. The two annotators worked in developing the annotation guidelines iteratively; after finalizing the guidelines, the whole corpus was annotated. For inter-annotator agreement, we used three measures for evaluation: overall sentence percent agreement (inspired by Wilbur et al. (Wilbur et al., 2006)), percent, and kappa agreement. Overall, the results for the inter-annotator agreement were high and the overall sentence-level percent agreement was high with 97.8%. One observation with the "EXC 3 – Insufficient evidence" dimension showed "substantial agreement" (see Table 2 for interpretation of Kappa in

(Viera & Garrett, 2005)) that was the lowest kappa score among all dimensions. This dimension indicates sentences with lack of evidence in any of the other nine dimensions. Overall, we annotated 3971 sentences extracted from methods sections of 86 articles and the inter-annotator agreement showed that the annotations and guidelines are valid.

Annotating a larger number of articles might generate more contextual patterns of a phenotyping definition in EHR-based studies. However, we also believe that we have a comprehensive coverage for several study types of studies. Here we report the study design terms as they appeared in our corpus and it here as it appears it the text:

- Observational Study
- Longitudinal study
- Cohort Study (retrospective cohort, prospective cohort, Nonrandomized retrospective cohort study)
- Case-Control Study
- Retrospective Study (retrospective cohort, nonexperimental retrospective, Nonrandomized retrospective cohort study, retrospective validation)
- Cross-sectional Study
- Comparative Study
- Descriptive Study
- Validation Study
- Prospective Study (prospective cohort study)
- Genome-Wide Association Study
- Epidemiology and/or Surveillance Study
- Follow-up Study

With the multi-study coverage, we believe that our corpus was sufficient to capture wide range of contextual cues representing a phenotyping case definition in the biomedical literature.

2.5.1 Sentence-level annotation and dimensions selection

Our decision in this work is to focus on the sentence-level annotations rather than entity-level annotations. There are several reasons for this decision. First, we believe that a phenotyping definition is best represented as full sentences rather than single concepts or terms. Entity-level annotations can be accomplished in future steps with the goal of text summarization. Second, we aimed to utilize a generalizable approach that serves as a foundational basis for annotating a phenotyping definition. The selection of ten proposed dimensions (Table 2) was based on identifying phenotyping definition contextual cues that were observed in published literature (Botsis & Ball, 2013; Kirby et al., 2016; Shivade et al., 2014; Yadav et al., 2018) as well as during our manual annotation process (Figure 3). Third, based on our analysis, contextual cues of a phenotyping definition are not only reliant only on biomedical concepts, but also it can be extended to other cues, such as “defined”, “inclusion criteria”, “exclusion criteria”, and “eligibility”. To our knowledge, contextual cues of phenotyping definitions in the literature that surround biomedical and medication entities were not studied previously.

2.5.2 Error analysis

We performed an error analysis on sentences where annotators had disagreements. We found that recognizing abbreviated terms was slightly challenging and it appeared problematic in seven dimensions shown in Table 6. Thus, it can be hard to determine if an abbreviated term is a biomedical, procedure, or medication. For example, the term ICD can mean “implantable Cardiac Defibrillators” or “International Classification of Diseases”. Therefore, we addressed this to the best of our abilities by returning to the full text article. In addition to the abbreviation challenge, we observed that natural human error could also generates some disagreements during the annotation process. For example, one of the annotators missed some keywords that were noticed during the consensus sessions. Such mistakes were not intentionally made. Furthermore, there was an ambiguity in some of the terms that the same term has more than one meaning. In this case, understanding the context around the text is necessary and helped in addressing this problem. Overall, annotating phenotyping definitions’ events e.g. a co-occurrence of more than one keyword, is challenging because they require the presence of more than one pattern. Table 6 provides common errors that led to some of the disagreements with examples.

Table 6 Error analysis of the annotation disagreements

Error	Dimensions	Examples (Sentences)
Abbreviated terms	Biomedical & Procedure	"Events that occurred during follow-up were identified from hospitalization records, and ARIC and CHS study" (PMID25104519)
	Standard codes	"Finally, the Apollo Data Repository provided data for ICDs" (PMID26961369)
	Medications	"‘‘common’’ side effects, e.g. headache, to judge the relevance of side effects associated with AZA." (PMID24177317)
	Use of NLP	"From this cohort, we identified 15,761 patients with HPI” (PMID25567824)
	Data	“Cohort with HPI data” (PMID25567824)
	EXC1 – irrelevant evidence	"190 patients completed the SCID assessment"(PMID25827034)
	EXC2 – Computational and statistical evidence	"The MCMC method" (PMID21931496)
One of the annotators missed keywords or/and criteria	Use of NLP	"The algorithm uses non-negated terms indicative of HF" (PMID17567225)
	Data	"If data on weight and height were available” (PMID21862746)
	EXC1 – irrelevant evidence	- EXC1 – irrelevant evidence (financial): "until termination of insurance coverage." (PMID12952547) - EXC1 – irrelevant evidence (Ethical): "To protect patient confidentiality, all personal identifiers are deleted” (PMID21051745) - EXC1 – irrelevant evidence (Location of the study): "We randomly sampled outpatient clinical encounters from October 1, 2003 through March 31, 2004 at VA Maryland (VAMHCS) and at VA Salt Lake City (VASLCHCS) Health Care systems." (PMID20976281)
	EXC2 – Computational and statistical evidence	"Characteristics were measured during the one-year baseline period (i.e., before time zero)." (PMID20112435)
Without co-occurrence	Use of NLP	"Humedica derives NLP items from text entries that correspond primarily to terms in two large dictionaries, SNOMED and MedDRA" (PMID26725697) NLP terms did not co-occur with biomedical/procedure/medication concept
	Data	"If the first record for a woman was either …" (PMID22071529)

Term ambiguity	Biomedical & procedures events	"Only acute conditions occurring during the first 24 hours of hospital admission were considered." (PMID24734124) The term “condition” by itself can have different meaning not relevant to disease. However, when the word "condition" is not supported with other keyword indicating it is a medical condition.
	Study design or IRB	"The nucleotide reference for this allele is guanine. 4." (PMID26221186) The term “reference” does not indicate gold standard reference.
	EXC2 – Computational and statistical evidence	"More points mean a higher risk of hyperkalemia." (PMID20112435)
Neither Biomedical nor Procedure (e.g. social status)	Biomedical & Procedure	"We created a binary variable for marital status, where “single” included those patients classified as divorced, single, widowed, or separated." (PMID25091637)
Not clear statement of using standard codes	Standard codes	"Outcomes were evaluated by administratively coded data” (PMID26370823)
Assigning terms as Biomedical & Procedure vs. medications (e.g. substances)	Biomedical & Procedure/Medications	"The most recent fasting lipid profile in patients with dyslipidemia and glycosylated hemoglobin level in patients with diabetes” (PMID11388131)
Spelling and short forms	Medications	"Asthma meds refilled regularly." (PMID12952547)
Biomedical/Procedure/Medication terms without supportive definition evidence	Biomedical & Procedure/Medications	"reports KD=9100 for bupropion and KD>10 000 for mirtazapine (vs 200 for nefazodone)." (PMID22466034)
“More than or less than” value, but not directly relevant to phenotyping	Clinical measurable values	"≥2 years of observation before period of interest; n = 50." (PMID23449283)
Adding new keywords for the dimension	EXC2 – Computational and statistical evidence	Example of new keywords describing “EXC2”, are: risk score, inter-rater variability, custom-designed data entry template, predictor variable, Tukey multiple comparison test, Web-accessible, teleconferences, propensity-matched, machine-implementable rule, Illumina Omni1_- QUAD, Illumina 660W, TaqMan, Illumina 660-Quad, and Illumina.

2.5.3 Limitations of the study

This work does not stand without limitations. The manual corpus annotation is time-consuming and labor-intensive. Only two annotators annotated the corpus; therefore, we tested the annotations with more than one measurement of agreement (overall percent, percent, and kappa). Both annotators were familiar with biomedical informatics concepts and text-mining approaches, but we note that some were more challenging than others. The results inter-annotator agreement showed high agreement indicating reliable annotations and guidelines. Generally, more annotators with clinical expertise could assist more during the task of annotations. In addition, automatic entity recognition to recognize biomedical entities can also improve the annotation process and decrease the time of annotation. As mentioned previously, the scope of this work is on capturing patterns of contextual cues surrounding a phenotyping definition.

For "Use of NLP" dimension, we decided to only annotate the presence or absence of NLP in a sentence with the goal to use it as a part of phenotyping. Going beyond this scope would complicate the annotation task, require detailed and full annotation of NLP methodology, and require a bigger corpus. Therefore, the number of sentences of this dimension is comparably lower than other dimensions. In addition, our aim in this work is to establish a foundational approach.

2.5.4 Applications of the corpus

To date, PheKB (Kirby et al., 2016) library provides around 50 definitions only for some phenotypes. A study of best practices for phenotyping of adverse events found that the re-utilization of existing definitions is crucial (Wiley, Moretz, Denny, Peterson, & Bush, 2015). This only works for case definitions that have been already published in the literature. Therefore, this work aimed to support the re-usability of published definitions (R. L. Richesson, Hammond, et al., 2013) by analyzing their contextual cues. Specifically, for using case definitions to establish EHR-based research, such drug safety surveillance. Availability of these definitions can also assist in the validation of them in several institutions to ensure cohort consistency (R. L. Richesson, Rusincovitch, et al., 2013). The ten dimensions in our annotation guidelines provide a foundational understanding of the basic contextual cues that represent a phenotyping case definition in the literature.

Therefore, we believe that this corpus can serve as a baseline for developing either automatic or manual approaches to annotate a larger corpus size and advancing our proposed guidelines. Furthermore, our main aim of developing this corpus is to use it for text-mining applications to automate mining of phenotyping definitions publish in the literature.

In conclusion, clinical research, such as drug discovery, is moving toward the use of EHRs that provides information about patient's variations, including comorbidities and co-medications. The corpus and annotation guidelines can serve as a foundational informatics approach for annotating and mining literature-based phenotyping definitions. Ten dimensions are proposed characterizing major contextual patterns and cues of a phenotyping definition in published literature. This is a step towards research to advance leveraging of phenotyping definitions from literature to support EHR-based phenotyping studies.

CHAPTER THREE: AN AUTOMATED TEXT MINING APPROACH OF PHENOTYPING DEFINITIONS IN THE BIOMEDICAL LITERATURE

In Chapter 2, we proposed an approach to annotate phenotyping definitions in published literature. In addition, we used our proposed annotation guidelines to create a corpus of annotated sentences from full texts. The main motivation behind developing the corpus is to develop a text-mining technology. In this chapter, we build an information retrieval (IR) and extraction (IE) systems for facilitating the use of published literature-based phenotyping definitions. In addition, we applied these systems on a large-scale literature. Similar to Chapter 2, we are using adverse drug reactions (ADEs) as our phenotypes of interest that can be used in many tasks, such as building lexica and dictionary. The final product of this chapter is a large collection of phenotype definitions-related abstracts and sentences. We note that our used approaches for mining ADE phenotype definitions-related abstracts and sentences can be generalized to other phenotypes and are not limited to ADEs.

3.1 Introduction

A major public problem is that many drug side effects appears in public, including deaths and hospitalizations, after the release of the drug to the market. These side effects reported to reach millions, where "5% hospital admissions, 28% emergency visits, and 5% hospital deaths" (Sarker & Gonzalez, 2015). Furthermore, the estimated cost is about seventy-five billion dollars yearly (Sarker & Gonzalez, 2015). There are several sources that have been used to conduct post-marketing ADE-based research, such as spontaneous reporting systems, electronic health records (EHRs), social media, and biomedical literature (Davazdahemami & Delen, 2018). Research shows that FDA Adverse Event Reporting System (FAERS) data has limitations where it either underestimates or overestimates some ADEs (Sarker & Gonzalez, 2015). On the other hand, repurposing of EHR for pharmacovigilance and clinical research (Newton et al., 2013) has increased where a number of approaches can be used for phenotyping (Banda, Callahan, et al., 2016; Newton et al., 2013; X. Wang, Hripcsak, Markatou, & Friedman, 2009). For example, several studies have used EHR for ADE signal detection and used literature for its validation (Iyer, Harpaz, LePendu, Bauer-Mehren, & Shah, 2014). The use of EHR provides several advantages, including “large scale, reduced cost, repeated observations, and the ability to observe rare events” (Mo et al., 2015). Furthermore, EHR for secondary use purposes mining is important because it offers a rich resource of accumulated clinical & patient’s data on variable disease levels, it provides opportunities for analyzing ADEs, and it supports answering of clinical research questions (Chiu & Hripcsak, 2017; Yadav et al., 2018). Examples of patient’s data in EHRs that are collected routinely in clinical practice are diagnoses, laboratory tests, billing records, medications, and medical history which can be either in structured (e.g. ICD9 codes) or unstructured format (e.g. clinical notes) (Chiu & Hripcsak, 2017). In contrast, EHR use generates new challenges.

The use of EHR does not stand without challenges. These challenges have opened new opportunities for informatics research. One of the challenges of EHR-based research is to accurately find cases and controls for a phenotype of interest (Carroll et al., 2011). This is called as cohort identification that has been widely used for various clinical and biomedical studies (Yadav et al., 2018). Cohort identification is an obstacle especially when phenotyping definitions are not readily available for performing clinical research studies (D. Li et al., 2012). Therefore, we have identified a gap, which is the absence of phenotyping definitions for some phenotypes of interest or sources that support its development.

To generate or obtain a phenotyping definition, there are several approaches: low-throughput or high-throughput approaches (R. L. Richesson et al., 2016). First, the low-throughput phenotyping is highly reliant on expert domain knowledge and rule-based algorithms, such as decision trees and boolean logic (R. L. Richesson et al., 2016). These methods for cohort identification tend to be time-consuming and labor-intensive (D. Li et al., 2012; Park & Choi, 2014) due to the need of an expert involvement. A multidisciplinary team works on developing and designing a phenotyping definition in which manual review, multiple iterations, and validation are needed (Carroll et al., 2011). For example, generation of a new phenotyping definition, especially when it is derived based on the EHR data of that institution, does not mean it is portable across other institutions. Therefore, a validation step of a phenotyping definition across multiple institutions is important to ensure that it is performing well across different populations (Liao, Ananthakrishnan, et al., 2015; Overby et al., 2013). Electronic Medical Records and Genomics (eMERGE) network has an effort to manually create, disseminate, and validate phenotyping definitions that they made them publicly available in Phenotype KnowledgeBase (PheKB)[1] (Newton et al., 2013; R. L. Richesson et al., 2016). However, these are still lacking standardized representations. For example, these definitions are stored in Microsoft® Word, Excel files, or other formats with no specific template for easing human interpretation (Chute et al., 2011). Another challenge is the development of a phenotyping definition for clinical notes that requires knowledge in Natural Language Processing (NLP) and human involvement (Park & Choi, 2014). Therefore, the development of expert-driven phenotyping definitions process is still very challenging, labor intensive, error-prone, and time-consuming (Lasko et al., 2013; Park & Choi, 2014; Xu et al., 2015).

Second, recent efforts (V. Agarwal et al., 2016; Banda, Callahan, et al., 2016; Halpern, Choi, Horng, & Sontag, 2014) are moving toward high-throughput phenotyping that uses statistical, machine learning, and data-driven approaches (Halpern et al., 2014; R. L. Richesson et al., 2016). Unlike low-throughput phenotyping that can be time-consuming and require high-effort, high-throughput phenotyping can be scalable to high-dimensional adverse events (V. Agarwal et al., 2016; Halpern et al., 2014; R. L. Richesson et al., 2016). However, it require multiple sources to support its scalability (Zhang et al., 2018). Some efforts suggest the use of machine-learning approaches to automate the development of a phenotyping definition using EHR data (Lasko et al., 2013). Such definitions are developed on specific populations in which generalization of models and algorithms can be infeasible due to EHR natural challenges. Moreover, EHR data can be sparse across patient data. EHR data usually reflects patient who are very ill, which generates bias (Castro et al., 2014; Malinowski et al., 2014; W. Q. Wei et al., 2016). Moreover, EHR data can be inconsistent (Castro et al., 2014; Frey, Lenert, & Lopez-Campos, 2014; Malinowski et al., 2014; W. Q. Wei et al., 2016), incomplete (Frey et al., 2014; Pathak et al., 2013; W. Q. Wei et al., 2016), fragmented, (Daniel & Choquet, 2014; W. Q. Wei et al., 2016), inaccurate, complex (Daniel & Choquet, 2014; Frey et al., 2014; Pathak et al., 2013), formatted in free text, from unknown sources, and variable in granularity (Daniel & Choquet, 2014). A challenge

[1] http://www.PheKB.org

is the variability across institutions in EHR data, which generate problem when creating or applying phenotyping definitions. For example, each institution might have its own usage of ICD-9 codes and drugs' brand names. This can affect the query of the definition that if generated in one institution would not work in another institution and will generate variable and inconsistent results (Chute et al., 2011). A study found that only half of the evaluated tools can be used portable among other EHRs that are different from where the phenotyping definition were originally developed (Xu et al., 2015). Therefore, these common issues can be problematic, especially when phenotyping definitions are derived from it.

Several systematic reviews have been performed to harmonize, compare, and validate phenotyping definitions in the literature. These studies (Claire Barber et al., 2013; Fiest et al., 2014; Leong et al., 2013; Lui & Rudmik, 2015; Macdonald et al., 2016; Pace et al., 2017; Souri et al., 2017) have systematically reviewed several case definitions for a number of phenotypes, including ADEs. Their goal was to validate or to compare performance of different case definitions. Moreover, these studies reported several reasons for performing these systematic reviews, such as the lack of widely used or validated definitions (Claire Barber et al., 2013; Fiest et al., 2014; Leong et al., 2013; Lui & Rudmik, 2015; Macdonald et al., 2016; Pace et al., 2017; Souri et al., 2017), and the need to improve reproducibility of observational studies (Fox et al., 2013). In addition to these efforts, the Observational Medical Outcomes Partnership (OMOP) (Fox et al., 2013), which is called today the Observational Health Data Sciences and Informatics (OHDSI), has developed a library source based on systematic literature review of a number of health outcomes of interest (HOIs) definitions for ADEs. Even though these efforts are very valuable in harmonization and validation of phenotyping definitions, the process of searching literature systematically for evidence-based phenotyping definitions lacks scalability, and can be difficult, slow, and time-consuming.

To summarize, we have identified several gaps related to the development of a phenotyping case definitions: (1) The lack of phenotyping definitions for several phenotypes; (2) The current approaches are labor-intensive and not scalable; (3) The need for high-throughput phenotyping with minimum expert involvement; and (4) The need of utilizing large-scale literature for knowledge discovery of phenotyping definitions. Therefore, in this study, we identified biomedical literature as a potential resource for text-mining, automated retrieval & extraction, and knowledge discovery of phenotyping definitions.

3.2 Background

3.2.1 EHR phenotyping

Large-scale EHR has become an enriched resource for secondary use research. In the United States, office-based physicians adoption of any EHRs (i.e. all or partially electronic records) has increased from 42% in 2008 to 87% in 2015 (C. Barber et al., 2013). The increase of EHR adoptions has led to an increase in EHRs longitudinal data providing new efficient and cost-effective resources for biomedical and clinical research. A number of other advantages of using EHRs data for research, such as big data, variety of data types, diverse populations, and real-world patterns of phenotypes. Moreover, EHRs enable new discoveries and hypothesis generations in areas like drug-adverse effect associations, phenotype-genetic associations, phenotype-disease associations, and comparing effectiveness of established therapies (Castro et al., 2014). Declerck et al. (Declerck et al., 2015) hypothesized that using EHR data can support the drug-related adverse events discovery. Examples of EHR datatypes, are demographics, drug history, symptoms, and laboratory tests (Declerck et al., 2015). Large number of these EHR based studies are already published in the literature.

As we introduced in Chapter 2 and this Chapter, the use of EHR requires an identification of cohort for a desired population with a data-driven approach called EHR phenotyping (Lasko et al., 2013; Park & Choi, 2014). EHR phenotyping is the process that involves the design, implementation, and execution of phenotyping algorithms for a phenotype of interest as well as the analysis of the queried results (Peterson & Pathak, 2014). Furthermore, EHR phenotyping process includes engineering, identifying, quantifying, and automating cohort and phenotype selection in EHR. This process is primarily achieved by using EHR-driven data (Frey et al., 2014; Glueck et al., 2016; Lasko et al., 2013). EHRs data is large by nature, phenotyping process involves dealing with massive amount of practice-based daily routine clinical data, such as clinical narratives, billing codes, and medications, and patient-generated data that both can be imperfect (Roden & Denny, 2016; W. Q. Wei et al., 2016). Moreover, data within the EHR can be structured, such as coded data (e.g. Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), and International classification of Diseases (ICD)) that can be used for billings and diagnosis (N. Alnazzawi et al., 2015; Ho et al., 2014). On the other hand, EHR data can be unstructured providing detailed information about clinical setting findings, vitals, symptoms, diagnosis, and signs (N. Alnazzawi et al., 2015), such as discharge summaries, radiology reports, and progress notes. However, with respect of phenotyping, there is a trade-off between the use of structured and unstructured data. Structured data can miss tremendous amount of clinical information about the patient, but it can be more interoperable, and machine-readable. On contrary, unstructured data is enriched with detailed clinical information that derived more knowledge about diseases, but it is more difficult to manipulate, and it needs new computational approaches.

3.2.2 Standardized terminologies for EHR phenotyping and literature mining

One of the biggest challenges in EHR secondary use is data interoperability. In fact, efforts of developing phenotyping definitions are known to lack standardization and portability (Fort, Wilcox, & Weng, 2014; Simonett et al., 2015). As a result, inconsistency creates a difficulty in using these definitions across different EHR systems (Declerck et al., 2015). In this section, a description of the common standardized terminologies is provided. There are many standard terminologies that are commonly used for different clinical purposes, such as Medical Dictionary for Regulatory Activities (MedDRA) for adverse events, and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) for clinical representation. Since we are using ADE as our example for phenotypes of interest, it is recommended to combine SNOMED CT, which is the most comprehensive terminology for clinical use, with MedDRA, which is used for adverse events but is not commonly used in clinical practice (Declerck et al., 2015). For literature uses, Medical Subject Headings (MeSH) terms were developed to index biomedical literature (S. T. Wu et al., 2012) which has been integrated in Merged disease vocabulary (MEDIC) (Davis, Wiegers, Rosenstein, & Mattingly, 2012). These terminologies are not only supportive for EHR phenotyping, but also for literature phenotyping and mining. Therefore, our dictionary integrates the above mentioned terminologies that are combined as a one dictionary to serve our text mining tasks. Here, we provide a brief description for each of these terminologies:

1. Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al., 1999) is the international medical terminology developed under International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). MedDRA has been widely used in classifying adverse events in clinical trials and event reporting systems (Reich, Ryan, Stang, & Rocca, 2012). MedDRA is characterized by its five levels hierarchy, which are System Organ Class (SOC), High Level Group Term (HLGT), High Level Term (HLT), Preferred Terms (PT), and Lowest Level Term (LLT). Furthermore, MedDRA covers pharmaceutical regulatory affairs terms, such as diagnoses, drug reactions, signs and symptoms, and procedures. Some of the advantages of using MedDRA is its completeness, accuracy, and flexibility (Brown et al., 1999).
2. Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is maintained by the International Health Terminology Standards Organization (IHTSDO). IHTSDO is a non-profit organization that owns SNOMED CT and was founded in 2007. SNOMED CT is considered as the most comprehensive health terminology in the world for clinical documentation in EHR. One of the features of SNOMED CT is that it can be mapped to other terminologies, such as ICD9 and ICD 10 codes. In fact, SNOMED CT is the largest resource that was developed specifically for clinical use (S. T. Wu et al., 2012). In addition, it supports data interoperability in healthcare settings. Therefore, mapping data from literature to SNOMED CT terminologies can be support for text mining tasks, particularly Named entity recognition (NER).
3. International classification of diseases (ICD) is the standard for classifying diseases and conditions for clinical care use. ICD is the official coding system for coding procedures and diagnosis in the United States. ICD-9-CM is based on the World Health Organization’s Ninth Revision, International Classification of Diseases (ICD-9). For our dictionary, we will incorporate the available ICD-9 procedure terms to combine them with SNOMED CT procedure terms.

4. Comparative Toxicogenomics Database (CTD) is a publicly available database that provides manually curated information about diseases, genes, and chemical. Merged disease vocabulary (MEDIC) (Davis et al., 2012) is a subset of diseases from the U.S. National Library of Medicine's Medical Subject Headings (MeSH) (Lipscomb, 2000), and a subset of genetic disorders from the Online Mendelian Inheritance in Man® (OMIM) database. MeSH vocabulary has been used to index MEDLINE/PubMed articles. On the other hand, OMIM has links of its diseases to many resources, such as MEDLINE. We believe that using MEDIC in our dictionary would enhance its coverage.

3.2.3 Biomedical literature text mining

Text mining was successful in several applications, such as protein-protein interaction, bio-entity tagging, normalization, and term extraction (Krallinger, Valencia, & Hirschman, 2008). Most of knowledge that requires analysis is represented in text. This knowledge provides a rich resource of scientific information (Fleuren & Alkema, 2015) that is mostly found within biomedical literature (Krallinger et al., 2008; Shatkay & Craven, 2012). For instance, PubMed offers over 24 million citations (Fleuren & Alkema, 2015) and is the most accessible database for the biomedical literature with more than 5000 biomedical journals, MeSH indexed, links to full text. Full texts can come in PDF or HTML formats where each of these possess its own challenges (Shatkay & Craven, 2012).

Text mining, literature mining, and text data mining are terms that have been used with the goal of making an effective use of the biomedical text with the utilization of computational tools. Text mining implies mining of valuable information within text. Text mining is not a single-step process, but rather it is a multi-task process which involves, user needs, accessibility to text source, text representation (e.g. PDF and XML), tools, and evaluation (Shatkay & Craven, 2012). Text mining automates the process of discovering and extracting knowledge from unstructured text to represent knowledge in a concise format and to generate hypotheses (Ananiadou et al., 2006; Rebholz-Schuhmann et al., 2012; Spasic et al., 2005). Most of text-mining tools consist of two major steps: information retrieval (retrieve relevant text and documents) and information extraction (extract information and knowledge from text) (Ananiadou et al., 2006; Rebholz-Schuhmann et al., 2012). Examples of the text-mining tasks and their descriptions (Shatkay & Craven, 2012), are:

- The process of segmentation is to segment a document into smaller units, such as sections. Within these units, there are paragraphs and sentences. The paragraphs and/or sentences are further tokenized into smaller segments or tokens that can be either sentences or words. This process possesses some challenges such as the recognition of the end of a sentence. For example, a period can mean the end of a sentence or an abbreviation. In addition, recognizing token boundaries of biomedical text is another challenge. For this, medical dictionaries can be used as a solution for this, but there can be other solutions that are out of our scope. Therefore, the overall objective of the tokenizer is highly reliant on the used application.
- Most of the literature documents are represented as PDF or HTML. Therefore, document conversion into free text is important but the structure of the final converted text is dependent on how the original format was presented. For example, PDF documents are more concerned on how the information look like rather than the structure of information. However, when converting PDF to text, the text might not presented in the correct order or characters. On the other hand, XML is more structured, but the structured can vary across publishers.
- Normalization is one of the tasks. There are several normalization approaches, such as converting all letters to lowercase. Stemming is another approach where we trim the end of the word without the context involvement. On the other hand, lemmatization is a linguistic-oriented approach and it considers parts of speech, morphological rules, and lemmas. The used method/s is highly dependent on the application.
- Chunking is the process of grouping words into phrases.
- Parsing is the process of analyzing a sequence of words (Shatkay & Craven, 2012).

In this work, we proposed building an automated text-mining approach to mine phenotyping definitions-related sentences in the biomedical literature. To achieve this goal, we performed several tasks. First, we build an annotated corpus for abstracts and full-text sentence-level (Chapter 2) and we used it to train and test machine learning algorithms. In these classifiers, we created features utilizing several text-mining approaches based on analyzing patterns and contextual cues of a phenotyping definition. Second, we performed a large-scale information retrieval and extraction using the trained and validated classifiers.

3.3 Methods

We developed a text-mining pipeline for classifying abstracts and full texts on a sentence-level. First, the Abstract-level classifier to retrieve and classify abstracts with relevant content of observational studies. Second, Full-text sentence-level classifier to identify and extract method sections, and to classify positive sentences in full texts with evidence of a phenotyping definition.

3.3.1 Building Lexica and Dictionary

Our research group is primarily interested in adverse drug events (Duke et al., 2012; H. Y. Wu et al., 2017). In Chapter 2, we proposed a list of 279 ADEs as phenotypes for data collection. The primary list of 279 ADE terms (Appendix 1) are represented as MedDRA PT level (Brown et al., 1999). In order to develop a text-mining suite for extracting phenotyping related sentences, we developed a comprehensive terminology that assists in information retrieval (IR) and information extraction (IE) tasks from both literature and medical records. Our aim is to increase the coverage of terms. In addition to MedDRA, a number of terminologies are integrated: SNOMED CT) (Stearns, Price, Spackman, & Wang, 2001), MEDIC (Davis et al., 2012), ICD-9 procedures, and DrugBank (Wishart et al., 2006). For the task of IE NER, we created four dictionaries:

This ADE dictionary is built by mapping our list of ADEs to all synonyms in MedDRA LLT, SNOMED CT, and MEDIC. The most recent version of SNOMED CT terms SNOMED CT was downloaded from Unified Medical Language System (UMLS) and MEDIC data were downloaded from CTD database. Within MEDIC, we used MeSH/OMIM terms, synonyms, and codes. SNOMED CT and MEDIC were mapped to MedDRA concepts (PT & LLT) using exact match method (Table 7). The clinical dictionary includes all clinical concepts excluding data from ADE dictionary for each of MedDRA (PT, LLT), SNOMED CT (diseases and disorders, body structure, clinical finding, clinical event, observable entity, organism, and the situation with explicit context), and MEDIC. The procedure dictionary is for procedures performed within a healthcare setting. For this dictionary, we included procedures from SNOMED CT and ICD-9 procedures. Finally, the drug dictionary is for drug terms from DrugBank[1].

Table 7 Mapping terms to ADE list of 279 phenotypes

Source	Number of mapped terms to ADE	Number of unmapped to ADEs
SNOMED CT	274	5
MEDIC	140	139

3.3.2 Corpus description

In this work, we followed a similar approach to the manual or human-based process of reviewing literature-based medical knowledge by an abstract selection and full texts retrieval (Cohen et al., 2010). For the abstract selection, we manually reviewed abstracts for their relevance to observational-based studies in EHR. PubMed articles were searched for the 279 ADEs. We manually reviewed abstracts and decided on their relevance to observational studies using EHR data (See Chapter 2). The negative abstracts (n=1079) were selected randomly from PubMed foe years between 1995 and 2017. In constructing the full-text sentence-level corpus, a random subset from the positive abstracts in abstract corpus were selected. Their full texts were retrieved, and sentences in the method sections were extracted. More details about the annotation guidelines and performance as well as the annotated dimensions in Chapter 2 (annotation examples were shown in (Binkheder, Wu, Quinney, & Li, 2018) and Table 2). A summary description of the corpus is provided in Table 8.

Table 8 Corpus summary

Corpus	Document type	Class	Number of documents	Total
Abstract-level	Abstract	Positive	799	1878
		Negative	1079	
Full-text sentence-level	Sentence	Positive	1923	3971
		Negative	2048	

[1] https://www.drugbank.ca/

The positive class, for either abstracts or sentence-level, means that they contain information about phenotype definitions. Sentences with a phenotype definition information can contain a description used for defining a phenotype or for building a cohort in an EHR that can include the inclusion and/or exclusion criteria or algorithmic criteria. Within an EHR context, “A phenotype is defined as a biochemical or physical trait of an organism, such as a disease, clinical or physical characteristics, or blood type” (Yadav et al., 2018). Several practices are used for defining phenotypes can be seen within the phenotyping definition descriptions, such as diagnostics terms or codes, clinical characteristics, laboratory tests values, use of medications, risk factors, use of standardized terminologies, and the use of NLP (e.g. list of keywords used) (Chute et al., 2011; R. Richesson et al.; Yadav et al., 2018). In addition, information about data sources (e.g. demographics, vitals, notes, electronic medical records) (Shivade et al., 2014) used in defining the phenotype can be potential for phenotyping, and it can appear in phenotype definitions-related sentences.

On the other hand, the negative class, for either abstracts or sentence-level, means that they do not contain relevant information for phenotyping or defining a phenotype, such as financial information, location of the study, and computational and statistical analyses. An example of a negative sentence,

> “Since nearly everyone residing in the target ZIP code for the current study receives their health care through Marshfield Clinic, this record is considered comprehensive.” (PMID17456828).

3.3.3 Information retrieval: the abstract-level classifier

The abstract classifier is a binary with two categories: positive for abstracts that satisfied the criteria for observational studies, and negative for abstracts that were not (Table 2). The abstract corpus was implemented in Waikato Environment for Knowledge Analysis (WEKA) as string attributes where each contains a title and an abstract. “StringToWordVector” module in WEKA was used to represent each text document as a set of attributes, using the following sub-specifications: “Lowercase tokens”, “wordsToKeep(1000)”,“IteratedLovinsStemmer”, “stopwordsHandler(MultiStopwords)”, "NGramTokenizer (1-3 grams)", "IDFTransform" (Inverse Document Frequency (IDF) Transformation), and "TFTransform" (Term frequency score (TF) Transformation).

After pre-processing of text data and defining these input features, we tested several classification approaches and trained our classifier on the best algorithm, including sequential minimal optimization (SMO) (Platt, 1999), logistic regression (LR) (Quinlan, 2014)), Naïve Bayes (NB) (John & Langley, 1995), and decision trees (C4.5 clone (Lecessie & Vanhouwelingen, 1992) called J48 in WEKA). All of the analyses were performed in WEKA software (Figure 4).

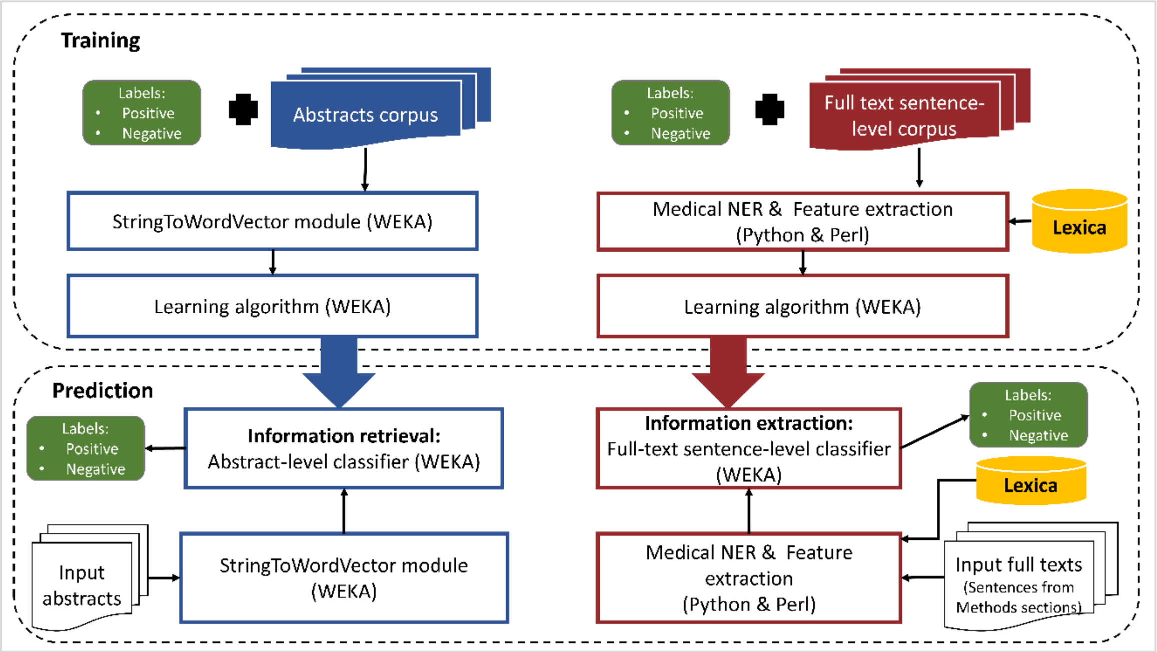

Training
Labels:
Positive
Negative
Abstracts corpus
StringToWordVector module (WEKA)
Learning algorithm (WEKA)
Labels:
Positive
Negative
Full text sentence-level corpus
Medical NER & Feature extraction (Python & Perl)
Lexica
Learning algorithm (WEKA)
Prediction
Labels:
Positive
Negative
Information retrieval:
Abstract-level classifier (WEKA)
Input abstracts
StringToWordVector module (WEKA)
Information extraction:
Full-text sentence-level classifier (WEKA)
Labels:
Positive
Negative
Lexica
Medical NER & Feature extraction (Python & Perl)
Input full texts (Sentences from Methods sections)

Figure 4 Classifiers training and prediction flowchart

NER: Named-entity recognition

3.3.4 Information extraction: the full-text sentence-level classifier

Document representation is a necessary pre-processing step for machine learning to represent text documents as vector of significant terms or patterns (Dalal & Zaveri, 2011; Khan, Baharudin, Lee, & Khan, 2010). There are several approaches that can be used for document representation utilizing different levels of linguistic processing, such as co-occurrence, single term or token, and/or phrase approach (Khan et al., 2010). We used our observations and intuitions to generate features (Kilicoglu, Rosemblat, Malicki, & ter Riet, 2018) that were inspired by features from the annotation guidelines in Chapter 2.

The sentence level classifier identifies phenotyping related sentences from full texts. This classifier is trained using full text sentence level corpus (Table 8). This corpus is a collection of documents (i.e. each document is a sentence from full texts' methods sections. We constructed 339 features from this corpus and converted the corpus into a matrix of numerical attributes: binary (0, 1), count of terms in a sentence, or sum of values multiple attributes (i.e. sum of attribute values for specific set of features). After the text pre-processing and features extraction, we trained the sentence-level full-text on four algorithms, SMO (Platt, 1999), J48 Decision Tree (Quinlan, 2014), Logistic Regression (Lecessie & Vanhouwelingen, 1992), and Naïve Bayes (John & Langley, 1995). All the sentence classifier is trained for binary classification: positive and negative (Figure 4).

Most of the extracted features used for representing each sentence in the corpus were based on Named-Entity Recognition (NER) technology. NER of medical terms for ADE, clinical, procedure, and drug entities (the dictionaries used for this task are shown in Table 10). For ADE entities, the 279 ADE phenotypes (listed in Appendix 1) were mapped to their exactly matched concepts and synonyms in other dictionaries which are Merged disease vocabulary (MEDIC) (Davis et al., 2012) and SNOMED-CT. With this, these ADEs terms and their synonyms were excluded from clinical and procedure entities' dictionaries (Table 10). For clinical entities, SNOMED CT dictionary includes terms for body structure, finding, event, observable entity, organism, and situation. For drug entities, we used DrugBank. Other NER features used for recognizing phenotype definitions' keywords (e.g. "defined as" and "identify"). These keywords were previously identified either manually during the annotation process, or using automated approaches such as n-grams, term frequency (TF-transform), and inverse document frequency (IDF-Transform) (Binkheder et al., 2018). Overall, there we used two ways to represent these features, which are described below.

Single-features refer to single term or pattern representations without rules. Several feature reduction techniques were used, such as word stemming (Dalal & Zaveri, 2011) and regular expression patterns. Regular expressions were used to capture some patterns, such as values of blood pressure, lab, age, height, weight, and body mass index (Appendix

6). NER of medical terms were used to represent single entities as single-features (Table 10 and examples are shown in Appendix 7). In addition, NER of phenotype definitions' keywords is used to represent phenotype definition-related information as one single feature. These definition keywords can be words or phrases, such as “defined", "definition", "classified", "defined as", "identification", "identified", "diagnosis of", "diagnostic criteria", and "case identification". For example, “Controls were patients without *evidence of (definition keyword)* PAD” (PMID20819866). In this sentence example, *evidence of* is recognized as a phenotype definition keyword and it can be represented as a single feature called “definition keywords”.

Compound-features (c-features) refer to the co-occurrence of terms without any order or distance specifications between these terms. This approach was introduced by Figueiredo et al. (Figueiredo et al., 2011) who showed that combining c-features with other e.g. single-features improved the performance of classification. For example,

> “Confirmed adult-onset *asthma (ADE entity)* (AOA) cases were *defined as (definition keywords)* those potential cases with either new-onset *asthma (ADE entity)* or reactivated mild intermittent *asthma (ADE entity)* that had been quiescent for at least one year” (PMID12952547).

In this example, asthma is recognized as an ADE entity, and *defined as* is recognized as a definition keyword. With this, the co-occurrence of ADE entity and definition keywords can be represented as one compound-feature indicates that this sentence has a phenotype definition for asthma. C-features can be also used, for instance, for the co-occurrence of “DRUG” entities with any of the medication-related terms, such as "initiat", "window", "dose", or "value". We hope that some c-features can represent important patterns of sentences in our corpus. Additional examples are shown in Appendix 7.

3.3.5 Classifiers performance evaluation

For each of the abstract-level and full-text sentence-level classifiers, the full corpus was divided into 70% for training and 30% for testing. These algorithms were evaluated using 10-fold cross validation. The training and validation were performed in WEKA. To assess the classifier performance, we used a number of matrices (Zaki, Meira Jr, & Meira, 2014), which are:

1. Recall (Sensitivity) The proportion of the correct predictions for the positive class, which also called as the true positive rate (TPR) or the recall for the positive class.

$$\text{TPR}=\frac{\text{TP}}{\text{TP+FN}}$$

2. Precision (Positive predictive value) is defined as

$$\text{FPR}=\frac{\text{FP}}{\text{FP+TN}}=1\text{-specificity}$$

3. F-measure The F-measure is the trade-off between precision and recall where the higher the F-measure value indicates the better the classifier.

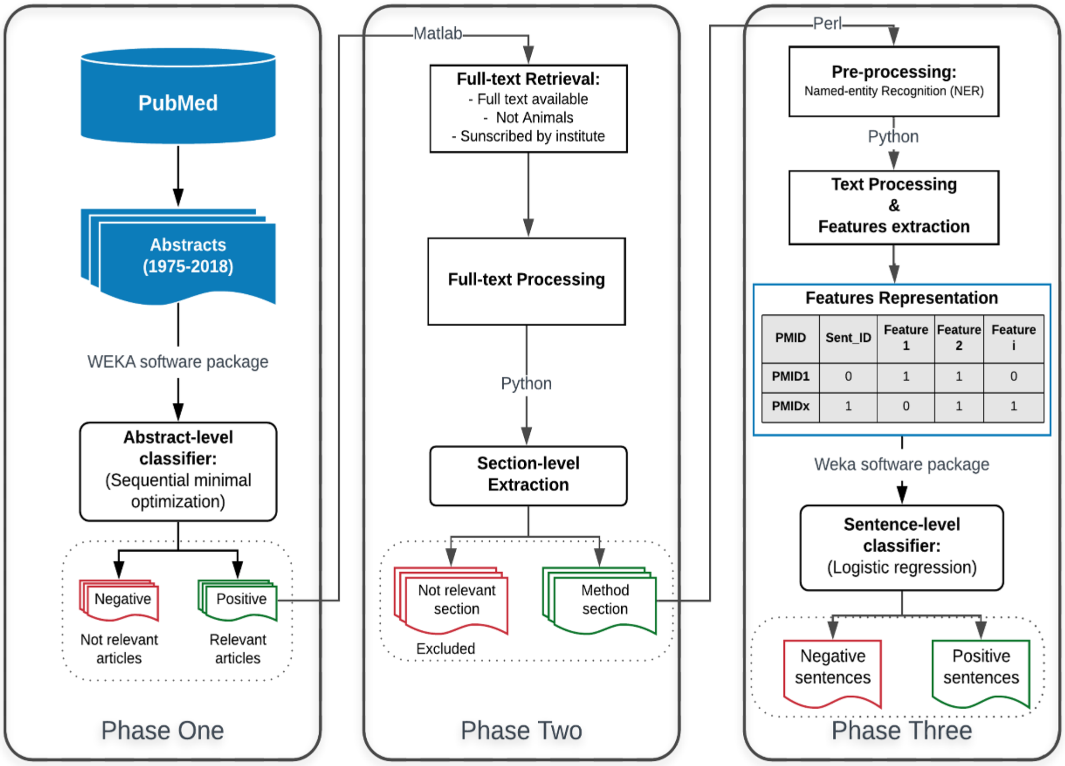
$$\text{F}=2\times\frac{\text{precision}\times\text{recall}}{\text{precision+recall}}$$

3.3.6 Large-scale literature screening

The large-scale screening of PubMed database is summarized into three major phases: large-scale screening of abstracts, full text data pre-processing, and large-scale screening of full-text sentence-level (Figure 5 & Figure 6).

3.3.6.1 Phase 1—Large-scale screening of abstracts

In this phase, we downloaded abstracts from PubMed database for years 1975-2018 (1st Quarter). We selected the machine-learning algorithm with the highest performance (Table 8) in Weka software package. Using abstract-level classifier, PubMed abstracts that were classified as positive are further processed for the large-scale full text screening.



PMID	Sent_ID	Feature 1	Feature 2	Feature i
PMID1	0	1	1	0
PMIDx	1	0	1	1

Figure 5 Flowchart for large-scale data processing

3.3.6.2 Phase 2—Full text data pre-processing

After the abstract-level retrieval, we retrieved full text articles in PDF or XML format. These files were pre-processed by converting them into text and then into GENIA format. The steps are as the following:

1. Retrieve positive full text PDF and XML documents. Using our positive set of abstract PMIDs, their PDF and XML documents were downloaded from PubMed repository if they are open access articles or from the subscribed publisher by our institute. We excluded abstracts that were not human studies.
2. Convert PDF format to text format. After retrieval of the PDF documents, they were converted to text format with pdftotext[1] tool. In addition, sentences are tokenized and

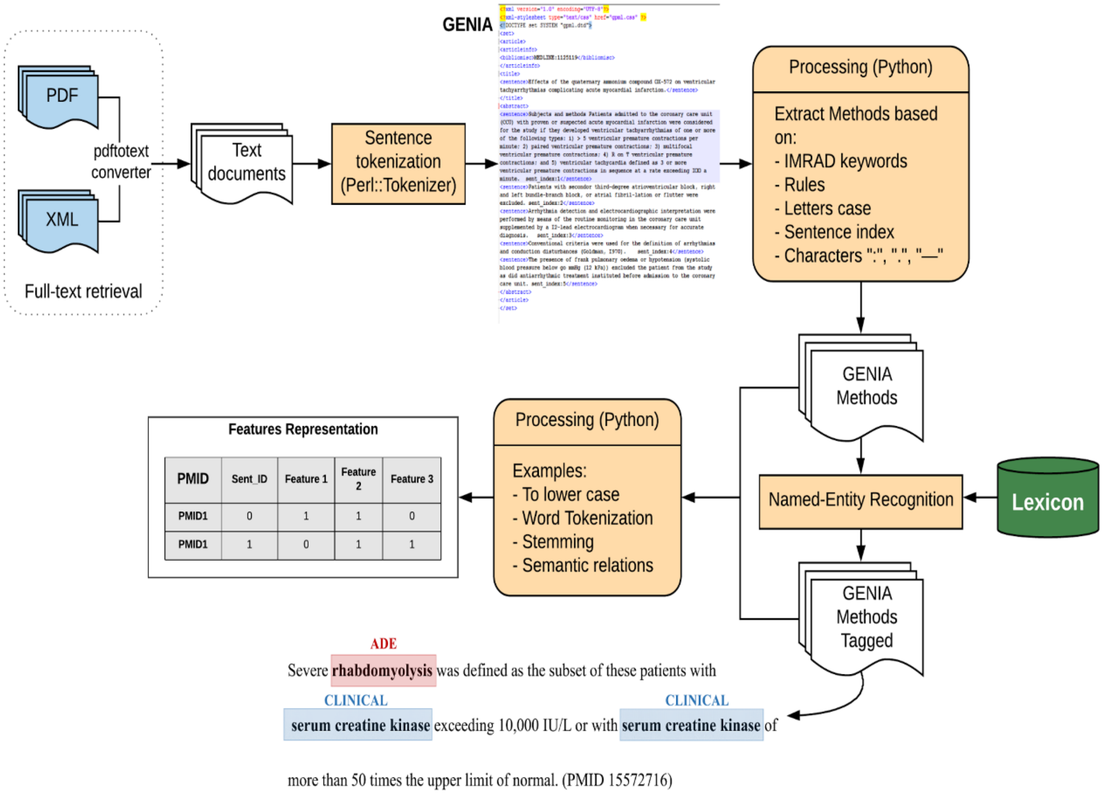
[1] The Xpdf is an open source project offers command line tools for processing PDF files (http://www.xpdfreader.com/about.html)

their boundaries are defined (Dalal & Zaveri, 2011) using a package called Perl::Tokenizer.

3. Convert from text format to GENIA XML format. GENIA format (J. D. Kim et al., 2003) has been used for bio-text-mining of the literature. GENIA is an XML format in which each article is annotated with PMID, Title, and full text sentences (Figure 6).
4. Extraction of method sections. Biomedical text in scientific papers are usually represented by four major sections; introduction, methods, results and discussion (IMRAD) (S. Agarwal & Yu, 2009). Using IMRAD keywords and rule-based methods, we were able to identify boundaries of the methods sections and extract them (Figure 6). IMRAD is a standard format that was recommended by American National Standards Institute (ANSI) since 1979 (American National Standards Institute. & Council of National Library and Information Associations (U.S.), 1979), where it is the most used format in many research journals (Nair & Nair, 2014). We utilized rule-based approach to extract these sections based on number of features (Figure 6). To automate extraction of method sections within the full texts according to IMRAD, we used "*Baseline*" classifier system which showed strong performance (S. Agarwal & Yu, 2009). "*Baselin*e" system is a simple classifier that works by assigning each sentence IMRAD category to its original IMRAD section in structured full texts (S. Agarwal & Yu, 2009). Therefore, we developed a rule-based program that assigns sentences to IMRAD headings. We used two categories for our system: relevant section and irrelevant section (Table 9). For example, if a keyword "Methods" appears in a sentence all subsequent sentences were assigned to "Methods" until "Results" keywords appear. Specifically, features that implies heading were considered, such as capitalization of first letter of the keyword, the presence of ":" or "—" after the keyword, or the presence of capital letter after the keyword. Keywords and rules used are shown in Appendix 8.

Table 9 Sections used for the "Baseline" extraction of full text articles

Category	**Sections**	**Reason**
Relevant section	Methods	Methods sentences are the sentences that we used for extracting phenotyping definitions information.
Irrelevant section	Introduction, Results, Discussion, Conclusion, References, Other sections	Sentences from these sections did not show a significant presentation of a phenotyping definition information.



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/css" href="gpml.css" ?>
<!DOCTYPE set SYSTEM "gpml.dtd">
<set>
<article>
<articleinfo>
<bibliomisc>MEDLINE:1125119</bibliomisc>
</articleinfo>
<title>
<sentence>Effects of the quaternary ammonium compound CK-572 on ventricular
tachyarrhythmias complicating acute myocardial infarction.</sentence>
</title>
<abstract>
<sentence>Subjects and methods Patients admitted to the coronary care unit
(CCU) with proven or suspected acute myocardial infarction were considered
for the study if they developed ventricular tachyarrhythmias of one or more
of the following types: 1) > 5 ventricular premature contractions per
minute: 2) paired ventricular premature contractions; 3) multifocal
ventricular premature contractions; 4) R on T ventricular premature
contractions: and 5) ventricular tachycardia defined as 3 or more
ventricular premature contractions in sequence at a rate exceeding 100 a
minute.  sent_index:1</sentence>
<sentence>Patients with secondor third-degree atrioventricular block, right
and left bundle-branch block, or atrial fibril-lation or flutter were
excluded. sent_index:2</sentence>
<sentence>Arrhythmia detection and electrocardiographic interpretation were
performed by means of the routine monitoring in the coronary care unit
supplemented by a 12-lead electrocardiogram when necessary for accurate
diagnosis.   sent_index:3</sentence>
<sentence>Conventional criteria were used for the definition of arrhythmias
and conduction disturbances (Goldman, 1970).    sent_index:4</sentence>
<sentence>The presence of frank pulmonary oedema or hypotension (systolic
blood pressure below go mmHg (12 kPa)) excluded the patient from the study
as did antiarrhythmic treatment instituted before admission to the coronary
care unit. sent_index:5</sentence>
</abstract>
</article>
</set>
```

PMID	Sent_ID	Feature 1	Feature 2	Feature 3
PMID1	0	1	1	0
PMID1	1	0	1	1

Severe **rhabdomyolysis** was defined as the subset of these patients with **serum creatine kinase** exceeding 10,000 IU/L or with **serum creatine kinase** of more than 50 times the upper limit of normal. (PMID 15572716)

Figure 6 Full text processing (Phase 2 & 3 in Figure 5)

3.3.6.3 Phase 3—Large-scale screening of full-text sentence-level

After full text processing and method section extraction, we followed similar steps of feature definitions for the sentence classifier when it was trained in the corpus (Figure 5 & Figure 6). First, NER was conducted to identify and normalize ADE/medical and drug terms (the dictionaries used for medical entities are described in 3.3.1 section). Second, sentences ware represented as a matrix. Each row represents one sentence, and each column represents a feature. This data matrix is ready for the sentence-level classification. We used our optimal full-text sentence-level classifier trained from our corpus for the prediction. It was conducted in in WEKA software package. Positive sentences, i.e. phenotyping related, were further used in the next analysis.

3.4 Results

3.4.1 The dictionary and lexica

Table 10 shows the dictionary and lexica that we developed for text-mining tasks by combining multiple standard terminologies. These were used mainly for the IE full-text sentence-level classifier. For example, a sentence with any of ADE entities is represented as a feature in the matrix of phase 3 (Figure 5). Similarly for CLINICAL, Procedure, and Drug entities.

Table 10 Dictionary for 279 adverse drug events (ADEs) and other medical terms used for extraction of full-text sentence-level features

Entity & Dictionary	Number of terms
279 ADEs (100%)	**5627**
MedDRA PT (4.9%)	279
SNOMED-CT (62.5%)	3517
MEDIC (MESH) (32.5%)	1831
CLINCAL (100%)	**471979**
SNOMED-CT (84.3%)	398077
MEDIC (MESH) (15.3%)	72167
MEDIC (OMIM) (0.4%)	1735
Procedure (100%)	**190399**
SNOMED-CT (98.8%)	188031
ICD-9 Procedures (1.2%)	2368
Drug (100%)	**21752**
DrugBank (100%)	21752
Total	**689751**

3.4.2 Optimal machine learning algorithms for classifying phenotyping related abstracts and full text sentences

The abstract-level classifiers were built upon the positive and negative abstract training dataset in our corpus. Table 11 shows the performance of the abstract-level classifier for SMO, J48 Decision Tree, Logistic Regression, and Naïve Bayes. SMO and J48 Decision Tree outperform the other algorithms, and their recall, precision, and F-measure are as high as 97%; while Naïve Bayes and logistic regression's performances are slightly lower.

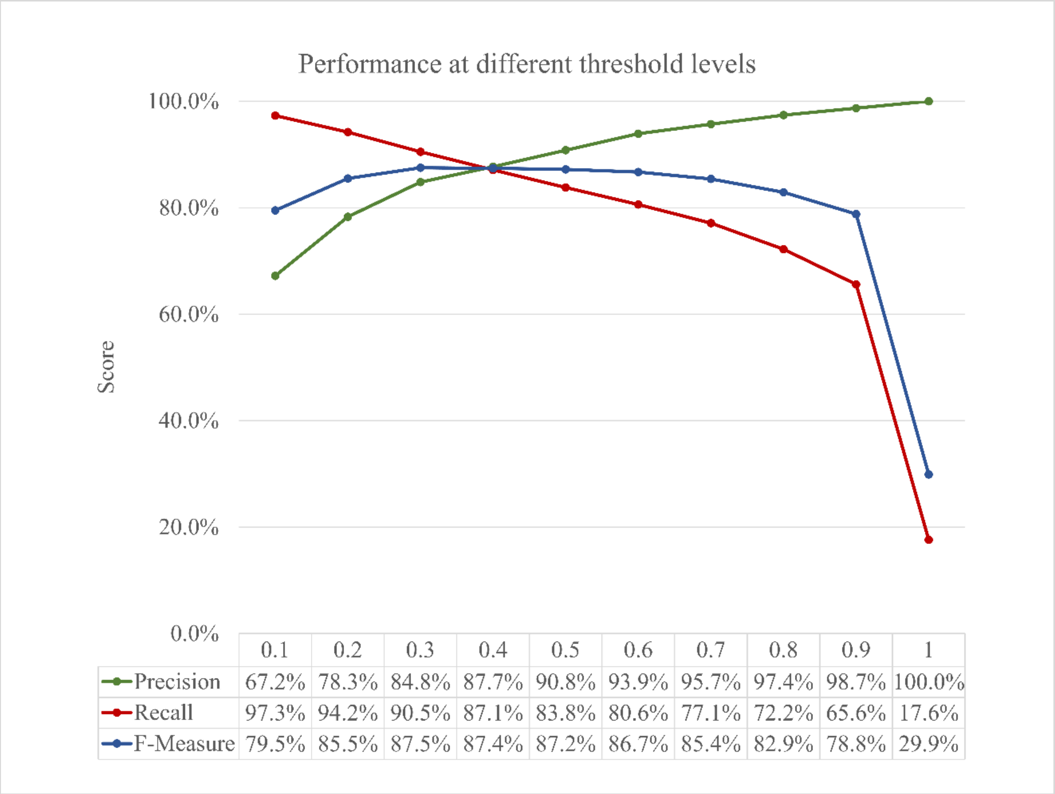
Similarly, full-text sentence-level classifiers were developed under the positive and negative sentence training dataset in our corpus. The classification performances of SMO, decision tree, logistic regression, and Naïve Bayes, are reported in Table 11. Overall, SMO and logistic regression showed the best performance, and their precision, recall, and F-measures reach as high as 0.84. Decision tree and Naïve Bayes's performances were a bit lower. Similar to the abstract level classifiers, sentence level classifiers were trained under the 10-fold cross-validation in the training set.

Table 11 Classifiers performance for abstract level classifiers and full sentence classifiers on 10-cross validation

Classifier	**Algorithm**	**Precision**	**Recall**	**F-Measure**
Abstract-level Classifier	***SMO**	0.972	0.972	0.972
	J48 Decision Tree	0.971	0.971	0.971
	Logistic Regression	0.953	0.953	0.953
	Naïve Bayes	0.924	0.908	0.909
Full-text sentence-level Classifier	**SMO**	0.846	0.844	0.843
	J48 Decision Tree	0.817	0.816	0.816
	***Logistic Regression**	0.840	0.838	0.837
	Naïve Bayes	0.799	0.796	0.794

*The selected algorithm for this classifier

For the full-text sentence-level classifier, we optimized the performance for recall. The default threshold that is used in WEKA is 0.5 where the predicted probability should be higher than 0.5 to be predicted as 'positive'. This threshold can be adjusted manually in Weka using "manualThresholdValue" for values between 0 and 1. Figure 7 shows the plot to visualize the threshold values of the predicted probability for logistic regression. Since we were interested in high recall, we selected 0.2 as our threshold for 'positive' category with a recall of 94.2%.



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Precision	67.2%	78.3%	84.8%	87.7%	90.8%	93.9%	95.7%	97.4%	98.7%	100.0%
Recall	97.3%	94.2%	90.5%	87.1%	83.8%	80.6%	77.1%	72.2%	65.6%	17.6%
F-Measure	79.5%	85.5%	87.5%	87.4%	87.2%	86.7%	85.4%	82.9%	78.8%	29.9%

Figure 7 Full-text sentence-level classifier performance using logistic regression (threshold selector)

Both the abstract-level and full-text sentence-level classifiers were further validated by a random subset of 30% of the corpus. Table 12 shows the number of documents for each of the training and testing dataset. Both abstract level classifier and sentence level classified have the comparable performance as in their training samples, and F-measures are 0.98 and 0.81 respectively.

Table 12 Classifiers validation results on testing dataset (70% validation results)

Classifier	**Training**	**Testing**	**Optimal Algorithm**	**Class**	**Performance measures**		
					Precision	**Recall**	**F-Measure**
Abstract-level	1315	563	SMO	**Positive**	0.97	**0.98**	0.98
				Negative	0.99	0.98	0.98
				Averaged	0.98	0.98	0.98
Full-text sentence-level	2780	1191	Logistic Regression	**Positive**	0.79	**0.86**	0.82
				Negative	0.85	0.77	0.81
				Averaged	0.82	0.81	0.81

3.4.3 Literature large-scale prediction results

For literature large-scale phenotyping case definitions discovery, we used our validated classifier for the automatic phenotype discovery (Table 11). PubMed abstracts were used for years between 1975 and first quarter of 2018. Using our abstract-level classifier (SMO machine-learning algorithm), the number of abstracts that were predicted as positive, i.e. phenotyping related, are 459,406 abstracts (the distribution of abstracts on years is shown in Appendix 9. For positive abstracts, we retrieved their full texts. Some filters were applied, such as institutional full text accessibility and exclusion of animal studies. We retrieved the full text only either as PDF or XML. Some scanned articles (i.e. pictures) cannot be converted into text file, and full texts with issues in either PDF and XML format were excluded. Therefore, the total number of the final set of full texts is 120,868. Using these full text articles, 6,129,574 sentences were extracted from their methods sections. Using our full-text sentence-level classifier (logistic regression machine-learning algorithm), the number of sentences that were predicted as positive were 2,745,416. Table 13 shows a summary of the results.

Table 13 Results for large-scale screening of abstracts and full texts sentences

Abstracts (Abstract-level classifier)	
Number of predicted positive abstracts (1975-2018 “Mid-March”)	459,406
Full-text Retrieval	
Number of full text retrieved (Filters: full text available, not animal studies)	141,511
Number of full text after data processing	120,868
Full-text sentence-level classifier	
Total number of sentences (Method section)	6,129,574
Number of predicted positive sentences	2,745,416
Number of predicted negative sentences	3,384,158

3.5 Discussion

This study proposed an automated approach to mine a phenotyping case definition in the biomedical literature. First, we built a dictionary that we used in text-mining tasks, such as NER. We used several standard terminologies to build dictionary has 689,752 terms for entities: ADEs, Clinical, procedure, and drug. Second, we built two classifiers and selected the optimal machine learning algorithms using our annotated corpus from Chapter

2, which are abstract-level and full-text sentence-level classifier with F-measures 0.98 and 0.81, respectively. Third, we used our validated classifiers for large-scale information retrieval and extraction in the literature. We predicted 459,406 abstracts as positive that were used for further analysis after applying some further filters that includes abstracts with full texts and exclude abstracts for animal studies. We were able to include 120,868 full texts, utilized their sentences within methods sections, and predicted 2,745,416 classified as positive. Using these sentences, we aimed to support research for phenotyping by deriving literature scientific information with evidence of a phenotyping definition.

3.5.1 Error analysis

We performed an error analysis to identify sources of error. We randomly selected 100 misclassified sentences and identified some possible reasons for the misclassifications of full-text sentence-level classifier. Table 14 shows examples of errors. These are challenges associated with text-mining. Ambiguity is one of the challenges where in some situations it is hard to correctly infer the meaning of character, symbol, or term. For example, periods can indicate the end of a sentence or a word abbreviation (Shatkay & Craven, 2012). An error we called as Negative atypical showed the highest number among others with 40% appeared in sentences. A possible reason is that we extracted features that focused on positive cues for evidence of a phenotyping definition. These sentences can be also very similar to positive sentences in features, but it is negative. On the other hand, positive atypical are sentences that shows no evidence for positive cues or few that were not sufficient to classify correctly them as positive. One of the reasons can be because these sentences are too short that few or no positive cues found, but they are still supportive for phenotyping. Clinical dictionary and keywords dictionary showed also percentages of 37% and 35%, respectively. Both of these errors are utilizing NER approaches. The difference is the clinical dictionary used clinical entities from standardized dictionaries, such as MedDRA. On the other hand, keywords dictionary are the terms that we derived from our corpus analysis for positive or/and negative cues that represent or does not represent a phenotyping definition. In addition, abbreviations were challenging. Even though we used general approaches of patterns recognition to recognize some of the abbreviations such as term length and preceded or succeeded terms, this was not one of the scope of this work.

Word boundary detection and semantic ambiguity were the lowest frequent errors among this subset of 100 sentences with 16% and 7%, respectively.

Table 14 Error analysis for full-text sentence-level classifier

Error	**Description**	**Example**	**Example description**	**Positive (n = 50)**	**Negative (n-=50)**	**Total (n=100)**
Abbreviations	Abbreviated phenotypes are harder to be recognized.	"A little more than one third of all patients identified by the NLP method were manually confirmed to have HF." (PMID17567225)	This sentence is actual positive and predicted as negative. HF was not recognized as a clinical phenotype in the sentence.	27(54%)	3 (6%)	30 (30%)
Word boundary detection	The word boundary detection of some keywords.	"First-line treatment." (PMID11388131)	This sentence is actual negative and predicted as positive. "men" is one of the keywords of patient. Here, "men" recognized from "treatment" due to incorrect word boundary identification.	4(8%)	12(24%)	16 (16%)
Semantic ambiguity	Entities that have same spelling, but different meaning depending on the context.	"Since nearly everyone residing in the target ZIP code for the current study receives their health care through Marshfield Clinic, this record is considered comprehensive." (PMID17456828)	This sentence is actual negative and predicted as positive. The term "code" was associated here with other positive cues	4(8%)	3(6%)	7 (7%)
Clinical dictionary	The dictionary needs updates for	"Uses inhaled steroids regularly." (PMID12952547)	This sentence is actual positive and predicted as negative.	25(50%)	12(24%)	37 (37%)

	inclusion terms.		"inhaled steroids" was not recognized as the term not in our dictionary.			
Keywords dictionary	The dictionary needs updates for inclusion terms.	“Baseline characteristics were compared using a X2 test for categorical variables and ANOVA for continuous variables.” (PMID15323063)	This sentence is actual negative and predicted as positive. X2 test and ANOVA should be recognized as a keywords for negative cues	8(16%)	27(54%)	35 (35%)
Positive atypical	Sentences with positive phenotyping context, but not clear evidence.	“Uses inhaled steroids regularly.” (PMID12952547)	This sentence is actual positive and predicted as negative. It does not have features for phenotype definitions.	23(46%)	0(0%)	23 (23%)
Negative atypical	Sentences with negative phenotyping context, but not clear evidence.	“Multivariate logistic regression identified patient factors associated with a correct diagnosis.” (PMID17712071)	This sentence is actual negative and predicted as positive. It has many features for positive sentences. In addition, “regression” was false positive as a clinical diagnosis (phenotype).	0(0%)	40(80%)	40 (40%)

3.5.2 Limitations of the study

This work does not stand without limitation. The annotated corpus might not be representative to all of the cases that either represent or does not represent a phenotyping case definition. However, based on the error analysis, these cases are not very problematic as we optimized the full-text sentence-level classifier for high recall for positive class. Another limitation is reliance on NER features for the automatic classification of sentence-level (S. N. Kim, Martinez, Cavedon, & Yencken, 2011). Our error analysis shows that some terms might not be correctly recognized or are not present in the dictionaries. Therefore, we note that these lexical-based methods require frequent updating of the dictionaries. Furthermore, hand-crafted features based on expert evaluations have also been used for sentence-level features, where it might not be optimal. We note that we utilized TF and TF-IDF for extracting some of these terms (Binkheder et al., 2018). Additional automatic representation and detection of features might be supportive in future studies (Kilicoglu et al., 2018) for better representation and performance.

Most of the literature documents are represented as PDF or XML. For example, PDF documents are more concerned on how the information look like rather than the structure of information. On the other hand, XML are more structured, but the structured can vary across publishers (Shatkay & Craven, 2012). This creates a challenge as we found some of the PDFs were in image format and/or some of XML were without content. Due to the challenges associated with full texts, most of the studies have focused on abstracts because they are free, easy to download, concise, and less challenging to mine (Shatkay & Craven, 2012). Here, our goal was to use full texts as we found that they contain most of the information for representing a phenotyping definition. Overall, conversions of full texts from PDF to text generated several unwanted characters, especially when using this information in GENIA XML format.

3.5.3 Future work

We believe that our large-scale corpus of predicted positive sentences provides a potential source for further applications of mining phenotyping definitions, such as implementing an evidence-based best practices for cohort identification research studies (Yadav et al., 2018). Several of studies shows that relying only on ICD-9 codes is not sufficient in building cohorts, and it is critical to utilize other sources like clinical notes (Iyer et al., 2014). Most efforts of automating the development of phenotyping definitions used EHR data, such as billing and administrative data and clinical text (Yu et al., 2015) can be biased. Developing data-driven approaches and more structured definitions using literature-mining of phenotyping definitions is recommended. Based on the challenges of developing DILI algorithm, such as identification of patients with rare conditions, incomplete knowledge about DILI and its translation to phenotyping definition expressions, and variations in interpretation of the definition, textual descriptions of definitions are still an issue (Overby et al., 2013). We believe that this is one of the steps towards standardization of phenotyping definitions. Literature showed that the differences across phenotyping definitions can affect their applications across studies as well as the interpretation of results (Chute et al., 2011) and standardization for better portability is still a challenge (Fort et al., 2014). There were some efforts in the standardization of the representation phenotyping definitions (Chute et al., 2011). Examples are eMERGE and OMOP where eMERGE supports portability (Ho et al., 2014). When a phenotyping definition is standardized, it can provide consistent inclusion and exclusion criteria to define a phenotype of interest across databases (R. Richesson et al.). Utilizing our collected sentences, future text-mining applications can be built for tasks, such as information extraction and text summarization. Furthermore, we believe that literature-based mining of a phenotyping definition supports future work of hypothesis generation to discover unknown and novel correlations and patterns for phenotypic associations hidden in text.

In conclusion, we proposed an automated approach to extract sentences with information of a phenotyping case definition. Two classifiers were built: abstract-level and full-text sentence-level. Both classifiers showed good performance in predictions and were applied to large scale literature. Future efforts are needed to support areas of text mining and knowledge discovery of phenotyping definitions information in the literature.

CHAPTER FOUR: DISCOVERY STUDY TO REPRESENT AND VALIDATE LITERATURE-BASED PHENOTYPING DEFINITIONS

In Chapters 2 and 3, we created corpora for abstracts and full texts' sentences to retrieve and extract phenotyping definition-related information. In this chapter, we utilize our large-scale corpus of over two million sentences that were predicted as positive for phenotype definition-related sentences. Our goal is to perform information extraction and a knowledge discovery study for some phenotypes of interest, such as type 2 diabetes and myopathy. After that, we provide some evaluations of the used approaches in this study.

4.1 Introduction

A phenotyping definition is critical for clinical and pharmacovigilance research. World health organization (WHO) and the centers of Disease Control (CDC) developed case definitions for some conditions (Botsis & Ball, 2013), these are referred by low-throughput or expert-driven definitions (R. L. Richesson et al., 2016) that we introduced in Chapter 3. Low-throughput or expert-driven has several challenges. For example, we introduced the PheKB example (in Chapter 2 & 3) that is capable of disseminating and validating definitions across institutions. However, they lack structured representation of a case definition (Chute et al., 2011). Low-throughput phenotyping is still a long process, labor-intensive not scalable, and does not cover all phenotypes of interest (Botsis & Ball, 2013; Henderson et al., 2017). Therefore, the main drawbacks with such manual processes of developing (Botsis & Ball, 2013) and representing (Rosenman et al., 2014) a phenotyping definition is affecting the progress of several research areas and surveillance.

Existing phenotyping definitions are useful to establish clinical study or to validate these definitions. However, existing definitions are more complete for some conditions, e.g. myocardial infarction, but less complete for others, e.g. osteoporosis (Rosenman et al., 2014). Additionally, most of these definitions are not capable of handling complex models, such as the ones for unstructured data (Xu et al., 2015) that lack flexible phenotyping definitions (Thompson et al., 2012). In validation challenges, an increase in definition complexity means harder validation across different institutions, e.g. HTCP definitions lack of standardization across EHRs (Simonett et al., 2015). Therefore, there is a need of informatics approaches to automate the general process of representing case definitions (Botsis & Ball, 2013).

The current direction is towards developing automated approaches for high-throughput phenotyping that uses data-driven approaches (Conway et al., 2011; R. L. Richesson et al., 2016). Machine learning algorithms are more capable of discovering unknown relationships because the prediction logic uses real-world data rather than prior knowledge. Phenotypes generated from such data-driven approaches are called computational phenotypes that can be rapidly generated in high volumes to scale up to the needs of high-throughput phenotyping (R. L. Richesson et al., 2016). However, current efforts of high-throughput phenotyping are mainly using EHR data to generate or derive computational phenotypes. The use of EHR data has several limitations. The nature of EHRs suffer from several issues, such as bias, confounding, missing/incomplete data, irregular data (Castro et al., 2014; Yadav et al., 2018), and more were discussed in Chapter 3. Furthermore, structured data, e.g. ICD-9 codes, have shown limited results in phenotyping; in comparison, when combined with NLP it showed better performance (Kotfila & Uzuner, 2015; Liao, Cai, et al., 2015; Roden & Denny, 2016) to obtain cases. Limestone (Ho et al., 2014) is an example of an EHR-based effort that uses data-driven approaches for deriving candidate computational phenotypes without the need of human supervision. Limestone investigated the interactions between diagnoses and medications using "tensors (a generalization of metrics)". They confirmed that 82% of the top 50 candidate phenotypes are clinical meaningful. Some of the limitations of Limestone, are: did not address portability, relied on only one medical expert, not all candidate were clinically meaningful (generating novel), and did not use text notes (Ho et al., 2014). Therefore, we believe that using EHR data to generate candidate computational phenotypes is not sufficient and should be supported with other sources such as biomedical literature.

There are several studies that used literature-based knowledge discovery. In "Automating case definitions using literature-based reasoning" (Botsis & Ball, 2013), Botsis and Ball used co-occurrence approach and network-graph to automate representation of "anaphylaxis" definitions from literature abstracts. Their aim is to replace the manual identification of synonyms and definitions by automation. They used case-based reasoning that utilize existing knowledge and build sematic similarity framework combined with machine learning approaches. Semantic relationships were used to build the graph network nodes and relationships in which semantically co-occurring terms with the term anaphylaxis. However, Botsis and Ball have only utilized abstracts rather than full texts that restricted the complete retrieval of information. They stated that generating a corpus of full text is time-consuming and expensive. In addition, their developed approach was for only one condition "anaphylaxis" where more generalization needs further research. Moreover, they did not consider all features of a phenotyping definitions where some features might not be necessary for anaphylaxis, such as laboratory values, but can be important for other conditions (Botsis & Ball, 2013).

PheKnow-Cloud (Henderson et al., 2017) is another study that used knowledge-discovery from literature. PheKnow-cloud leveraged clinical expertise from PubMed Open Access Subset by using the evidence of co-occurrence in sentences as an automatic validation. The user needs to specify candidate phenotypes that were derived from EHR data. Then, these candidate phenotypes can be validated by screening the literature as a validation tool. The phenotype significance metric called lift was used to measure the clinical significance of candidate phenotypes (Henderson et al., 2017). The limitation of this tool is the need to have a potential candidate generated from other source rather than generating potential candidates, which is our goal in this study.

Identification of cohorts of chronic diseases, such as diabetes, has critical value for the “clinical quality, health improvement, and research” as well as the development of patients’ registries and research datasets (R. L. Richesson, Rusincovitch, et al., 2013). With the utilization of standard phenotyping definitions, it enables comparison and aggregate analysis of patients on several levels, including population characteristics, risk factors, and complications (R. L. Richesson, Rusincovitch, et al., 2013). Recent results support the evidence that a phenotype or disease, like asthma and heart failure, are not single entity, but rather a collection of phenotypes. Data-driven approaches are unbiased and able to reveal unknown knowledge. Such analysis tools supported with large datasets are capable of discovering, for example, unknown clinical sub-phenotypes of diseases (Lasko et al., 2013). The hypothesis generation of risk factors involves the use of statistical models to “describe the relationship between a condition and phenotypic and clinical data” (Ouyang, Apley, & Mehrotra, 2016). These methods are considered unsupervised methods with the goal of minimizing human involvement (Lasko et al., 2013).

Text mining analysis, especially dictionary-based approaches, using literature text is well-studied in the biomedical field (Kilicoglu et al., 2018). One of the text mining tasks is the information extraction and knowledge discovery through development of an automated approaches for identification of co-occurrence concepts (Krenn, 2000). In the medical field, co-occurrence association measures have been used to identify similar diseases, predict disease-causing genes (Henry, McQuilkin, & McInnes, 2018), and perform literature-based discovery (Henry et al., 2018; Yetisgen-Yildiz & Pratt, 2006). Lexical-based approaches can be used to recognize co-occurrence of two biomedical concepts. Recurrent concept combinations are expected to contain co-occurrence candidates than low ranking combinations (Krenn, 2000). The association measures are used to evaluate the likelihood of significant pattern of co-occurrences between any two concepts. After that, networks can be built using co-occurrence information; for example, Davazdahemami and Delen (Davazdahemami & Delen, 2018) used literature to build a network for drug-ADE associations. They reported that when this approach if replicated on a larger scale it can generate better results (Davazdahemami & Delen, 2018).

4.2 Background

4.2.1 Phenotypes in EHRs

Phenotype is an observable property that result from interaction of an organism’s characteristics and environmental factors (N. Alnazzawi et al., 2015; Frey et al., 2014). In the medical domain, clinical phenotypes can be defined as one or collection of observable and measurable patient’s characteristics within a population (Frey et al., 2014; Glueck et al., 2016). The current phenotyping definitions are phenotype-specific, and are composed of an application of decision logics using EHR-based features and specifications (Carroll et al., 2011; Lasko et al., 2013; Mo et al., 2015; Yu et al., 2015). The phenotype features can come from EHR data, including structured data (coded data), occurrences of two coded temporal events, unstructured narrative text, occurrences of clinical concepts (Peterson & Pathak, 2014; Yu et al., 2015), or relations between events (Park & Choi, 2014). Some studies showed an improved accuracy when EHR phenotyping definitions are combining coded and clinical narratives than using one of them (Liao, Cai, et al., 2015; Roden & Denny, 2016; Yu et al., 2015). While other studies showed that structured data can be more effective in retrieving cohort data (Denny, 2012).

EHR phenotyping can be complex; for example, recent studies showed that heart failure or asthma are not composed of single entities, but rather a collection of phenotypes that can overlap with historical diseases (Lasko et al., 2013). Current technologies are not practical in deepen our understanding of phenotypes. Thus, one of the major challenges is the insufficient phenotype granularity, which can result in uncertainty during the process of EHR phenotyping (Glueck et al., 2016). Therefore, the analysis of EHR phenotypes requires a deep understanding of all aspects of phenotypes. Computational algorithms capable of dealing with massive amount of data are necessary to generate new discoveries from EHR data (Kotfila & Uzuner, 2015).

An understanding of phenotype-disease associations helps in diagnosing of diseases, improving treatments, identification of disease's etiology (N. Alnazzawi et al., 2015; Glueck et al., 2016). A phenotype can appear as an abnormal observation of one or a combination of the following: physiological, behavior, genetic, and physical traits. In addition, other factors can play an important role in the origin phenotypes, such as ethnicity, gender, and environment. Therefore, spectrums of phenotypic abnormalities are highly considered to better understanding of the phenotype-disease associations (Glueck et al., 2016). The use of EHR data has facilitated "novel clinical decision support, biomedical association studies, auditing and EHR security, and the cost effectiveness of treatments" (Chen et al., 2015). The research on the EHR data enable conversion of this data into knowledge (Chen et al., 2015).

Phenotype definitions are used for cohort identification utilizing risk factors, clinical or medical characteristics and complications (Yadav et al., 2018). For example, coronary artery disease (CAD) study on three cohorts discovered that the risk of CAD is 63.68% lower in rheumatoid arthritis and inflammatory bowel disease than diabetes mellitus. They developed CAD algorithm to compare the risk factors across diseases. However, their use of EHR data affected its generalization. The authors stated that one of the major limitations of the study is generalization of the algorithm due to ascertainment and bias on recording risk factors in EHR data facilities. They recognized the need to identify risk factors across other population (Liao, Ananthakrishnan, et al., 2015). Therefore, we believe that our approach provides more generalized evidence that can be applied over different populations.

4.2.2 Co-occurrence and graph-based representation

In Natural Language Processing (NLP), words combinations or associations are an important source of information (Evert, 2005; Kolesnikova, 2016), such as knowledge generation, text analysis and generation, knowledge extraction, text summarization, and information retrieval (Kolesnikova, 2016). Such collection of co-occurrences of concepts derived from real-world data e.g. sentences from literature creates important source of knowledge. A database can be created to represent these collocations of co-occurrences (Krenn, 2000) where frequency information of co-occurrences, when interpreted, can indicate a statistical association (Evert, 2005). Section 4.2.3 provides a summary about these measures of associations.

Lexical-based approaches are used for recognition of co-occurring concepts rather than semantics (Krenn, 2000). One of the advantages of this technique is that it does not involve complicated linguistic theories (Chung & Lee, 2001). Mainly, we used positional co-occurrence when the terms co-occur within a certain distance (Evert, 2005). For example, the distance that we used in this study is the co-occurrence of two terms within the same sentence, and we called it direct co-occurrence. The direct co-occurrence can be extended using the network graph approaches into indirect co-occurrences.

4.2.3 Measures of associations

Association measures are statistical tests that help to distinguish between random co-occurrences and true associations. These frequencies and measures are used for ranking of pairs and/or selection of cutoff threshold. Co-occurrences are called candidates until specified criteria is employed (Evert, 2005). There are several associations' measures models. Frequency counts associations is the simplest co-occurrence measure for association. However, frequency performance is weak because it only considers positive co-occurrences and does not consider the frequency of single words/terms. On the other hand, statistical lexical-based co-occurrence approaches provides better results, such as

DICE coefficient (Krenn, 2000). Association coefficients, such as Jaccard's and Dice's, are also called similarity coefficients (Chung & Lee, 2001).

Dice coefficient (Smadja, McKeown, & Hatzivassiloglou, 1996) is one of the simplest association measures that considers the significance of each combination of words. Dice coefficient sums up the conditional probabilities p(X|Y) and p(Y|X) with equal weights and it considers the significance of individual words. This lowers the bias when data frequency is relatively low (Krenn, 2000). The degree of an association between two words or linguistic elements can be measured by "coefficients of association strength from the observed data" (Kolesnikova, 2016). Dice coefficient outperformed other association measures as well as it showed better performance for dictionary-based co-occurrences (Kolesnikova, 2016).

In this work, we aimed to build data-driven approaches and hypothesis-driven investigation for high-throughput phenotype representations. We utilized our large-scale corpus data predicted from literature mining of full text sentences. The aim for our work is to use the corpus of sentences (from Chapter 3) with evidence of phenotyping information to identify potential set of computational phenotype candidates. Biomedical concepts were used for statistical identification of co-occurrences within the sentences. Reducing terms to their preferred terms is one approach for the identification of recurrent concepts with the aim to increase co-occurrences (Krenn, 2000). Co-occurrence approaches and association measures were used for extraction and ranking of the biomedical and procedure concepts. Finally, we compared our results with existing Gold/Silver standards, such as PheKB and UpToDate.

4.3 Methods

4.3.1 Co-occurrence analysis of phenotypes

Co-occurrence analysis is a functional relationship and occurrence of two medical terms within a sentence (Evert, 2005; Fleuren & Alkema, 2015). Table 15 shows the dictionary used to recognize medical entities in sentences. After that, co-occurrences of the medical entities were extracted regardless of each entity length (Krenn, 2000) for further co-occurrence analysis.

Table 15 Dictionary used to extract co-occurrence and MedDRA normalization

Terminology	Number of terms in the dictionary	Percent of terminology in the dictionary	Normalization to MedDRA Preferred terms	
			Mapped	Not mapped
MedDRA LLT	69955	9.5%	69955	0
SNOMED CT	580580	79.2%	15062	565518
MEDIC (MESH)	78650	10.7%	44100	34550
MEDIC (OMIM)	1643	0.2%	67	1576
ICD9 Procedure	2354	0.3%	0	2354
Total	733182	100%	129184 (17.6%)	603998 (82.4%)

After the extraction of co-occurrence terms, they were represented as a document-term matrix (DTM). DTM is a matrix that the rows are the sentences and the columns are the terms. This was further converted into co-occurrence matrix (n x n); n is the total number of terms. The matrix has the frequency of co-occurring phenotypes in columns and rows (Figure 8). After that, all terms were normalized by mapping them to MedDRA preferred terms.

4.3.2 DICE scores for ranking phenotypes

DICE coefficient is one of the statistical methods used in measuring the degree of the association between words x and y in an observed dataset (Evert, 2005). Olga Kolesnikova study (Evert, 2005) showed that DICE coefficient outperformed other association measures. The DICE coefficient is calculated as: $D = \frac{2f(xy)}{f(x)+f(y)}$

Where:

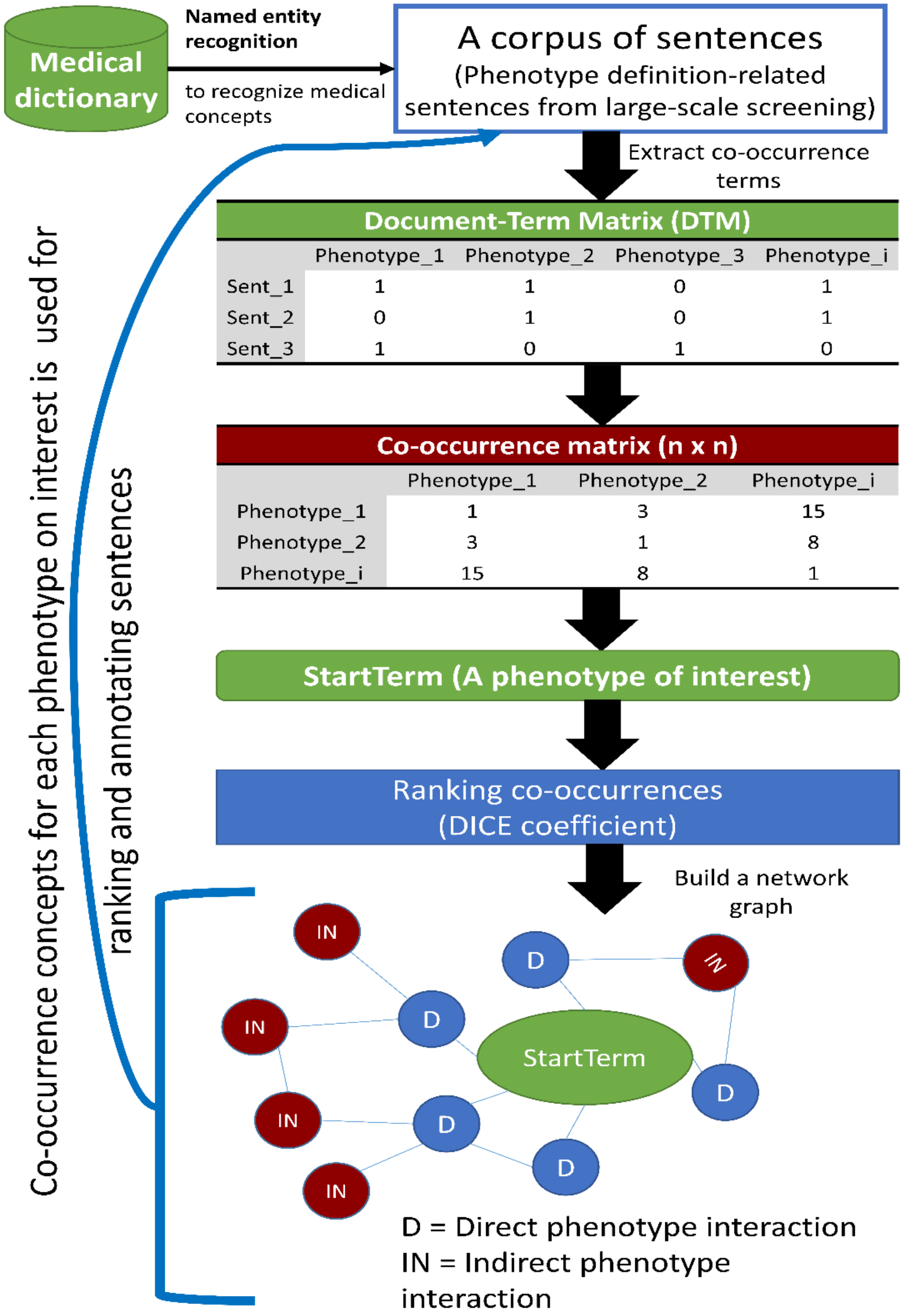
- F(x) Number of occurrences of x (StartTerm)
- F(y) Number of occurrences of comparison term y
- F(xy) Number of joint occurrences of x (StartTerm) and y

To calculate DICE coefficient, we start with an ADE or phenotype of interest which we called StartTerm and calculated its associations. All terms with zero DICE scores were eliminated as these indicate that they appeared in the corpus but did not co-occur with StartTerm. Terms were ranked in descending order where the highest means more significant associations. Directly co-occurred terms are terms that co-occurred with StartTerm in the same sentence. Indirectly co-occurred terms are terms that did not co-occurred with StartTerm in the same sentence. Indirectly co-occurred terms were derived by finding the co-occurred terms for the top-ranked 5% directly co-occurred terms with StartTerm. The 5% is our selected threshold, but this can be changed. The relation here is when A is related to B and B is related to C, so A is related to C. Both direct and indirect terms were utilized in building the network graph of the starting term. NER and term extraction were performed in Python and Perl. The co-occurrence analysis and DICE ranking (sections 4.3.1 & 4.3.2) were performed in R[1] (Wiedemann & Niekler).

4.3.3 Network graphs

To visualize the co-occurrence results, we used DICE coefficient scores for a selection of significant phenotype terms with the StartTerm (Figure 8). Terms were normalized to MedDRA PT (Table 15) to facilitate the generation of the network. Both direct and in-direct terms were used to generate the co-occurrence network. Network graphs were evaluated and visualized in Cytoscape (Shannon et al., 2003).

[1] The code and R packages can be found in "Tutorial 5: Co-occurrence analysis by Andreas Niekler and Gregor Wiedemann": https://tm4ss.github.io/docs/Tutorial_5_Co-occurrence.html


Medical dictionary
Named entity recognition
to recognize medical concepts
A corpus of sentences
(Phenotype definition-related sentences from large-scale screening)
Extract co-occurrence terms
Document-Term Matrix (DTM)

	Phenotype_1	Phenotype_2	Phenotype_3	Phenotype_i
Sent_1	1	1	0	1
Sent_2	0	1	0	1
Sent_3	1	0	1	0

Co-occurrence matrix (n x n)

	Phenotype_1	Phenotype_2	Phenotype_i
Phenotype_1	1	3	15
Phenotype_2	3	1	8
Phenotype_i	15	8	1

StartTerm (A phenotype of interest)
Ranking co-occurrences
(DICE coefficient)
Build a network graph
IN
D
IN
D
IN
D
StartTerm
D
IN
D
D
IN
D = Direct phenotype interaction
IN = Indirect phenotype interaction
Co-occurrence concepts for each phenotype on interest is used for
ranking and annotating sentences

Figure 8 Co-occurrence analysis, DICE ranking, and network graphs

4.3.4 Evaluation of literature-based co-occurrence results

To validate and evaluate the phenotypes generated from co-occurrence analysis, we performed some further analysis. The goal is to compare the phenotype terms from literature with existing sources for terminologies, phenotyping definitions, and clinical guidelines.

4.3.4.1 Evaluation of derived co-occurrences with 50/50 sample split of articles

To evaluate if the co-occurrences are reproducible when derived from two independent sets of articles, we randomly divided the dataset into two subsets based on the number of articles per set. We used "Myopathy" as the StartTerm. The co-occurrence analysis and DICE ranking were performed on each of the datasets separately. Paired T-test in R was used to compare the co-occurred terms with "Myopathy" in the two samples.

4.3.4.2 Comparing T2DM concepts with existing sources for standard terminologies

The goal is to compare phenotypes representations between our literature-based results and other existing terminology systems. We used the significant co-occurrence terms for "Myopathy" extracted from the full dataset. For comparison, we selected two terminology systems, MedDRA and SNOMED-CT. We compared our co-occurring myopathy terms to the myopathy related terms in two different terminology systems.

4.3.4.3 Comparing with existing sources phenotyping definitions and clinical guidelines

A number of existing sources are available providing information for case definitions or clinical guidelines. Here we selected PheKB[1] and UpToDate[2]. PheKB is a "collaborative environment to building and validating electronic algorithms to identify characteristics of patients within health data." On the other hand, UpToDate is "an evidence-based, physician-authored clinical decision support resource which clinicians trust to make the right point-of-care decisions." Documents from both sources were

[1] https://phekb.org/ (accessed on September-October 2018)

[2] https://www.uptodate.com (accessed on September-October 2018)

converted into raw text, and terms were extracted using our dictionary and lexica (Table 15). All of the extracted terms were compared across the three sources literature-based results, PheKB, and UpToDate. Our phenotypes selection for comparison was based on the overlap between our list of ADEs (in Chapter 2) and phenotypes that have case definitions available in PheKB. We evaluated 10 phenotypes that were existed in the three sources, which are: Diabetes Mellitus Type 2 (T2DM), acute coronary syndrome, aneurysm, arthritis, cardiac failure, cough, dementia, High-Density Lipoprotein (HDL) decreased, hypercholesterolaemia, hypothyroidism.

4.3.4.4 Manual analysis of Type 2 Diabetes Mellitus (T2DM) co-occurred terms

In this analysis, we utilized the terms for T2DM as a case study that were extracted from literature-based results, PheKB, and UpToDate. We manually evaluated all the terms for their clinical significance. For each source, we manually evaluate the phenotype by utilizing text documents that were used to extract the terms. For literature-based results, we used DICE scores of the T2DM significant terms to rank all positive sentences in our dataset. After that, articles were ranked in descending order using the sum of DICE score of sentences within that article. All of the text documents were used for manual categorizing and validation of all T2DM's phenotypes (Figure 9).

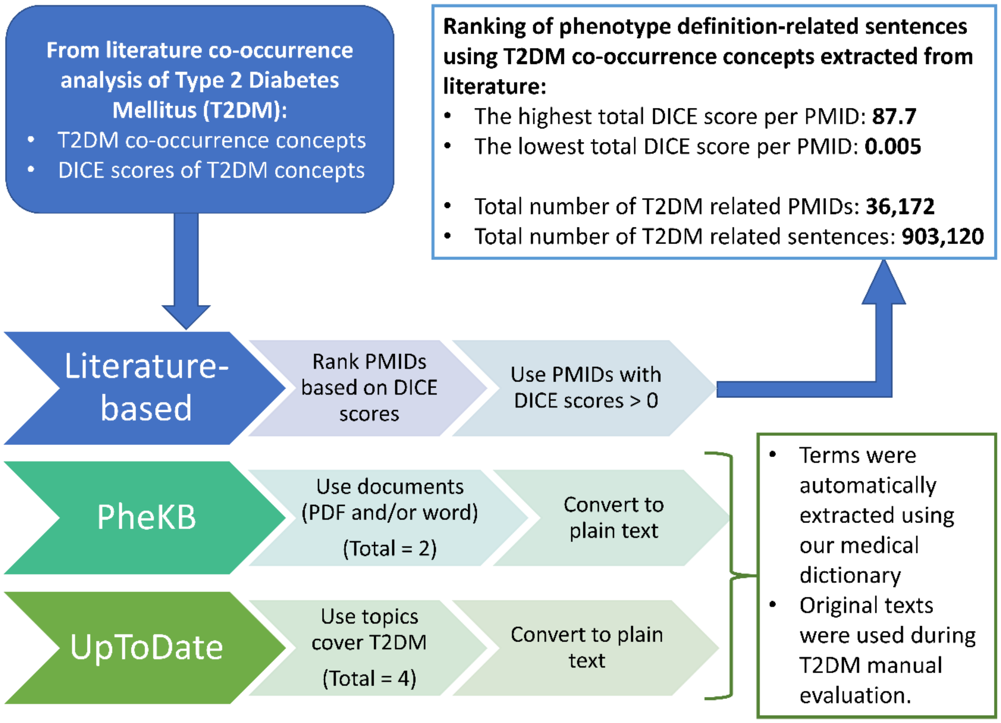

From literature co-occurrence analysis of Type 2 Diabetes Mellitus (T2DM):
• T2DM co-occurrence concepts
• DICE scores of T2DM concepts
Ranking of phenotype definition-related sentences using T2DM co-occurrence concepts extracted from literature:
• The highest total DICE score per PMID: 87.7
• The lowest total DICE score per PMID: 0.005
• Total number of T2DM related PMIDs: 36,172
• Total number of T2DM related sentences: 903,120
Literature-based
Rank PMIDs based on DICE scores
Use PMIDs with DICE scores > 0
PheKB
Use documents (PDF and/or word) (Total = 2)
Convert to plain text
UpToDate
Use topics cover T2DM (Total = 4)
Convert to plain text
• Terms were automatically extracted using our medical dictionary
• Original texts were used during T2DM manual evaluation.

Figure 9 The process of processing type 2 diabetes mellitus data and ranking of T2DM definition-related sentences

4.4 Results

4.4.1 Co-occurrence analysis results

We used the 2,745,416 positive sentences (Chapter 3) to extract the co-occurrences terms. The number of sentences with co-occurrence terms is 1,414,380. The number of co-occurrences is 12,616,465 and the number unique terms is 19,423 (Table 16). These phenotypes are shown in Figure 10 as a word cloud based on their frequency of unique terms in our dataset.

Table 16 Co-occurrence analysis results

Co-occurrence analysis	
Number of sentences with co-occurrences	1,414,380
Number of co-occurrences	12,616,465
Number of unique concepts	19,423

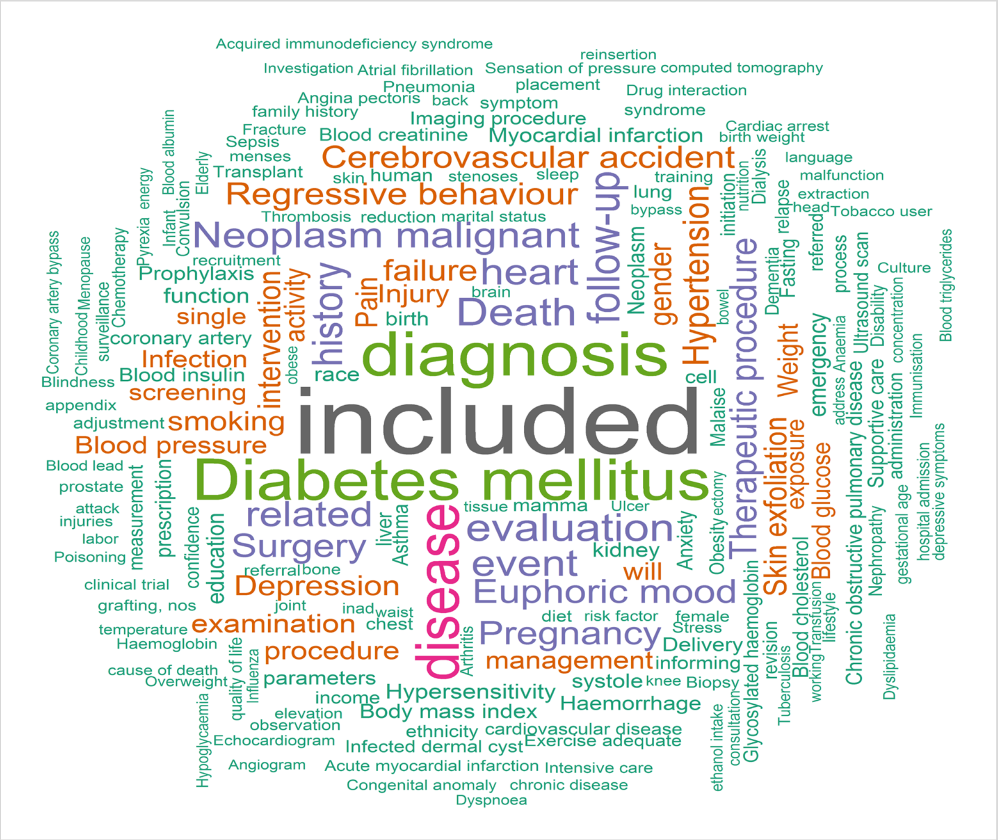


Figure 10 Word cloud based on the frequency of unique concepts (Table 16)

4.4.2 DICE ranking and network graphs of co-occurred terms

For each StartTerm (a phenotype of interest), we calculated the frequencies of co-occurrences with the StartTerm and DICE coefficient. If the DICE coefficient is 0, we eliminated these terms as they are not significant. For example, Table 17 shows the top 20 terms when using “Myopathy” as StartTerm. “Rhabdomyolysis” showed the highest score (DICE = 0.1948) to co-occur with “Myopathy”. The number of significantly co-occurred terms to “Myopathy” is 573. These terms are directly co-occurred in the same sentence. We further extended to indirect terms by utilizing the top 5% of the direct co-occurred terms, and it led to additional indirect 29 terms. Table 17 shows the top 20 co-occurred terms for Myopathy.

Co-occurred terms are visualized in network graphs. For example, Figure 11 shows the network for “Myopathy” that includes both direct and indirect phenotypes. For visualization purposes, we selected the top 69 co-occurred terms with “Myopathy”. Further, for the top 5% (n=29) indirect terms, we selected their top 10 co-occurred terms.

The largest node with the red color is the StartTerm, “Myopathy”. The bigger nodes have more edges than smaller nodes. For T2DM, Appendix 10 shows the top 20 co-occurred terms with T2DM and Appendix 11 shows T2DM network.

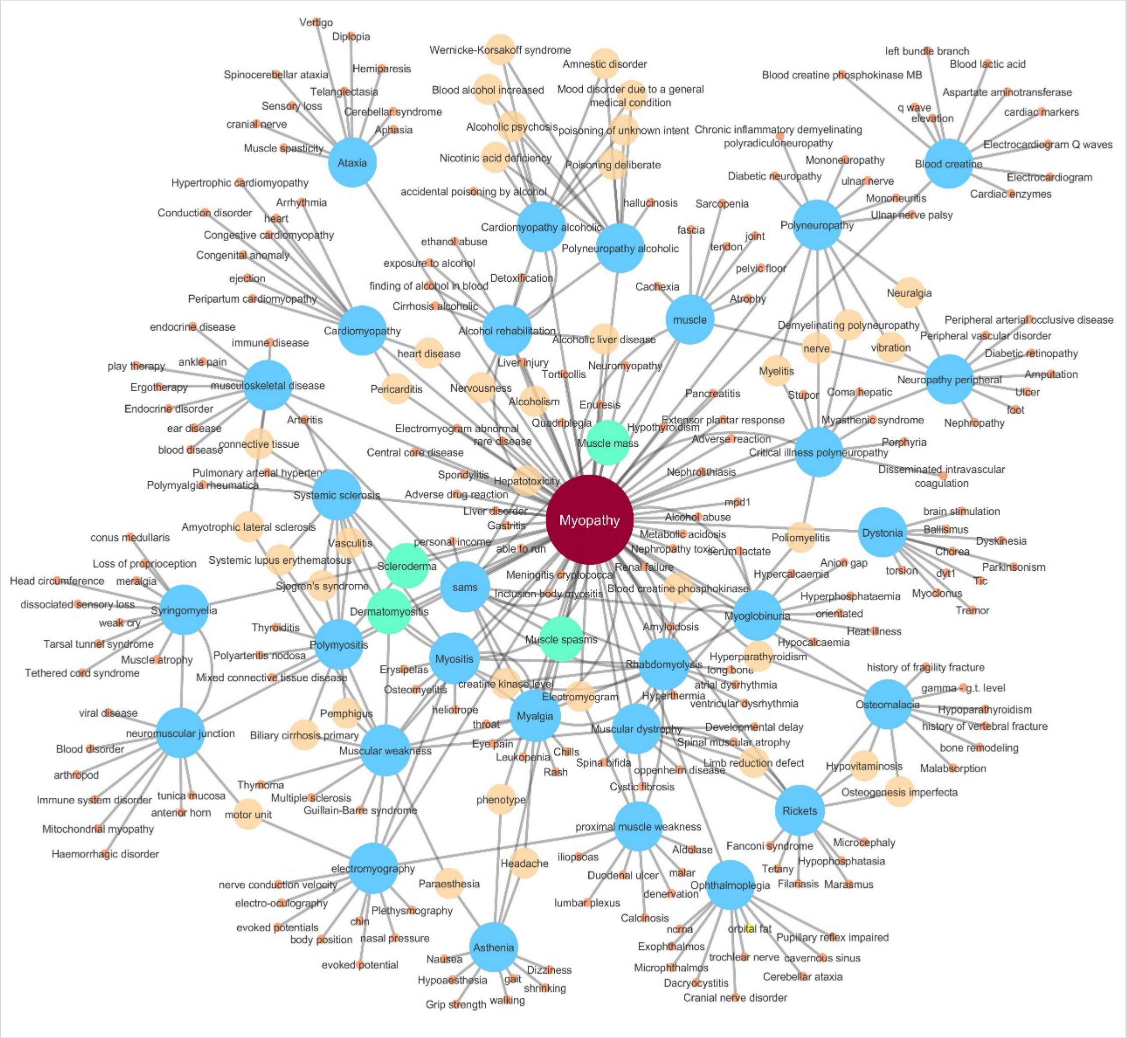


Figure 11 Myopathy Network (full dataset)

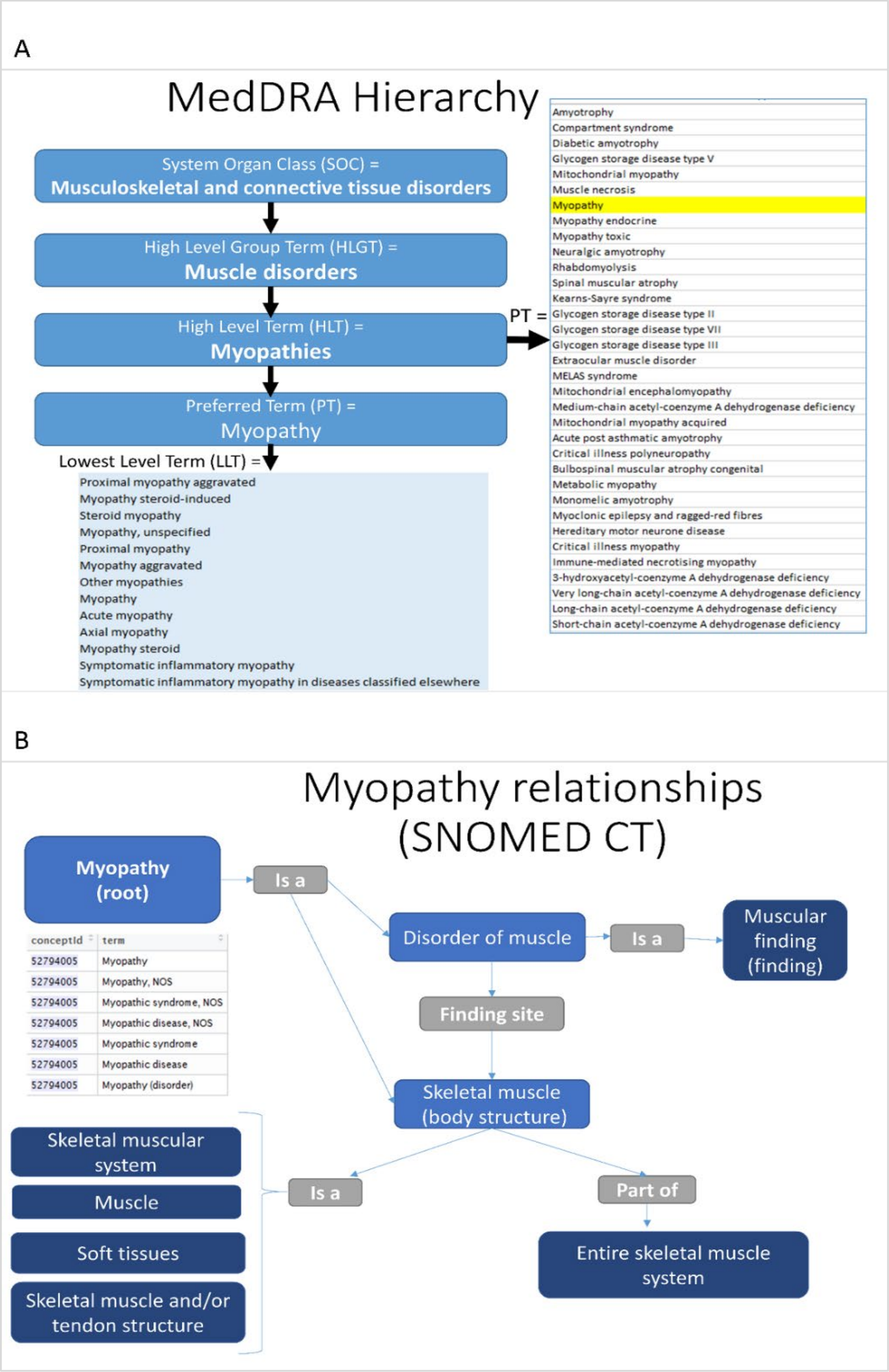
Table 17 Top 20 terms for co-occurred terms with Myopathy in the two divided datasets (50/50) and combined (full dataset)

Rank	Myopathy (Sample 1)			Myopathy (Sample 2)			Myopathy (Combined)		
	Concept	**Dice**	**Frequency**	**Concept**	**Dice**	**Frequency**	**Concept**	**DICE**	**Frequency**
1	Rhabdomyolysis	0.2188	36	Rhabdomyolysis	0.1640	21	Rhabdomyolysis	0.1948	57
2	Myalgia	0.0842	20	Myositis	0.0642	9	Myositis	0.0719	21
3	Myositis	0.0789	12	Myalgia	0.0456	9	Myalgia	0.0667	29
4	Muscular dystrophy	0.0592	8	Proximal muscle weakness	0.0370	3	Muscular dystrophy	0.0337	8
5	SAMS	0.0555	6	Electromyography	0.0246	3	SAMS	0.0315	6
6	Muscular weakness	0.0306	6	Dystonia	0.0238	3	Muscle	0.0268	43
7	Muscle	0.0297	25	Muscle	0.0236	18	Muscular weakness	0.0227	8
8	Critical illness polyneuropathy	0.0295	3	Cardiomyopathy alcoholic	0.0235	2	Critical illness polyneuropathy	0.0219	4
9	Polyneuropathy alcoholic	0.0294	3	Ophthalmoplegia	0.0235	2	Polyneuropathy alcoholic	0.0217	4
10	Neuropathy peripheral	0.0259	17	Syringomyelia	0.0229	2	Proximal muscle weakness	0.0217	4
11	Systemic sclerosis	0.0259	4	Rare disease	0.0218	3	Systemic sclerosis	0.0213	6
12	Polyneuropathy	0.0258	4	Electromyogram	0.0208	2	Cardiomyopathy alcoholic	0.0209	4
13	Motor end plate	0.0195	2	Torticollis	0.0208	2	Neuropathy peripheral	0.018	24
14	Myoglobinuria	0.0193	2	Meningitis cryptococcal	0.0206	2	Myoglobinuria	0.0163	3
15	Cardiomyopathy alcoholic	0.0188	2	Enuresis	0.0173	2	Polymyositis	0.0162	4
16	Liver injury	0.0164	3	Polymyositis	0.0170	2	Dystonia	0.0157	4
17	Polymyositis	0.0155	2	Systemic sclerosis	0.0156	2	Syringomyelia	0.0153	3
18	Liver disorder	0.0152	3	Ataxia	0.0155	2	Polyneuropathy	0.0143	4
19	Hepatotoxicity	0.0143	2	Pericarditis	0.0150	2	Cardiomyopathy	0.0122	10
20	Asthenia	0.0148	11	Asthenia	0.0148	11	Asthenia	0.0117	19

4.4.3 Evaluation and validation analysis of literature-based co-occurred phenotypes

To better demonstrate the clinical significance of the extracted terms, we performed following validation analyses. We compared the performance of co-occurrence approach for “Myopathy” by splitting the dataset into two 50/50 samples based on the number of PMIDs randomly. “Myopathy” co-occurred terms in each of the two samples were matched and we conducted Paired t-test to compare their DICE coefficient scores. The results showed no significant difference (t =1.036, df =566, p-value=0.301, 95% CI: -0.0003 - 0.0009) between the first and the second sample (mean of the difference = 0.0003). This suggests the consistency of results in smaller subsets of full-text articles. Three terms appeared in the first sample but not in the second: “Electromyogram abnormal” (DICE=0.01), “History of hepatocellular carcinoma” (DICE=0.009), and “MPD1” (DICE=0.009). On the other hand, three terms appeared in the second sample only: “Inclusion body myositis” (DICE=0.01), “able to run” (DICE=0.01), and “unable to run” (DICE=0.01). All of these 6 terms have co-occurred with “Myopathy” only one time, which explains why they did not appear in both subsets. Table 17 shows the top 20 terms in each sample.

In the second analysis, we compared the co-occurred terms with myopathy to the existing sources of terminologies, MedDRA and SNOMED-CT. Figure 12 (A & B) shows the representation of terms for “Myopathy” in MedDRA and SNOMED-CT, respectively. In MedDRA, for example, “Myopathy” is one of the preferred terms for (“Myopathies”, MedDRA HLT). On a lower level, “Myopathy” has 14 LLT terms that serve are synonyms for it. On the other hand, in SNOMED-CT, “Myopathy” is considered as the root term, which has around seven synonym terms. SNOMED-CT describes relationships where in Figure 12 B “Myopathy” is also described as “Is a” “Disorder of muscle” and “Is a” “Skeletal muscle (body structure)”. This example illustrates the classification structure for “Myopathy” in each standard. We note that in our NER tasks for recognizing terms, we used both dictionaries. Overall, these terms are already extended to the same levels as the examples in Figure 12. Literature-based phenotypes adds concepts associations (direct and in-direct) of other biomedical and procedure terms.


A
MedDRA Hierarchy
System Organ Class (SOC) =
Musculoskeletal and connective tissue disorders
High Level Group Term (HLGT) =
Muscle disorders
High Level Term (HLT) =
Myopathies
Preferred Term (PT) =
Myopathy
Lowest Level Term (LLT) =
PT =

Proximal myopathy aggravated
Myopathy steroid-induced
Steroid myopathy
Myopathy, unspecified
Proximal myopathy
Myopathy aggravated
Other myopathies
Myopathy
Acute myopathy
Axial myopathy
Myopathy steroid
Symptomatic inflammatory myopathy
Symptomatic inflammatory myopathy in diseases classified elsewhere

Amyotrophy
Compartment syndrome
Diabetic amyotrophy
Glycogen storage disease type V
Mitochondrial myopathy
Muscle necrosis
Myopathy
Myopathy endocrine
Myopathy toxic
Neuralgic amyotrophy
Rhabdomyolysis
Spinal muscular atrophy
Kearns-Sayre syndrome
Glycogen storage disease type II
Glycogen storage disease type VII
Glycogen storage disease type III
Extraocular muscle disorder
MELAS syndrome
Mitochondrial encephalomyopathy
Medium-chain acetyl-coenzyme A dehydrogenase deficiency
Mitochondrial myopathy acquired
Acute post asthmatic amyotrophy
Critical illness polyneuropathy
Bulbospinal muscular atrophy congenital
Metabolic myopathy
Monomelic amyotrophy
Myoclonic epilepsy and ragged-red fibres
Hereditary motor neurone disease
Critical illness myopathy
Immune-mediated necrotising myopathy
3-hydroxyacetyl-coenzyme A dehydrogenase deficiency
Very long-chain acetyl-coenzyme A dehydrogenase deficiency
Long-chain acetyl-coenzyme A dehydrogenase deficiency
Short-chain acetyl-coenzyme A dehydrogenase deficiency

conceptId	term
52794005	Myopathy
52794005	Myopathy, NOS
52794005	Myopathic syndrome, NOS
52794005	Myopathic disease, NOS
52794005	Myopathic syndrome
52794005	Myopathic disease
52794005	Myopathy (disorder)

Figure 12 Myopathy in MedDRA and SNOMED CT

The third analysis was to compare our literature-based co-occurred phenotypes to the phenotypes in PheKB and UpToDate. Table 18 shows the total count of terms for each phenotype in each of the sources. If we treat the total terms from three databases as total information, our literature based co-occurred phenotypes always possess the most the information, i.e. they have much more terms than the other two phenotype definition sources. For each of these ADEs, we provide the Venn diagram (Figure 13) for the three sources, the top terms for each ADE, and the network graph.

Table 18 The number of terms from text sources (Literature, PheKB, UpToDate) in ten selected phenotypes

Phenotypes of Interest	**Total count of terms**			**Total combined**
	Literature (%)	**PheKB (%)**	**UpToDate (%)**	
Diabetes Mellitus Type 2 (T2DM)	508 (63.8%)	106 (13.3%)	498 (62.6%)	796
Acute coronary syndrome	1203 (96.5%)	36 (2.9%)	197 (15.8%)	1247
Aneurysm	1487 (89.4%)	120 (7.2%)	432 (26.0%)	1664
Arthritis	2776 (93.9%)	169 (5.7%)	539 (18.2%)	2956
Cardiac failure	412 (58.0%)	119 (16.8%)	406 (57.2%)	710
Cough	2299 (93.9%)	31 (1.3%)	526 (21.5%)	2449
Dementia	2484 (90.5%)	38 (1.4%)	793 (28.9%)	2744
High-Density Lipoprotein (HDL) decreased	405 (78.9%)	33 (6.4%)	197 (38.4%)	513
Hypercholesterolaemia	1307 (87.7%)	340 (22.8%)	147 (9.9%)	1491
Hypothyroidism	1418 (81.8%)	137 (7.9%)	678 (39.1%)	1733

Finally, we manually compared T2DM phenotypes among literature-based discovery, PheKB and UpToDate. Figure 9 shows the data processing steps. All of the terms were extracted using our lexica (Table 15) and NER method. For literature-based, we used the DICE scores for T2DM co-occurred terms and their synonyms to recognize all sentences with presence of these terms. The DICE scores for each article's sentences were summed. The total number of articles that showed DICE score more than zero was 36,172 articles with a total of 903,120 sentences (Figure 9). Articles with higher DICE scores suggest that they contain more relevant T2DM definition-related sentences. For PheKB and UpToDate, the relevant T2DM texts were extracted and relevant terms were recognized using NER method.

Figure 13 shows a Venn diagram for the overlap of T2DM terms across the three sources. There were around 59 terms overlapped across the three sources. The overlap between Literature-based and PheKB was on 78 terms. Therefore, the literature-based T2DM covered 73.6% (n= 78 out of 106) of the PheKB terms. Then, we further looked into all of the terms (n = 796) from three sources combined. We evaluated each of the terms manually and assigned a category for it. Table 19 shows the categories, their counts in each of the sources, and examples for some overlapped terms.

Examples of risk factors that we found in our literature-based discovery, but not in PheKB and UpToDate are in Appendix 12. All of these concepts were direct relationship with T2DM, which means that they appeared in the same sentence. Appendix 12 shows examples of sentences for some T2DM risk factors or complications that appeared only in literature-based definitions. These sentences were selected from our dataset for illustration and might not have T2DM terms within the same sentence. However, these terms provide sentences with phenotyping definition information for phenotypes with relation to T2DM. Each of these concepts was mapped to its MedDRA PT.

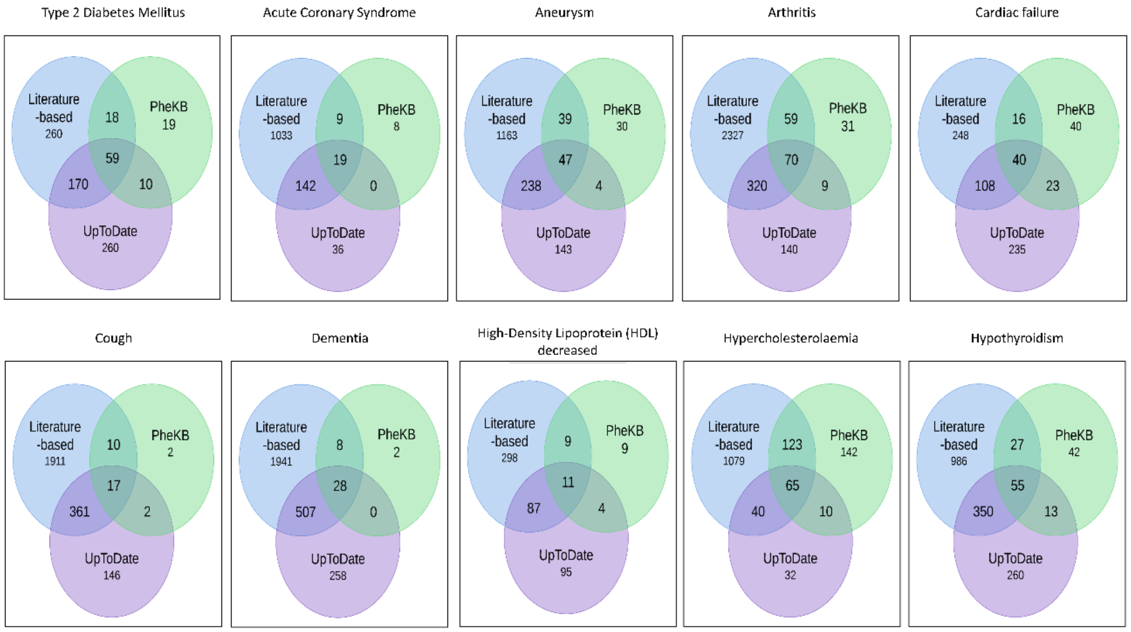


Figure 13 Venn diagrams of overlapping concepts for the 10 phenotype

Table 19 Categories of candidate phenotypes for Type 2 Diabetes Mellitus in the three sources

Type	**Description**	**Number of terms in each source**			**Number of terms appeared in the three sources**	**Terms appeared in the three sources**	
		Literature	**PheKB**	**UpToDate**		**Term**	**Term rank in literature co-occurrences (DICE score)**
Clinical diagnosis or symptom	Refers to an illness, condition, disease, disorder, or clinical features that describes patient's status to establish a clinical diagnosis. It can be also other related conditions, such as co-morbidities.	68 (13.4%)	11 (10.4%)	83 (16.6%)	6 (10.2%)	Type 2 diabetes mellitus	0
						Type 1 diabetes mellitus	1 (0.23267)
						Diabetes type	4 (0.01513)
						Hyperglycaemia	15 (0.00949)
						Diabetes mellitus	19 (0.00812)
						Glycosuria	Indirect relationship (T2DM | Gestational diabetes | Glycosuria)
Risk factor or complication	Factors that increase the risk of diabetes and reduce the risk of diabetes, or are diabetic complications.	225 (44.4%)	41 (38.7%)	199 (39.9%)	23 (39%)	Gestational diabetes	10 (0.01092)
						Diabetic retinopathy	41 (0.00494)
						Caucasian	43 (0.00475)
						Neuropathy peripheral	47 (0.00412)
						Family history	49 (0.00399)
						Retinopathy	55 (0.00339)
						Diabetic complication	82 (0.00317)
						Female	200 (0.00195)
						Nephropathy	255 (0.00136)
						Metabolic syndrome	265 (0.00131)
						Infected dermal cyst	281 (0.00115)
						Sensation of pressure	297 (0.00101)

						Pregnancy	314 (0.00085)
						Hypertension	355 (0.00064)
						Weight	366 (0.00057)
						Kidney	371 (0.00052)
						Smoking	374 (0.00052)
						Fall	401 (0.00043)
						Ethnicity	432 (0.00032)
						Race	435 (0.00031)
						Gender	439 (0.00029)
						Birth	449 (0.00023)
						Polycystic ovaries	Indirect relationship (T2DM | Gestational diabetes | Polycystic ovaries)
Measurements (e.g. laboratory tests)	Refers to the terms for clinical measurements, or laboratory tests or results.	38 (7.5%)	14 (13.2%)	58 (11.6%)	9 (15.2%)	Blood insulin	6 (0.0137)
						Drug tolerance	22 (0.0067)
						Blood glucose	46 (0.0041)
						Glycosylated haemoglobin	51 (0.0037)
						Fasting	122 (0.0029)
						Body mass index	289 (0.0010)
						Haemoglobin	316 (0.0008)
						Blood pressure	428 (0.0003)
						Glucose tolerance test	Indirect relationship (T2DM | Tolerance | Glucose tolerance test)
Procedure	Refers to the procedures that are performed as a part of a healthcare delivery, such as surgery.	33 (6.5%)	4 (3.8%)	16 (3.2%)	2 (3.4%)	Pancreatectomy	21 (0.00746)
						Therapeutic procedure	298 (0.00099)
Definition criteria	Terms that describe the medical entity within the phenotype definition, but not specific enough to be categorized as a diagnosis or a procedure	15 (3%)	8 (7.5%)	8 (1.6%)	6 (10.2%)	Diagnosis	266 (0.00131)
						Problem	319 (0.00082)
						History	359 (0.00062)
						Screening	370 (0.00054)
						Included	393 (0.00045)
						Measurement	394 (0.00045)

General / remove	Terms that are either too general for describing a phenotype or disease, or can be a noise.	128 (25.2%)	28 (26.4%)	135 (27.1%)	13 (22%)	Mass	165 (0.00233)
						Syndrome	177 (0.00222)
						Concentration	312 (0.00089)
						Related	324 (0.00079)
						Management	334 (0.00076)
						Education	341 (0.00073)
						Disease	346 (0.00068)
						Euphoric mood	384 (0.00049)
						Will	386 (0.00048)
						Single	434 (0.00032)
						Observation	442 (0.00025)
						Counseling	Indirect relationship (T2DM | supervision | counseling)
						Ovary	Indirect relationship (T2DM | Gestational diabetes | ovary)
Grand Total		507 (100%)	106 (100%)	499 (100%)	59 (100%)		

Terms with ranking are the terms that have direct relationship with T2DM. On the other hand, terms without ranking are the indirect relationships terms i.e. they did not co-occur in the same sentence with T2DM.

4.5 Discussion

4.5.1 Primary findings

The corpus of sentences predicted as positive (from Chapter 3) with evidence of phenotyping information were used for further discovery. Using these sentences, we used lexical-based named entity recognition to extract co-occurrences (n=12,616,465) and identify unique phenotypes (n=19,423). The statistical co-occurrence approach called DICE coefficient was used to rank co-occurrence concepts for a phenotype of interest. We used several approaches to validate the co-occurrences. First, for myopathy phenotype, we divided the dataset into 50/50 randomly, extracted co-occurrences concepts for each half with "Myopathy". We applied paired t-test on the co-occurrence concepts and it showed no significant difference between the two datasets. Second, we compared the co-occurrence concepts with myopathy from the whole dataset with other existing terminologies, such as MedDRA and SNOMED CT. According to our observations, existing terminologies do not provide the relationships that were observed in our literature-based phenotypes. We believe that literature-based concepts would generate unknown relationships for a phenotype of interest and serve better for the task of phenotyping and cohort identification. Third, literature-based phenotypes were further compared to phenotypes in other existing data sources: PheKB and UpToDate. When considering the total terms from three data sources combined as total information, literature-based co-occurred phenotypes always showed the most information across the ten tested phenotypes. Furthermore, we showed that T2DM concepts that were derived from literature-based co-occurrence analysis covered terms with 73.6% of PheKB. Finally, we manually evaluated T2DM concepts using raw text or online searches.

PheKB provides phenotyping definitions for several phenotypes through collaboration between institutions. The process of building these definitions are still manual, and relied on experts and multidisciplinary teams from several institutions. Since PheKB has expert-driven phenotype definitions, we further explored the missing terms by our method. For T2DM, we found 28 terms appeared in PheKB, but not in our literature concepts. We note that we are looking for exact concept match. Upon further manual assessment of the 28 terms contained within PheKB T2DM definition (Table 20), we found that 18 terms of them were directly relevant to the T2DM definition. We looked further in our literature concepts, we found that all of terms co-occurred with T2DM as direct relationships as single terms except one terms, which is “Diabetes with Hyperosmolarity” co-occurred as indirect relationship. The indirect relationship between T2DM and “Hyperosmolarity” means that they did not co-occur in the same sentence. Instead, T2DM co-occurred directly with a complication called “Ketoacidosis”, “Ketoacidosis” co-occurred directly with “Hyperosmolarity”, and the final indirect relationship is “T2DM| Ketoacidosis| Hyperosmolarity”. For the 17 terms that co-occurred with T2DM as direct relationships, these terms appeared as single terms. For example, “Cataract diabetic” appeared as “Cataract” and “Diabetic coma” appeared as “Coma” in which both terms in this example co-occurred with T2DM. For the 17 terms that co-occurred with T2DM as direct relationships, these terms appeared as single terms. For example, “Cataract diabetic” appeared as “Cataract” and “Diabetic coma” appeared as “Coma” in which both terms in this example co-occurred with T2DM. For the term “Diabetes mellitus uncontrolled”, even though its exact match is not in our literature-based terms, we found diabetic complications that can be caused by uncontrolled diabetes as single concepts. According to the American Diabetes Association (ADA)[1], uncontrolled diabetes can lead to several complications including foot complications, (Ketoacidosis) & Ketones, Kidney Disease (Nephropathy), High Blood Pressure, Stroke, and infections. These conditions appeared as single terms in our literature-based concepts that co-occurred with T2DM.

On the other hand, 7 terms were found in PheKB were not relevant directly to T2DM phenotyping definition, such as “Blue” and “Circling”. We found that these terms appeared out of context of T2DM definition such as a description of using blue highlight pen during the manual extraction process. Lastly, we found that “American Indian” appeared as a datatype in PheKB for manual data collection. “American Indian” did not appear in neither directly or indirectly co-occurred terms with T2DM in literature. Instead, we found the term “Race” co-occurred with T2DM, which is more generalized term. Overall, we found that most them co-occurred with T2DM as single terms. This shows that the concepts contained within T2DM phenotyping definitions-related sentences overlapped with most concepts contained within T2DM PheKB definition excluding the 7 terms (Table

[1] http://www.diabetes.org/living-with-diabetes/complications/

10) that are not relevant to phenotyping definitions as well as “Chart review”. Table 10 shows the 28 concepts with their assessment for their presentation in PheKB. Table 20 shows the 28 concepts with their assessment for their presentation in PheKB.

Table 20 Missing terms in our literature-based concepts and existed in PheKB

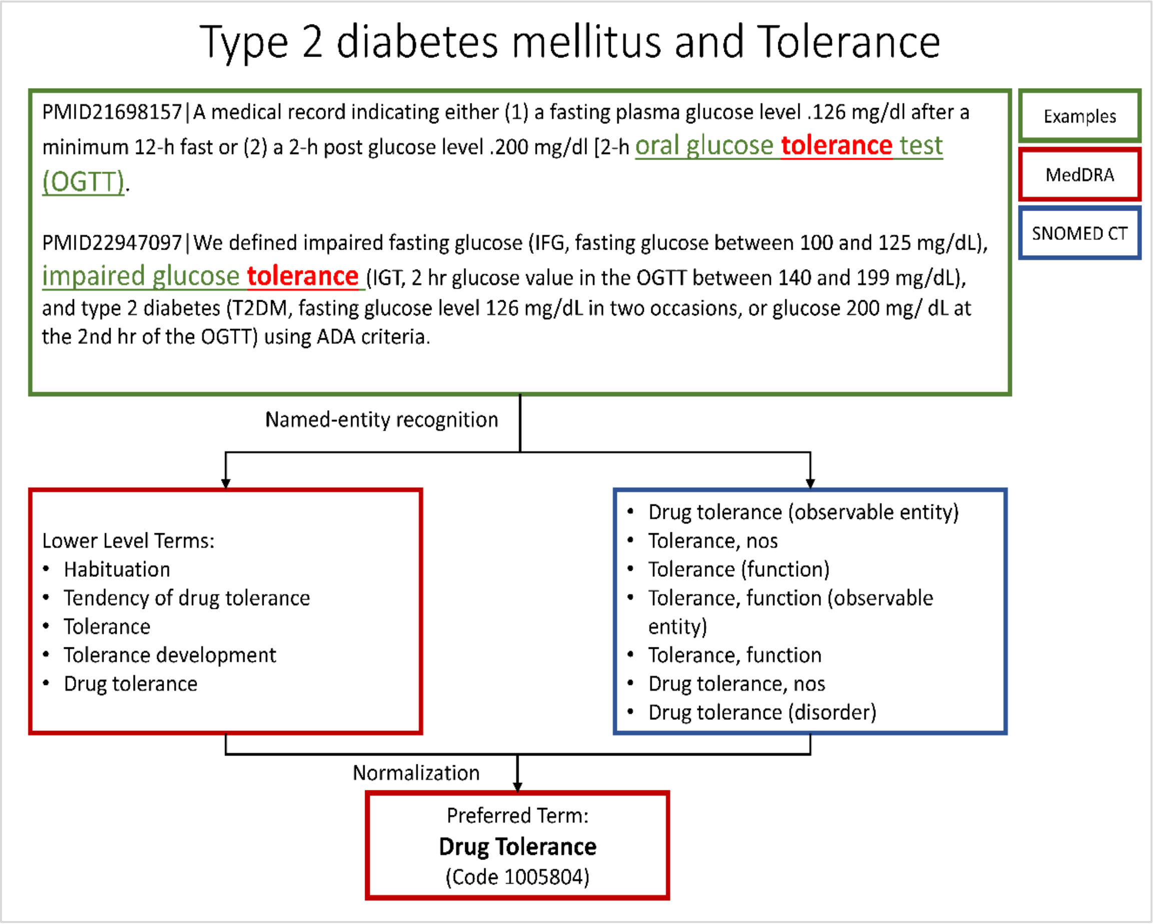
T2DM terms in PheKB, but not in literature-based co-occurrences	The presence of the term within PheKB phenotype definition	Reasons for missing the terms in T2DM literature-based co-occurrences
1. Cataract diabetic 2. Diabetes self management 3. Diabetes with hyperosmolarity 4. Diabetic arthropathy 5. Diabetic coma 6. Diabetic nephropathy 7. Diabetic neuropathy 8. Glycosylated hemoglobin measurement 9. Macular oedema 10. Neurological symptom 11. Familial risk factor 12. Family history of 13. Glucose measurement 14. Impaired fasting glucose 15. History of 16. Random blood glucose 17. Screening for diabetes 18. Tolerance test	Presented with T2DM definition/algorithm	These terms are presented in the literature as single terms that co-occurred with T2DM as either direct (n=17) or indirect relationship (n=1): ▪ “Cataract diabetic” appeared as “Cataract” ▪ “Diabetes self management” appeared as "Diabetes self-management" ▪ “Diabetes with hyperosmolarity” appeared as indirect relationship “T2DM | Ketoacidosis | hyperosmolarity”. ▪ “Diabetic arthropathy” appeared as “Arthropathy” ▪ “Diabetic coma” appeared as “coma” ▪ “Diabetic nephropathy” appeared as “Nephropathy” ▪ “Diabetic neuropathy” appeared as “Neuropathy peripheral” ▪ “Glycosylated hemoglobin measurement” appeared as “Glycosylated haemoglobin” and “measurement” ▪ “Macular oedema” appeared as “Oedema” ▪ “Neurological symptom” appeared as “Neuropathy peripheral” ▪ “Familial risk factor” appeared as “family history” and “family medical history” ▪ “Family history of” appeared as “family history” and “family medical history” ▪ “Glucose measurement” appeared as “measurement” and “Blood glucose” ▪ “Impaired fasting glucose” appeared as “Fasting” and “Blood glucose” ▪ “History of” appeared as “history" and "family history ▪ “Random blood glucose” appeared as “Blood glucose” ▪ “Screening for diabetes” appeared as “screening”

		▪ “Tolerance test” appeared as “Glucose tolerance test”
19. Diabetes mellitus uncontrolled	Presented with T2DM definition/algorithm	“Diabetes mellitus uncontrolled” is not in literature-based concepts, but terms for diabetes complications are present.
20. American Indian	Presented with datatypes used for data extraction, not T2DM definition/algorithm itself.	“American Indiana” appeared in PheKB T2DM datatypes for extraction of patient’s data. In our literature, we found more generalized terms, such as “Ethnicity” and “Race, co-occurred with T2DM.
21. Chart review	Presented within the method the validation criteria, not T2DM definition/algorithm itself.	“Chart review” is not in our literature terms. This term is usually appear within information used for creating or building a gold standard.
22. Blue 23. Circling 24. Digit 25. Does not fall 26. Interested 27. Sign 28. Separated	Presented within other text. These terms are not contained within T2DM definition.	These terms were either noise or not relevant to T2DM phenotyping definition itself. In our manual evaluation, these terms were categorized as “general/remove”.

4.5.2 Limitations of the study

One limitation is the manual process of assigning categories for T2DM co-occurred concepts. T2DM was selected for manual validation of the co-occurred phenotypes because both PheKB and UpToDate provide sufficient information. For each candidate phenotype for T2DM, we manually reviewed the original text in each of the three sources. We identified six categories, and each of them was assigned to a concept. These concepts include diagnosis and/or symptom, risk factor, laboratory, procedure, definition criteria, and general and/or remove. We note that false associations, i.e. noise, we called it “general and/or remove” in literature-based and UpToDate is higher than in PheKB. A possible reason is that literature and UpToDate sources have more textual information than PheKB that has finalized & post-processed T2DM definition information. Besides, when some of the phenotype candidates were still ambitious, we consulted a physician with specialization in diabetes. She reviewed the co-occurred T2DM concepts, confirmed the clinical significance of some, and denied others, including the overlapping concepts in the three sources.

Another limitation is the NER matching that it may match shorter terms rather than longer ones (e.g. one word rather than two). It has been reported that vocabulary mapping can generate some error in creating cohorts (Hripcsak, Levine, Shang, & Ryan, 2018), such as mapping to preferred terms or string matching. There is no standard or agreement on the best method for normalization, but dictionary-based methods are the best (Botsis & Ball, 2013). However, for example, we found that the preferred term 'Drug tolerance' is the MedDRA PT for the term 'Tolerance' (see Figure 14). In this example, the correct matched concept is "oral glucose tolerance test". The normalization of terms helped in increasing the weight of some of the co-occurrences by extending the synonyms of a phenotype. There are still some other challenges when using lexicon based methods such as describing the occurrence of hypokalemia or hyperglycemia in quantitative terms, such as changes in potassium or sugar levels (Iyer et al., 2014). Our network graphs helped to improve our understanding about the relations between concepts. It also helps in observing the patterns and clusters of phenotypes and the in-direct phenotypes, which provides a vision of unknown relationships. This was a more technical issue that can be addressed with advanced NLP approaches such as dependency approaches (Abacha & Zweigenbaum, 2011) as a future work.



Type 2 diabetes mellitus and Tolerance

PMID21698157|A medical record indicating either (1) a fasting plasma glucose level .126 mg/dl after a minimum 12-h fast or (2) a 2-h post glucose level .200 mg/dl [2-h oral glucose tolerance test (OGTT).

PMID22947097|We defined impaired fasting glucose (IFG, fasting glucose between 100 and 125 mg/dL), impaired glucose tolerance (IGT, 2 hr glucose value in the OGTT between 140 and 199 mg/dL), and type 2 diabetes (T2DM, fasting glucose level 126 mg/dL in two occasions, or glucose 200 mg/ dL at the 2nd hr of the OGTT) using ADA criteria.

Examples

MedDRA

SNOMED CT

Named-entity recognition

Lower Level Terms:

- Habituation
- Tendency of drug tolerance
- Tolerance
- Tolerance development
- Drug tolerance

- Drug tolerance (observable entity)
- Tolerance, nos
- Tolerance (function)
- Tolerance, function (observable entity)
- Tolerance, function
- Drug tolerance, nos
- Drug tolerance (disorder)

Normalization

Preferred Term:
Drug Tolerance
(Code 1005804)

Figure 14 Named-entity recognition for “Tolerance” and examples from literature sentences

4.5.3 Impact and future work

One of the major goals of this work is to decrease the need for human involvement during the process of developing a phenotyping definition. We used literature to derive evidence that supports information extraction of these definitions. We showed that we were able to decrease the expert involvement during this process. A researcher or an expert role can come later by either selecting a subset of candidate phenotypes or using all of them. This replaces the need to manually search for this information to define these terms in literature and medical guidelines sources. Besides, our data-driven approach provides less bias criteria for selection of phenotypes in comparison to expert involvement that their prior knowledge and experience might reflect their selection. We believe that utilizing this domain-specific corpus of sentences (from Chapter 3) with evidence of a phenotyping definition information helped in generating more accurate co-occurrence concepts. Such evidences were not built on a single definition, but on a large-scale set of data. We were able to extract over 12 million co-occurrences that the more recurrence of a co-occurrence means the highest its significance (Krenn, 2000). This provides research-based evidence to promote certain science and to derive consistent and generalizable findings built across several studies (Greenhalgh). We believe that our approach will support the generation and advancement of phenotyping definitions that were not represented previously in other sources, such as PheKB. In addition, it will support developing machine learning algorithms for automatic identification of cohorts of patients. With this, the goal is to help to transform the data to answer different research questions because different studies require different questions, and consequently designs (Yadav et al., 2018).

One of the advantages of utilizing literature-based definitions is the availability of definitions that were already used in research studies; in comparison, UpToDate is only providing diagnosis descriptions and guidelines developed primarily for clinical use. In addition, PheKB did a great achievement in creating a collaborative environment for developing, disseminating, and validating phenotyping definitions; however, it does not provide the definitions for all phenotypes of interest. In fact, the already existing definitions might not be efficient on all research purposes (R. Richesson et al.)[1]. Moreover, the designation of medical knowledge mapping, such as co-occurrences of terms extended to common dictionaries, can support phenotyping. For example, a patient with rheumatoid arthritis with an elevated liver function test might indicate either an ADE or a result of viral infection, heart failure, sepsis, or other causes (Mo et al., 2015). In this example, single phenotypes are not sufficient to understand or identify the cohort, but rather evidence from other supportive sources is required to define a phenotype. The terms contained within literature-based phenotyping definitions is capable of providing not only phenotype synonyms, but also other terms with some correlation evidence in the literature, such as risk factors and complications. These provide more flexibility for the user in designing the study of interest. For example, infectious disease is one of the candidate phenotypes with T2DM, but not in other sources (PheKB and UpToDate). Studies showed that there are

[1] http://rethinkingclinicaltrials.org/resources/ehr-phenotyping/

associations between infectious diseases and diabetes (Casqueiro, Casqueiro, & Alves, 2012; Shah & Hux, 2003). Furthermore, the procedure of splenectomy is also shown in our phenotypes, but not in other sources. We found that there is literature-evidence that splenectomy has an association with diabetes (S. C. Wu, Fu, Muo, & Chang, 2014). More examples are shown in Appendix 12.

In conclusion, data-driven approaches were used for extracting and ranking candidate phenotypes, including co-occurrence and network graphs, named-entity recognition, and DICE coefficients. Our main contribution is to decrease the human effort and involvement during the process of deriving phenotyping information from literature. Furthermore, our candidate concepts offer potential resource to support phenotyping and hypothesis generations, and open opportunities for EHR-based studies and validation. Overall, data-driven approaches are supportive for the areas of knowledge discovery of phenotyping definitions.

CHAPTER FIVE: DISCUSSION AND CONCLUSION

This dissertation presents an innovative informatics approach for mining phenotyping definitions in biomedical literature. Phenotyping definitions are often not available for many phenotypes of interest, especially when performing high-throughput phenotyping. A phenotyping case definition helps in the identification of cohorts of patients (Q. Li et al., 2014). We further discussed current approaches to develop a phenotyping definition including low-throughput phenotyping (expert-driven) and high-throughput phenotyping (data-driven). These methods can be time-consuming, labor-intensive, biased, and not scalable. Therefore, we developed a text-mining pipeline combining rule-based and machine-learning methods to automate retrieval, classification, and extraction of phenotyping definitions-related sentences from literature. To our knowledge, there is no existing work for mining literature-based phenotyping definitions using full texts on a large scale. We proposed three Aims to build our text mining and knowledge discovery approach for mining literature-based phenotyping definitions. In this chapter, we summarize major findings in each Aim of this dissertation, limitations of the study, and future work.

In Aim 1, we developed two corpora, abstracts and sentence-level full texts, as a first step for building a text-mining pipeline. Our selected phenotypes of interest were based on our research group's interests in adverse drug reaction phenotypes, and 279 phenotypes were selected. The list of phenotypes was used for several tasks, including data collection and lexica construction. For abstract-level corpus construction, two searching criteria were performed to retrieve abstracts relevant to EHR-based studies. A random set of these abstracts were selected that consists of 86 abstracts to build the full text corpus. Top phenotypes in these 86 abstracts are diabetes, hypertension, and heart failure. We downloaded their full texts, tokenized text into sentence tokens, and extracted sentences within methods sections boundaries. We proposed a new generalizable approach that serves as foundational basis for sentence-level annotations. The annotation guidelines aimed to annotate sentences that show contextual cues of a phenotyping definition (Botsis & Ball, 2013; Kirby et al., 2016; Shivade et al., 2014; Yadav et al., 2018) and PheKB modalities (Kirby et al., 2016), such as laboratories and standard codes. To our knowledge, contextual cues of phenotyping definitions in the literature that surround biomedical and medication entities were not studied previously. Two annotators with degrees in biomedical informatics have annotated the corpus. Several inter-annotator agreement measurements were used to assess the reliability of the annotations and guidelines. These measurements are overall sentence-level percent agreement (inspired by Wilbur et al. (Wilbur et al., 2006)), percent agreement, and Kappa. 3971 sentences were annotated. The overall sentence-level percent agreement was as high as 97.8%.

In Aim 2, we constructed a text-mining approach to automate extraction of phenotyping definitions' information. Two tasks were performed to accomplish this goal: information retrieval of abstracts and information extraction of sentences from methods sections of full texts. First, we used the 279 ADEs from Aim 1 to build our lexica and dictionary composed of 689,752 concepts that were used in several text-mining tasks. Second, we trained and validated two classifiers: abstract-level and full-text sentence-level. These classifiers utilized informatics approaches of text-mining, machine learning, and rule-based. For building the abstract-level classifier, we utilized a corpus of 799 positive abstracts (manually reviewed from Aim 1) and 1079 negative abstracts (randomly selected from PubMed between 1995 and 2017). We used WEKA to test and train the abstract-level classifier on several classification algorithms including sequential minimal optimization (SMO) (Platt, 1999), logistic regression (LR) (Quinlan, 2014)), Naïve Bayes (NB) (John & Langley, 1995), and J48 Decision Tree (Lecessie & Vanhouwelingen, 1992). The SMO and J48 Decision Tree algorithms outperformed the others, and their recall, precision, and F-measure were as high as 97%. For building the full-text sentence-level classifier, we used the corpus of 3971 sentences from Aim 1. NER and feature extraction (n=339) were performed followed by training and testing the classifier on SMO, LR, NB, and J48 Decision Tree algorithms. SMO and logistic regression showed the best performance, and their recall, precision, and F-measures were 84%. Both classifiers, the abstract-level (SMO algorithm) and the full-text sentence-level (LR algorithm), were used for predictions on large-scale literature text data. After optimizing the classifiers, we performed a large-scale screening of PubMed for years between 1975 to early March 2018. Using the abstract-level classifier, we predicted 459,406 abstracts as relevant to phenotyping. We retrieved their full texts, and our final set of full texts is 120,868. We processed the documents that were either PDF or XML formats into text format. Sentences within methods sections were extracted for predictions (n=6,129,574). We used the full-text sentence-level classifier on these sentences and were able to predict 2,745,416 sentences to be relevant to phenotyping. We believe that these sentences provide important phenotyping information, and were used for further knowledge discovery in Aim 3.

In Aim 3, we performed a discovery-based study to evaluate and validate literature-based phenotyping case definitions of selected phenotypes. We utilized sentences with phenotyping information from Aim 2 (n=2,745,416). Using lexical-based approaches we extracted concepts (n=19,423) and their co-occurrences in the same sentence (n=12,616,465). We used DICE coefficient scores to rank associated concepts with a phenotype of interest from the most significant to the least. We showed examples for myopathy and Type 2 Diabetes Mellitus (T2DM). Furthermore, we performed several validation tests. First, we compared the performance of co-occurrence approach for "Myopathy" by dividing the PMIDs into two subsets randomly. From each of the two subsets "Myopathy" co-occurred terms were extracted. We performed a paired t-test to compare the DICE coefficient scores between the two subsets, and the results showed no significant difference between them (p-value=0.301). Second, we compared candidate concepts for myopathy with concepts in MedDRA and SNOMED CT. We found that our candidate concepts provide additional information about myopathy phenotype such as risk factors and comorbidities. We note that we used both MedDRA and SNOMED CT for recognizing concepts in sentences. Third, we compared the candidate concepts for ten phenotypes in three resources: our literature-based results, PheKB, and UpToDate. We found that our literature-based phenotypes generally generated the largest number of concepts. We further manually reviewed T2DM candidate concepts for their clinical significance. We identified six categories that were each assigned to a T2DM concept, diagnosis and/or symptom, risk factor, laboratory, procedure, definition criteria and general and/or remove. Literature and UpToDate provided the most information about the risk factor category phenotypes followed by the diagnosis and/or symptom category. The diagnosis and/or symptom category was the highest in PheKB followed by the risk factor category.

In this work, we aimed to provide a scalable approach that is capable of deriving large number of concepts relevant to a disease, a phenotype, or an ADE with a minimum need to experts' involvement. We were able to collect large number of sentences containing phenotyping definitions information using text-mining and machine learning classifiers. The collection of sentences, if collected manually, can be time-consuming and labor-intensive. With this, we were able to extract over 2 million sentences that were predicted to contain phenotype definitions information. This collection of sentences can be used for several tasks, in future work, including information extraction and text summarization.

We utilized these sentences for extracting knowledge relevant to phenotyping definitions that uses data-driven approach. Unlike traditional methods, data mining and data-driven approaches provide new opportunities to use several data sources for knowledge discovery and identification of significant associations (R. Harpaz et al., 2012). We compared our data-driven approach to PheKB, which their phenotype definitions are considered expert-driven. Expert-driven approach, as we mentioned previously, requires, in many cases, experts collaboration and multidisciplinary teams involvement from one or multiple institutions. In addition, generating new definitions in PheKB is still a long process. To date, PheKB contains definitions for less than 65 phenotypes, which does not cover all phenotypes or ADEs of interest in high dimensions. Therefore, we identified the need of a more scalable approach that can accelerate the process of identifying concepts that are relevant to a phenotype. We further found that our literature results is capable of deriving most terms that were presented in PheKB. Additionally, our literature-based concepts are not limited to the expert knowledge, which can be sometimes bias, but are derived from evidence presented in literature supporting the goal of discovering unknown knowledge. Moreover, we believe that our literature-based concepts can provide phenotyping candidates in large numbers that supports high-dimensional research of ADEs and other phenotypes. This approach of extracting and ranking terms from full texts showed that we were able to present terms and concepts that are related to a phenotype or ADE of interest. The evidence used is the co-occurrence of the concepts within the same sentence that we called direct relationships. We further built and extended relationships of our network to the terms that did not co-occur within the same sentence and we called it indirect relationships. Our literature-based concepts included not only phenotypes, but also others such as procedures, definition criteria, and laboratories. With this, we are able to generate candidate concepts for any phenotype or ADE of interest if its data is contained in our literature-based concepts. These phenotype candidates facilitate standardized development of definitions using common terminologies. In addition, they can provide potential lists of concepts and relationships that can be later filtered according to research needs. The corpus of sentences combined with our candidate concepts can provide potential data sources for supporting EHR-based phenotyping research. A summarized overview of literature-based phenotyping definitions mining and knowledge discovery is shown in Figure 15.

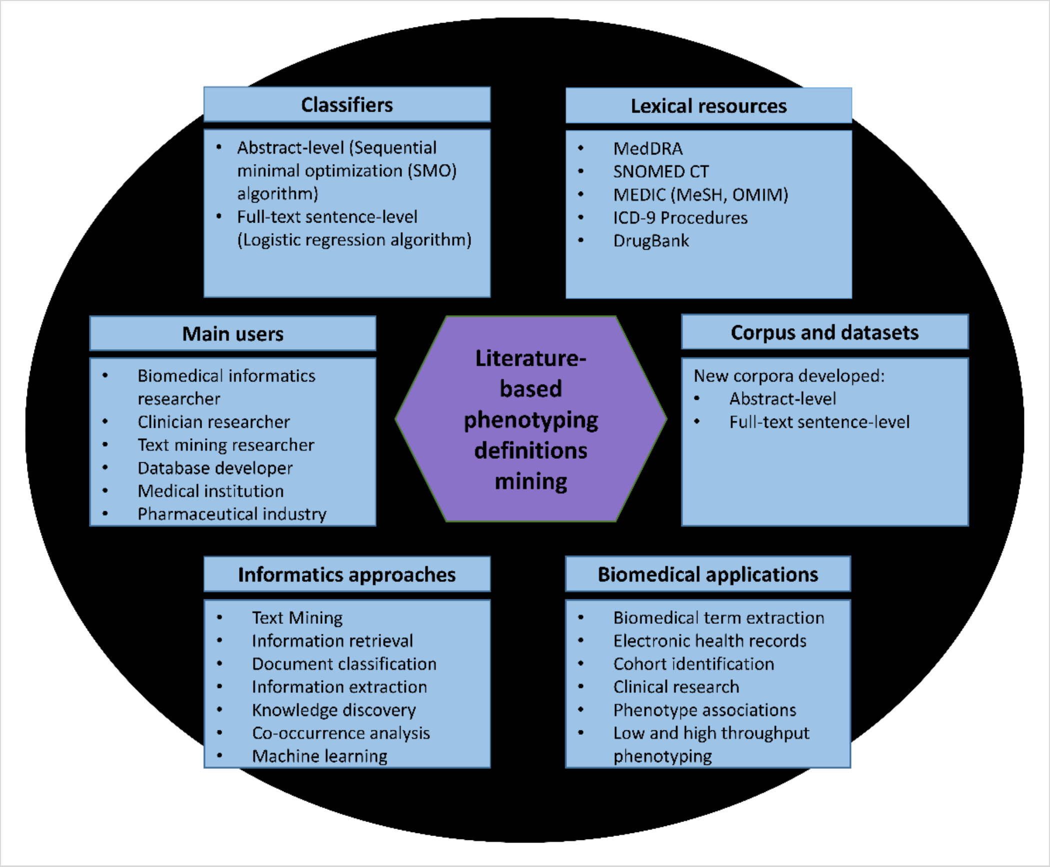


Figure 15 Overview of literature-based phenotyping definitions mining and knowledge discovery of this dissertation

This work does not stand without limitations. First, the annotation process is expensive, time-consuming, and labor-intensive. Therefore, only two annotators annotated the sentences of the corpus. Testing the guidelines on more annotators with clinical expertise is highly recommended. Second, a number of tasks were reliant on dictionary or lexical-based approaches such as feature extraction and co-occurrence extraction tasks. Such lexical-based approaches require frequent updates of the used dictionaries. Third, challenges accompanied with processing full texts are more than with abstracts. For example, texts converted from PDF documents had some issues with structure and unreadable special characters. Fourth, the manual validation process of the T2DM requires time and effort as well as expertise in several clinical specialties. Fifth, we note that one of the limitation of this study in our annotation criteria and text-mining classifier is not addressing the differentiation between the inclusion and exclusion criteria of a phenotyping definition. At this level of work in this dissertation, our goal was to extract all sentences that contain phenotyping definitions information. Future work can include further categorization and negation handling (J. J. Kim, Zhang, Park, & Ng, 2006) in order to differentiate between inclusion and exclusion sentences and terms contained within these sentences. Here are examples of definitions with inclusion and exclusion criteria, consecutively:

> “Those with specific diagnoses were included based on the following criteria: PA: diagnosis confirmed by pathological SIT i.e., PAC > 140 pmol/L post the infusion of 2 L of normal saline (0.9% NaCl) over 4 h” (PMID28924583), and “Patients with gestational diabetes mellitus, secondary diabetes (steroid-induced, cystic fibrosis, hemochromatosis, and chronic pancreatitis), or type 1 diabetes were excluded” (PMID25986070).

There are several opportunities for future work. We developed a corpus that can serve as a gold standard for future text-mining applications. In addition, the annotation guidelines can serve as a foundational basis for mining a phenotyping case definition and can be tested on a bigger corpus with more annotators. Furthermore, we recommend annotating entities including biomedical and phenotyping modalities. Such annotations can assist in several text-mining tasks, such as information extraction and summarization. For co-occurrence analysis, the candidate associations need some further validation and testing in other data sources such as EHR. These candidate concepts support several future research opportunities within areas of EHR-based research. For a phenotype of interest, indirect associations contribute in hypothesis generation and knowledge discovery.

Negative outcomes, such as drug-drug interactions (DDIs) and ADEs, have triggered the expansion of drug discovery research to detect relationships between drugs and outcomes at different levels (Zeng, Deng, Li, Naumann, & Luo, 2018). There are several data sources for mining ADEs such as literature, FAERS, social media, and EHR, that evaluation of ADEs can use more than one source (Tafti et al., 2017). Mining ADEs in EHR requires phenotyping definitions, especially when dealing with large number of ADE phenotypes. A phenotyping definition has several research applications, which include diagnosis categorization, novel phenotype discovery, clinical trial screening, pharmacogenomics, DDIs and ADEs, and genomic studies (Zeng et al., 2018). Therefore, our work has the potential to build a database for ADEs phenotyping definitions and their associated concepts that serve as dictionaries and potential related candidates. In addition, the collection of sentences can support the process of future text annotation and summarization of ADEs phenotype definitions from literature to build this database. Harmonization of the definitions in one source can help in a better understating of how an ADE has been defined previously across different studies in the literature. By creating such a source, ADEs phenotyping definitions information can be available to use for the EHR-based drug discovery research.

In conclusion, the contribution of this dissertation is in building specific corpus for mining a phenotyping definition and advancing knowledge about contextual cues surrounding these definitions. Abstract-level and full-text sentence-level classifiers were built to recall relevant sentences with phenotyping information. Furthermore, this work is different than previous work because it uses full texts rather than abstracts to represent co-occurrences of phenotypes. In addition, it used literature rather than EHR that suffers from bias. Validation of the co-occurrence candidates were performed with several methods. For empirical validation, text from different sources was used that differs in origin and style. For statistical validation, a paired t-test was used for comparing the co-occurrences derived from two subsets of data and showed no significant difference. Finally, this work is an effort to build scalable data-driven approach to represent computational phenotypes that can serve in several high-throughput phenotyping applications.

APPENDICES

Appendix 1

Phenotypes of interest: list of 279 potential adverse drug events (ADEs)

ADE-related phenotypes (1-150)	**MedDRA PT**	**ADE-related phenotypes (151-279)**	**MedDRA PT**
Abscess	10000269	Insomnia	10022437
Acne	10000496	Irritability	10022998
Acute coronary syndrome	10051592	Ischaemia	10061255
Affect lability	10054196	Ischaemic stroke	10061256
Aggression	10001488	Jaundice	10023126
Agitation	10001497	Lethargy	10024264
Akathisia	10001540	Leukocytosis	10024378
Alopecia	10001760	Leukocyturia	10050791
Anaemia	10002034	Leukopenia	10024384
Aneurysm	10002329	Lipoatrophy	10024604
Angina pectoris	10002383	Lipodystrophy acquired	10049287
Anxiety	10002855	Liver disorder	10024670
Anxiety disorder	10057666	Liver injury	10067125
Arrhythmia	10003119	Lung disorder	10025082
Arteriosclerosis	10003210	Lymphocele	10048642
Arthritis	10003246	Lymphoproliferative disorder	10061232
Asthenia	10003549	Malaise	10025482
Asthma	10003553	Mania	10026749
Ataxia	10003591	Menorrhagia	10027313
Atrial fibrillation	10003658	Methaemoglobinaemia	10027496
Atrioventricular block	10003671	Miosis	10027646
Atrioventricular block second degree	10003677	Mitochondrial toxicity	10053961
Azotaemia	10003885	Multi-organ failure	10028154
Back pain	10003988	Muscular weakness	10028372
Bipolar disorder	10057667	Musculoskeletal pain	10028391
Blood cholesterol increased	10005425	Musculoskeletal stiffness	10052904
Blood creatinine increased	10005483	Mutism	10028403
Blood pressure decreased	10005734	Myalgia	10028411
Bone marrow failure	10065553	Myocardial infarction	10028596
Bradycardia	10006093	Myocardial ischaemia	10028600
Bundle branch block left	10006580	Myoclonus	10028622
Cachexia	10006895	Myopathy	10028641
Cardiac arrest	10007515	Myositis	10028653
Cardiac failure	10007554	Nail disorder	10028694

Cardiac failure congestive	10007559	Nephrolithiasis	10029148
Cardiac fibrillation	10061592	Nephropathy	10029151
Cardiomegaly	10007632	Nephropathy toxic	10029155
Cardiotoxicity	10048610	Nephrotic syndrome	10029164
Cerebrovascular accident	10008190	Nervousness	10029216
Chest discomfort	10008469	Neuralgia	10029223
Chills	10008531	Neuropathy peripheral	10029331
Cholelithiasis	10008629	Neurotoxicity	10029350
Cholestasis	10008635	Neutropenia	10029354
Chronic allograft nephropathy	10063209	Nightmare	10029412
Cognitive disorder	10057668	Obsessive-compulsive disorder	10029898
Coma	10010071	Oedema	10030095
Completed suicide	10010144	Oliguria	10030302
Confusional state	10010305	Osteopenia	10049088
Constipation	10010774	Overdose	10033295
Convulsion	10010904	Pain	10033371
Coronary artery disease	10011078	Palpitations	10033557
Cough	10011224	Pancreatitis	10033645
Crying	10011469	Pancytopenia	10033661
Cyanosis	10011703	Panic attack	10033664
Delirium	10012218	Panic disorder	10033666
Delusion	10012239	Paraesthesia oral	10057372
Dementia	10012267	Parkinsonism	10034010
Depression	10012378	Peptic ulcer	10034341
Dermatitis	10012431	Peripheral sensory neuropathy	10034620
Diabetes mellitus	10012601	Peripheral vascular disorder	10034636
Diarrhoea	10012735	Pharyngitis	10034835
Dissociation	10013457	Poisoning	10061355
Dizziness	10013573	Polyuria	10036142
Drug intolerance	10061822	Poor quality sleep	10062519
Drug tolerance	10052804	Pregnancy	10036556
Drug tolerance decreased	10052805	Presyncope	10036653
Dry mouth	10013781	Productive cough	10036790
Duodenal ulcer	10013836	Proteinuria	10037032
Dysarthria	10013887	Prothrombin time prolonged	10037063
Dyslipidaemia	10058108	Pruritus	10037087
Dysphagia	10013950	Psoriasis	10037153
Dyspnoea	10013968	Psychosomatic disease	10049587
Dystonia	10013983	Psychotic disorder	10061920
Electrocardiogram qt interval	10014385	Pulmonary toxicity	10061924
Electrocardiogram qt prolonged	10014387	Pyelonephritis	10037596

Electrocardiogram st segment	10014389	Rash	10037844
Embolism	10061169	Renal failure	10038435
Epistaxis	10015090	Renal failure chronic	10038444
Erectile dysfunction	10061461	Renal impairment	10062237
Erythema	10015150	Renal tubular necrosis	10038540
Erythema multiforme	10015218	Restlessness	10038743
Essential hypertension	10015488	Rhabdomyolysis	10039020
Euphoric mood	10015535	Rhinitis	10039083
Extrapyramidal disorder	10015832	Salivary hypersecretion	10039424
Fatigue	10016256	Schizoaffective disorder	10039621
Fluid retention	10016807	Schizophrenia	10039626
Flushing	10016825	Sedation	10039897
Formication	10017062	Serotonin syndrome	10040108
Gait disturbance	10017577	Sexual dysfunction	10040477
Gastric ulcer	10017822	Shock	10040560
Gastrointestinal haemorrhage	10017955	Sinus bradycardia	10040741
Gastrooesophageal reflux disease	10017885	Sinusitis	10040753
Gingival hyperplasia	10018283	Skin toxicity	10059516
Glomerulonephritis	10018364	Sleep disorder	10040984
Glucose tolerance impaired	10018429	Social avoidant behaviour	10041243
Glycosuria	10018473	Somnolence	10041349
Gout	10018627	Stomatitis	10042128
Graft dysfunction	10059677	Stress	10042209
Graft loss	10048748	Sudden cardiac death	10049418
Graft versus host disease	10018651	Sudden death	10042434
Grand mal convulsion	10018659	Suicidal ideation	10042458
Gynaecomastia	10018800	Suicide attempt	10042464
Haematoma	10018852	Syncope	10042772
Haematuria	10018867	Tachycardia	10043071
Haemolysis	10018910	Tardive dyskinesia	10043118
Haemorrhage	10055798	Tension	10043268
Haemorrhagic diathesis	10062713	Thinking abnormal	10043431
Hallucination	10019063	Thrombocytopenia	10043554
Hemiparesis	10019465	Thrombosis	10043607
Hemiplegia	10019468	Thrombotic thrombocytopenic purpura	10043648
Hepatic cirrhosis	10019641	Torsade de pointes	10044066
Hepatic encephalopathy	10019660	Transaminases increased	10054889
Hepatic enzyme increased	10060795	Tremor	10044565
Hepatic failure	10019663	Type 2 diabetes mellitus	10067585
Hepatic function abnormal	10019670	Ulcer	10045285
Hepatic steatosis	10019708	Upper gastrointestinal haemorrhage	10046274

Hepatitis cholestatic	10019754	Urinary incontinence	10046543
Hepatotoxicity	10019851	Urticaria	10046735
High density lipoprotein decreased	10020060	Vasoconstriction	10047139
Hostility	10020400	Ventricular arrhythmia	10047281
Hot flush	10060800	Ventricular extrasystoles	10047289
Hyperbilirubinaemia	10020578	Ventricular failure	10060953
Hypercalcaemia	10020583	Ventricular fibrillation	10047290
Hyperchlorhydria	10020601	Ventricular tachycardia	10047302
Hypercholesterolaemia	10020603	Vision blurred	10047513
Hyperglycaemia	10020635	Visual impairment	10047571
Hyperhidrosis	10020642	Weight decreased	10047895
Hyperkalaemia	10020646	Weight increased	10047899
Hyperlipidaemia	10062060	Withdrawal syndrome	10048010
Hypersensitivity	10020751		
Hypertension	10020772		
Hyperthyroidism	10020850		
Hypertriglyceridaemia	10020869		
Hypertrophic cardiomyopathy	10020871		
Hyperuricaemia	10020903		
Hypoalbuminaemia	10020942		
Hypochondriasis	10020965		
Hypoglycaemia	10020993		
Hypokalaemia	10021015		
Hypomagnesaemia	10021027		
Hypomania	10021030		
Hyponatraemia	10021036		
Hypophosphataemia	10021058		
Hypoprothrombinaemia	10021085		
Hypotension	10021097		
Hypothyroidism	10021114		
Idiopathic thrombocytopenic purpura	10021245		
Immunodeficiency	10061598		
Incontinence	10021639		
Infarction	10061216		

Appendix 2

PMIDs selected by searching criteria explained in (Table 1 Abstract Inclusion-Exclusion criteria)

12952547
16765240
17456828
20112435
20819866
21051745
21156884
22071529
22737097
23449283
23471929
23574801
23740530
24377421
24780720
25024246
25567824
25725597
25827034
25991397
26524702
27112538
27969571
28081941
20362271
20504370
20976281
21182790
21722567
21727258
21862746
21931496
23193215
23873756
23913737
23940245
23969148
24177317
24297547
24349080
24636641
24658100
24734124
24882379
25104519
25431293
26167484
26209741
26365338
26370823
26725697
26961369
27151343
27621120
27749702
25851993
28222112

Appendix 3

Abstracts selected by using the other searching criteria (Not included Table 1 Abstract Inclusion-Exclusion criteria)

PMID	Search Criteria
11388131	hypertension electronic medical record diagnosis guideline
12461305	hypertension and "electronic medical record" and algorithm
15323063 22051424 17712071 24283597	hypertension electronic medical record (code OR retrospective)
15572716 26116311	myopathy electronic medical record
15758007 20655691	thrombosis electronic medical record
16567608 17162144 27252874 23439167 23445773	diabetes electronic medical record
17269833	electronic medical records adrenal cohort
17567225	electronic medical records heart failure
22466034	stroke electronic medical record (cohort* OR retrospective)
27940627	(cardiotoxicity OR cyanosis OR "peripheral vascular disorder" OR shock OR vasoconstriction OR "hypertrophic cardiomyopathy" OR "acute coronary syndrome" OR "angina pectoris" OR "cardiac arrest" OR infarction OR ischaemia OR" myocardial infarction" OR "myocardial ischaemia" OR "sudden cardiac death" OR arteriosclerosis OR "coronary artery disease" OR arrhythmia OR "atrial fibrillation" OR "atrioventricular block" OR "atrioventricular block second degree" OR bradycardia OR "bundle branch block left" OR "cardiac fibrillation" OR "electrocardiogram st segment" OR palpitation* OR presyncope OR "sinus bradycardia" OR syncope OR tachycardia OR "ventricular arrhythmia" OR "ventricular extrasystoles" OR "ventricular failure" OR "ventricular fibrillation" OR "ventricular tachycardia" OR "torsade de pointes" OR electrocardiogram qt interval OR" electrocardiogram qt prolonged" OR "cardiac failure" OR cardiac failure congestive OR cardiomegaly OR "blood pressure decreased" OR "blood pressure increased" OR thrombotic thrombocytopenic purpura OR "sudden death") electronic health record (cohort OR surveillance or case-control or epidemiological or Longitudinal Studies) (code or billing or algorithm)(code OR codes OR algorithm* or case definition)
23781409	(Diabetes type II) AND ("Electronic Health record" OR "Electronic Medical Record" OR "Electronic Health records" OR "Electronic Medical Records") AND validation
23968235	electronic health records and anemia
24303267	(diabetes mellitus OR glucose tolerance impaired OR glycosuria OR hyperglycaemia OR hypoglycaemia OR type 2 diabetes mellitus OR DM2) electronic health record

	(cohort OR surveillance or case-control or epidemiological or Longitudinal Studies) (code or billing or algorithm) (code OR codes OR algorithm* or case definition)
24507049	renal failure electronic medical record
25091637	myocardial infarction electronic medical record (retrospective OR cohort*)
25933736	electronic medical record Arrhythmia cohort
26221186	electronic medical record Arrhythmia algorithm
26283069	electronic health records and anemia and validation
27317850	(drug-induced OR "adverse events" OR "DDI" OR drug drug interaction OR "adverse reaction") electronic health record (cohort OR surveillance or case-control or epidemiological or Longitudinal Studies)
26082655	(asthenia OR chest discomfort OR chills OR dry mouth OR dysphagia OR fluid retention OR flushing OR formication OR haematoma OR hot flush OR hypercalcaemia OR hyperkalaemia OR hypokalaemia OR hypomagnesaemia OR hyponatraemia OR hypophosphataemia OR mitochondrial toxicity OR multi-organ failure OR oedema OR overdose OR pain OR poisoning) electronic health record (cohort OR surveillance or case-control or epidemiological or Longitudinal Studies) (code or billing or algorithm)(code OR codes OR algorithm* or case definition)

Appendix 4

Annotation guidelines to annotate a phenotyping definition in the literature

Inclusion category (five dimensions)

Inclusion dimension 1 – Biomedical & Procedure:

Criteria 1: The sentence should include two entities:

Rule for criteria 1: [Biomedical|Procedrue] AND [Definition criteria]

A. [Biomedical|Procedrue]: any biomedical or procedure terms, or any of the following: disease stages, symptoms, outcome of interest, diseases, laboratory & vital tests, diagnosis, procedure, clinical observation, person-time, Bed rest, Height, race, comorbidity, weight, sex or gender (Males, females, women, woman, man, men), hospitalization, birth date, surgery, chronic condition, BMI, age, ADE(drug-induced side effect), medication adherence, drug intolerance, cell level (gene/allele/SNPs/ homozygotes/dna).

B. [Definition criteria]: can be any of the following:
   - Verbs to define a phenotype: defined, identify, identified, included, excluded, calculate, having, undergoing, underwent, who had, documented, diagnosed, classified, consider, selected, counted, captured
   - Nouns to define a phenotype: inclusion, exclusion, definition, case identification, eligible, presence, criteria, Algorithm, diagnostic criteria, presence, absence, parameter, incident, sign, history, diagnosis, diagnoses, initiation, onset, occurrence, referral, guideline, category or categories, stage, outcome, outcome of interest, history, endpoint, examination, severity, adverse event, condition of interest
   - Phrase: “Patients/case/subject/child with”, “Patients/case/subject/child had”, “Patients/case/subject/child who”, “primary diagnosis”, “secondary diagnosis”, “primary procedure”, “secondary procedure”, “based on”, evidence of.

Criteria 2: Definitions in table, figure, box, or appendix. The sentence provide evidence of a phenotyping definition information presented in other sources, rather than text, such as table, figure, box, or appendix:

Rule for criteria 2: [Table terms] AND [Definition terms]

A. [Table terms]: Table, figure, Box, appendix.

B. [Definition terms]: inclusion, exclusion, definition, case identification, inclusion criteria, criteria, phenotyping algorithm, exclusion criteria.

Examples for inclusion dimension 1:

- “[identification] of [syndromic conditions]” (PMID17567225) – criteria 1
- “Such phenocopies [include] several [vasculitides, Buerger’s disease, embolism, trauma to leg arteries and other rare arteriopathies].” (PMID20819866) – criteria 1
- “We first [calculated] the prevalence of [prehypertension], [stage] 1 [hypertension], and stage 2 [hypertension] in the cohort.” (PMID17712071) – criteria 1 with categories or stages
- “The [categories] of [race] were ‘white’, ‘black or African American’, ‘American Indian or Alaskan’, ‘Asian’, ‘other’, and ‘unknown’” (PMID20819866) – criteria 1 with categories or stages
- “ Six comorbid disease conditions were selected and validated using the definitions reported in Table 3.” (PMID21051745) – criteria 2

- "The diagnoses are presented in hierarchical order in the first column of the Table." (PMID23449283) – criteria 2

Inclusion dimension 2 – Standard Codes:
Criteria 1: Mention of standard terms (e.g. International classification of diseases), such as ICD, CPT, UMLS, SNOMED, RxNorm, billing code, Read codes, diagnostic code¸ procedure code. Accepted formats: Short or long form (ICD or International Classification of Diseases) or list diagnostic or procedure codes.
Criteria 2: able terms with evidence of diagnostic or procedure codes list/ code definitions/algorithms: Table, figure, Box, appendix

Example for inclusion dimension 2:

- "a primary or any secondary discharge diagnosis (International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] code) of myoglobinuria (791.3)" (PMID15572716)

Inclusion dimension 3 – Medications:
Criteria 1: Keywords describe medications: e.g. generic drug names, prescribing, medication regimens, recommended agent, medication prescribed, drug dosage, drug frequency, drug route, medications, prescriptions.

Criteria 2: Drug name o-occurs with any of the following: medication, dose, treatment, therapy, drug, receiving, received, prescrib, using, use, use of, inclusion, include, exclusion, exclude, definition, case identification, identify, eligible, presence, criteria, presence, initiation, window, dose, guideline, history, started, agent(s), drug, medication, exposure, who had, treated with, indication, drug dosage, drug frequency, drug route, cohort.

Criteria 3: Medication terms co-occur with table, figure, box, or appendix: Table terms with evidence of a list of medication terms (Table, figure, Box, appendix).

Examples for inclusion dimension 3:

- "Other risk factors and comorbidities were ascertained based on ICD-9-CM codes, medication use and laboratory data." (PMID20819866) – criteria 1
- "Prior antihypertensive therapy was defined as the use of any AHDs before the start of amlodipine, which were not discontinued on or before the start of amlodipine therapy." (PMID15323063) – criteria 2
- "Table 1 outlines the recommended agents for specific comorbid conditions, as stated in our guideline." (PMID12952547)- criteria 3

Inclusion dimension 4 – Laboratories:
Criteria 1: The sentence should provide evidence of using clinical measurable values (i.e. laboratory values, vital values, procedures, clinical) combined with real values. The sentence should include [Clinical or procedure] AND [Measurable value]:

A. [Clinical or procedure]: Clinical can be any of the following: disease stages, symptoms, outcome of interest, diseases, laboratory & vital tests, diagnosis, procedure, clinical observation, Height, weight, BMI, age.
B. [Measurable value]: Any of the following:
   - Terms or numbers indicate measurable values: >, <, ≥, numerical values, more than, less than.
   - Other = ["mg", "ml", "mg/dL", "years old"]

- Other clue words combined with real values: value, measure, measurement, reference range, normal range, reading, level.

Example for inclusion dimension 4:

- “Achievement of lipid goals was defined as recommended by the National Cholesterol Education Program Adult Treatment Panel III guidelines16 as follows: LDL-C less than 100 mg/dL, triglyceride level less than 150 mg/dL, HDL-C greater than 40 mg/dL, and non–HDL-C less than 130 mg/dL.” (PMID16765240)

Inclusion dimension 5 – Use of Natural Language Processing (NLP):
Criteria 1: The sentence provides evidence of using NLP in a phenotyping definition.
Rule for criteria 1: [Phenotype, procedure, medication] AND [NLP terms]

A. [Phenotype, procedure, medication]: Terms can be any of the following: disease stages, symptoms, outcome of interest, diseases, diagnosis, procedure, clinical observation, drugs, medications. Other terms can be considered: person-time, Bed rest, Height, race, comorbidity, weight, sex, birth date, surgery, diseases, signs/symptoms, anatomical sites, procedure, drug, medication.
B. [NLP terms]: Natural Language processing, Natural language, nlp, text mining, "wildcard character", “bag of words”, parses, “named entity”, rule-based, NLP algorithm, n-grams, Regular Expression, tokenization, normalization, stemming, Lemmatization, named entity, named entity recognition (NER).

Criteria 2: NLP evidence in a phenotyping definition and this information explained in a table, figure, box, or appendix.
Rule for criteria 2: [NLP terms] AND [Table terms]

A. [NLP terms]: Natural Language processing, Natural language, nlp, text mining, "wildcard character", “bag of words”, parses, “named entity”, rule-based, NLP algorithm, n-grams, Regular Expression, tokenization, normalization, stemming, Lemmatization, Named entity, synonym, Named entity recognition (NER).
B. [Table terms]: Table, figure, Box, appendix.

Examples for inclusion dimension 5:

- “Example of a Clinical Note Represented as a “Bag of Words” . . . HF status positive negative Covariate #1 "heart" 3 1 Covariate #2 "pulmonary"” (PMID17567225) – criteria 1
- “Rule-based and machine learning techniques were applied to clinical narratives and smoking status was classified as ‘past’, ‘current’, ‘smoker’, ‘non-smoker’, or ‘unknown’.” (PMID20819866) – criteria 1
- “Structuring free text into useable coded data Text mining techniques were used to code diagnoses and drug prescriptions into ICD10 and ATC classification systems, respectively.” (PMID26209741) – criteria 1 medications
- “Details of text mining for identifying diagnoses are contained in a supplementary technical document” (PMID26209741) – criteria 2

Intermediate category (two dimensions)

Intermediate dimension 1 – Data sources:
Criteria 1: mention of used sources.

- Electronic health records keywords: Electronic health records, EHR, electronic medical record, EMR, database, registry, biobank, biospecimen, biorepositories

Criteria 2: mention of datatypes/variables used in EHR/EMR.

- Medical records keywords: progress notes, clinical notes/reports, laboratory records/data, radiology report/data, pharmacy records/data, administrative records/data, insurance claims/records/data, patient record, patient chart, hard copy report, medical chart, computerized charts.
- Clinical data keywords: referral, encounter, immunization, consultation report, laboratory, dismissal summaries, Height measurement, weight measurement, genetic data, serological data, problem list, scanned image, free-text, diagnoses list
- Procedure data keywords: claim, discharge, hospitalization, visit, admission, outpatient, inpatient, billing, hospital report, note.
- Other data mentions with evidence in EHR/EMR: demographic, sociodemographic, patient characteristics, abnormal measurement, follow-up data/measurement., encounter identifier.

Criteria 3: Any clinical or procedure followed with data keywords.
Rule for criteria 3: [Biomedical|Procedrue] AND [Data keywords]

A. [Biomedical|Procedrue]: any biomedical or procedure terms, or any of the following: disease stages, symptoms, outcome of interest, diseases, laboratory & vital tests, diagnosis, procedure, clinical observation, person-time, Bed rest, Height, race, comorbidity, weight, sex or gender (Males, females, women, woman, man, men), hospitalization, birth date, surgery, chronic condition, BMI, age, ADE(drug-induced side effect), medication adherence, drug intolerance, cell level (gene/allele/SNPs/ homozygotes/dna).
B. [Data keywords]: data, measures, measurement, value, values, datamart, dataset.

Examples for intermediate dimension 1:

- "Computerized medical and pharmacy records were reviewed for patient demographics, antihypertensive medications, comorbid conditions, and BP readings." (PMID11388131) – criteria 1 & 2
- "The data were all based on pharmacy claim records from the KP electronic prescription system." (PMID17269833) – criteria 1 & 2
- "This detailed information includes medical history, clinical assessments, consultation reports, dismissal summaries, laboratory and radiology results, and correspondence." (PMID17162144) – criteria 2

Intermediate dimension 2 – Study design or Institutional Review Board (IRB):
Criteria 1: Institutional Review Board (IRB) or Study design. If any of the following is in the sentence:

- Study design keywords: Retrospective, observational study, longitudinal study, case-control study, "random/ly sample/d", Inception cohorts, matched controls, matched cases, intervention group, control group, matched pairs, case-control pairs, cohort, negative cohort, positive cohort, pilot study, stratified, stratification, prospective, Surveillance Study, control.
- Gold standard keywords: chart review, manual review, notes reviewed, records reviewed, manual abstraction, expert panel, validation study, gold standard, standardized abstraction, standardized protocol.
- IRB keywords: IRB, Institutional Review Board, study protocol

Criteria 2: Any of the following co-occurrences the same sentence:

- "case" and "control"

- “chart” and “review”, “record” and “review”, or similar.

Examples for intermediate dimension 2s:

- “IRB approval The Institutional Review Board (IRB) at the Birmingham VA Medical Center approved this study.” (PMID24377421) – criteria 1
- “STUDY DESIGN: Retrospective chart review.” (PMID11388131) – criteria 1
- “To establish the control group, all active patients in the practice for less than 12 months were excluded.” (PMID11388131) – criteria 1
- “Chart review Confirmation of case status by manual review” (PMID12952547) – criteria 1

Exclusion category (three dimensions)

Exclusion dimension 1 – Irrelative evidence:
Criteria 1: Evidence of information relevant to other components of the study that might not assist in phenotyping. Each of the following sub-dimensions shows examples of keywords:

- Physical location (geographic) of the study only: information about the location (country, county, city, zip code, region, geographic).

  Note: We exclude from this criterion general location names because it can cause ambiguity with other institution names that are not physical location. Examples: institute, office, clinic, department…etc. In addition, we exclude from this criterion: if the state name is referring to the hospital.
- Ethical: consent, ethics, patient approval, patient denial, HIPAA.
- Financial: Funding, financial support, copayment, charged, sponsor, cost, insurance coverage, fee-per-service.
- Patient direct contact or enrollment: The sentence that shows evidence of a direct contact or enrollment of patients in the study. Example keywords: Surveys, questionnaire, interviews, instructions, recruitment, recruit, enrollment, enroll patients, 9-item Patient Health Questionnaire (PHQ-9)
- Providers & researchers: provider, physician, medical student, nurse, team, staff, clinician, fellow, --ologist, resident, general practitioner (GP), team, psychiatric, principal investigator, case manager.

  Note: We exclude from this criterion: author
- Performance: performance evaluation, training, performance measure, CPOE intervention, human error
- Quality of care: Quality of care, Quality Assurance, Quality Improvement

Examples for exclusion dimension 1:

- “Patients from the Department of Neurology, the Newborn Service, and the Neonatal Intensive Care Unit were excluded, as were patients receiving mechanical or pharmacologic prophylaxis.” (PMID15758007) – (Type: Location)
- “Reasons for exclusion were as follows: 6 persons denied permission to use their medical records for research” (PMID17162144) – (Type: Ethical)
- “All patients were members of the managed care system and incurred a significant financial advantage from having their prescriptions filled within the system.” (PMID16765240) – (Type: Financial)

- “Patients overdue for specific screening services received personalized letters recommending the needed service (e.g., cholesterol testing or dilated eye examinations) on a quarterly basis.” (PMID16567608) – (Type: Patient direct contact or enrollment)
- “Physicians received training on the use of the electronic medical record system and associated tools, such as reminders, from consultants working for the vendor company.” (PMID16567608) – (Type: Provider)
- “Physicians received training on the use of the electronic medical record system and associated tools, such as reminders, from consultants working for the vendor company.” (PMID16567608) – (Type: Performance)
- “Quality of care was determined by measuring the same parameters designed to measure the awareness, treatment, and control of hypertension.” (PMID12461305) – (Type: Quality of care)

Exclusion dimension 2 – Computational and statistical evidence:
Criteria 1: Evidence of information relevant to computational and statistical that might not assist in phenotyping. Each of the following sub-dimensions shows examples of keywords:

- Alerts: computer alerts, reminders, intranet tracking, continuously updated, robust, automated
- Software or tool: software, platform, plugin, computer, tool
- Statistical methods (usually toward the end of the method section). Any of the following statistical terms (or similar):
    - Analysis of covariance (ANCOVA)
    - Area under the receiver operating characteristic curves (AUC)
    - Bayes
    - Bias
    - Bivariate
    - Calculate (calculated, calculates, calculations)
    - Charlson’s comorbidity index
    - Chisquare
    - Chi-square
    - Cluster
    - Coefficient
    - Compute (computed, computes) - verb
    - Confidence interval, CI
    - Correlation
    - Covariance
    - Cox
    - Degrees of freedom
    - Descriptive statistics
    - Equation
    - Fisher exact test
    - Fisher’s exact
    - Fishers test
    - General linear model
    - Goodness-of-fit
    - Graphic
    - Imputat
    - Independent samples t-test
    - Measure (measured) - verb
    - Model (modeled, modeling)
    - Multiplication [x]
    - Multivariate
    - Normally distributed
    - Odds
    - Over-fitting
    - P value
    - Package
    - Permutation
    - Poisson distribution
    - Poisson regression
    - Predict
    - Predicted
    - Predictive value
    - Probability
    - Propensity score
    - R statistical language
    - Regression
    - Relative risk (rr)
    - Risk score
    - SAS
    - Sensitivity, specificity
    - Simulation
    - SPSS
    - Statistically significant

  - Kaplan-meier
  - Kolmogorov– smirnov
  - Likelihood
  - Logistic
  - Logistic regression
  - Mantel–Cox (log-rank)
  - Mean, median, mode
  - Statistics, statistical
  - T-test
  - Two-tailed
  - Univariate statistical analysis
  - Variance
  - Welch and Brown–Forsythe
  - Weighted
- Structured Query Language (SQL)

Examples for exclusion dimension 2:

- “We used logistic regression models with generalized estimating equations to adjust for race, year, race x year interactions, age, and sex.” (PMID16567608) – (Type: Statistics)
- “We used the proportional-hazards model to estimate the relative hazard of clinical end points associated with the computer alert and obtained 95 percent confidence intervals from this model.” (PMID15758007) – (Type: Alerts)

Exclusion dimension 3 – Insufficient evidence:
Criteria 1: Sentences with insufficient evidence. We mean by insufficient evidence is a sentence that does not met any of the dimensions in all categories (inclusion, intermediate, exclusion 1 & 2).

Example of exclusion dimension 3:

- “BSA= beclomethasone-salmeterol; COPD= chronic obstructive pulmonary disease; FSA= fluticasone-salmeterol; ICS= inhaled corticosteroid.” (PMID17162144)

Appendix 5

Entities and terms in the 86 abstracts using PubTator annotation tool

Entity	Term	Count of PMID
Chemical		**60**
	1RA	1
	alcohol	1
	aminosalicylates	1
	amlodipine	1
	amlodipine besylate	1
	atorvastatin	1
	beclomethasone	1
	bilirubin	1
	calcium	1
	cerivastatin	1
	cerivastatin-fibrate	1
	chloride	1
	cholesterol	6
	creatinine	3
	Cys	1
	cystatin C	1
	DVT	1
	fatty acid	1
	ferritin	1
	fibrate	1
	fluticasone	1
	gabapentin	1
	glucose	1
	Hg	1
	irbesartan	1
	iron	2
	lisinopril	1
	losartan	1
	N	1
	Neurontin	1
	olmesartan	1
	PIO	1
	potassium	1
	PPV	2
	pravastatin	1
	rivaroxaban	1

	SABA	1
	salmeterol	1
	serotonin	1
	simvastatin	1
	statin-fibrate	1
	statins	1
	steroid	1
	TGL	1
	thiopurines	1
	triamcinolone acetonide	1
	triglyceride	1
	triheptanoin	1
	uric acid	1
	valsartan	1
	venlafaxine	1
Disease		**264**
	AAA	1
	abdominal aortic aneurysm	1
	acute gout, chronic gout	1
	acute kidney injury	1
	acute liver failure	1
	acute myocardial infarction	1
	acute renal failure	2
	adult-onset asthma	1
	agranulocytosis	1
	AHDs	1
	ALD	1
	allergic reaction	1
	allergic reactions	1
	allergies	2
	allergy	1
	AMI	1
	anemia	1
	aneurysm	1
	anxiety	1
	anxiety symptoms	1
	AOA	1
	AOA to infection	1
	ARDS	1
	ARF	1
	ARI	1

	arrhythmia	1
	arthritis	1
	ASCVD	1
	aspiration	1
	asthma	3
	atherosclerotic	1
	Atrial Fibrillation	1
	beta-lactams	1
	bipolar disorder	1
	bleeding	3
	BP reduction	1
	breast cancer	1
	CAD	1
	cancer	3
	cancers	1
	cardiomyopathy	1
	cardiovascular disease	1
	Cardiovascular Health Study	1
	CAS	1
	catheter-directed thrombolysis	1
	CDT	1
	cellulitis	1
	cerebrovascular disease	1
	CHD	2
	CHF	1
	chronic disease	1
	chronic diseases	1
	chronic kidney disease	1
	chronic obstructive pulmonary disease	1
	CKD	3
	CLIA	1
	cognitive impairment	1
	congestive heart failure	1
	COPD	1
	coronary heart disease	3
	CRC	1
	Crohn disease	1
	CRT-D	1
	CVD	1
	death	2
	deep vein thrombosis	1

	deep-vein thrombosis	1
	defined as high blood pressure	1
	dementia	2
	depression	1
	Device failures	1
	diabetes	16
	diabetes care	1
	diabetes mellitus	8
	diabetic	3
	diabetics	1
	DM	4
	DVT	1
	ectopic pregnancies	1
	epilepsy	1
	GAD	1
	generalized anxiety disorder	1
	GI and other bleeding complications	1
	GI bleeds	1
	gout	2
	gout flares	1
	gout-related visits	1
	heart disease	1
	heart failure	7
	Hemorrhage	1
	HEP	1
	Hepatic encephalopathy	1
	hepatocellular carcinoma	1
	HF	2
	HH	1
	hip fractures	1
	HLD	1
	HLMs	1
	HSD	1
	HTN	1
	hyperkalemia	1
	hypertension	11
	hypertensive	3
	hypertensives	1
	hypoglycemia	1
	hyporesponsive	1
	hyporesponsiveness	1

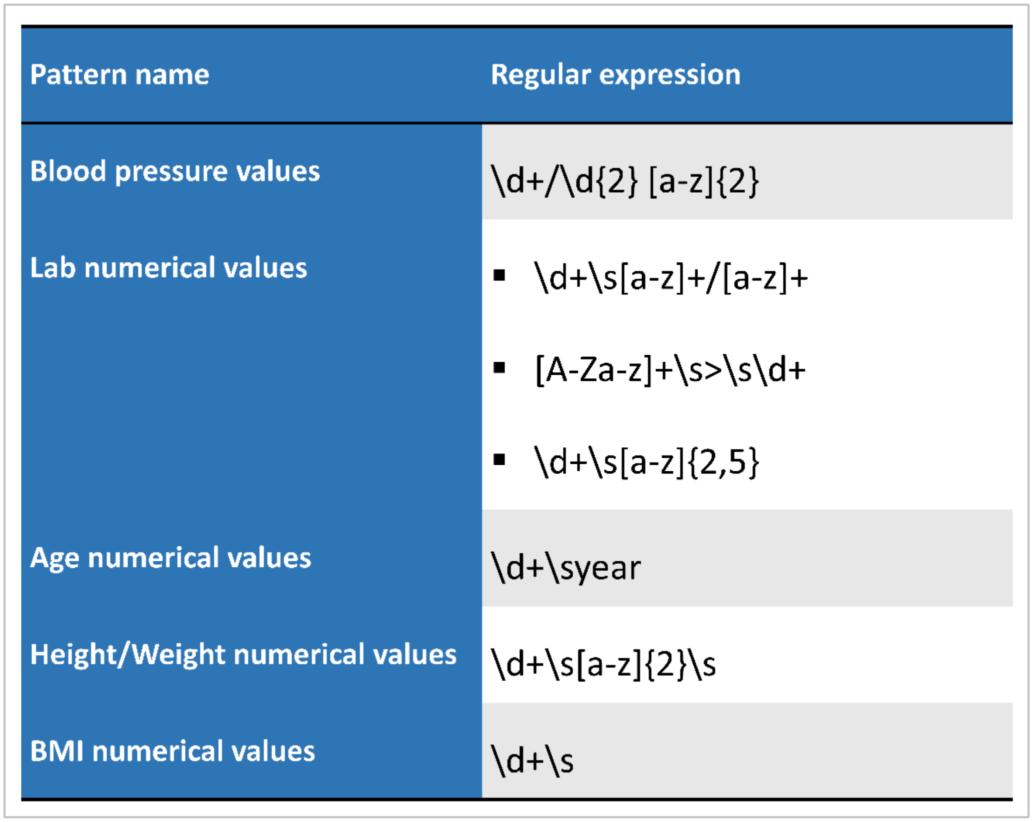
	IBD	1
	ICD	2
	ICS	1
	IDA	1
	IHD	1
	iliofemoral DVT	1
	incremental systolic BP reduction	1
	infection	1
	injury research	1
	injury type definitions	1
	injury types	1
	ischaemic heart disease	1
	LBBB	1
	LE PAD	1
	major bleeding	1
	MB	1
	MDD	2
	MELD	1
	mineral abnormalities	1
	Model for End-Stage Liver Disease	1
	myocardial infarction	2
	nephrolithiasis	1
	neutropenia	1
	neutrophilia	1
	NVAF	1
	obese	1
	obesity	3
	osteoarthritis	3
	PAD	2
	pain	1
	pancreatitis	1
	parkinsonism	1
	peripheral arterial disease	1
	pneumonia	3
	poisoning	1
	postoperative complications	1
	postoperative myocardial infarction	1
	prehypertension	1
	Preoperative anemia	1
	pulmonary embolism	2
	RBBB	1

	reduced kidney function	1
	rhabdomyolysis	2
	rheumatoid arthritis	2
	right bundle branch block	2
	rupture	1
	SCD	1
	Scotia	1
	sepsis	2
	SIRS	1
	SSS	1
	stroke	2
	systemic inflammatory response syndrome	1
	T2D	2
	TBI	1
	TBI-related condition	1
	thrombosis	1
	thrombus	1
	tophaceous gout	1
	trauma	1
	TSAT	2
	tumor	1
	type	3
	type 2 diabetes	3
	type of injury	1
	UC	1
	ulcer prophylaxis	1
	Ulcerative Colitis	1
	ULT	1
	urate-lowering therapy	1
	venous thromboembolism	1
	weight loss	1
	white-black disparity	1
Gene		**18**
	ACE	1
	Angiotensin-converting enzyme	1
	ARNO	1
	CLNK	1
	eGFR	1
	Epoetin	1
	HFE	1
	HSD	1

	K77	1
	KCNH2	1
	LDLR	1
	RYR2	1
	serotonin transporter	1
	transferrin	3
	zip	1
	3/5/2019	1
Mutation		**3**
	Cys282Tyr	1
	His63Asp	1
	p.Cys282Tyr	1

Appendix 6

Regular expressions for numerical patterns



Pattern name	Regular expression
Blood pressure values	\d+/\d{2} [a-z]{2}
Lab numerical values	▪ \d+\s[a-z]+/[a-z]+ ▪ [A-Za-z]+\s>\s\d+ ▪ \d+\s[a-z]{2,5}
Age numerical values	\d+\syear
Height/Weight numerical values	\d+\s[a-z]{2}\s
BMI numerical values	\d+\s

Appendix 7

Examples of features with sentences

Description	Semantic relationship	Example
Single features examples		
Patterns	Regular expression to capture lab values	"diabetes was diagnosed if a patient had fasting plasma glucose of 126 mg/dl or greater, or a random glucose greater than 200 mg/dl" (PMID20819866)
	Regular expression to capture BMI values	"Height and weight were used to calculate BMI, with BMI of 30 kg/m2 or greater defined as obese." (PMID17162144)
Single term (complete, stemmed, or multi-word)	defin	"Confirmed adult-onset asthma (AOA) cases were defined as those potential cases with either new-onset asthma or reactivated mild intermittent asthma that had been quiescent for at least one year." (PMID12952547)
	code	"Cerebrovascular disease was defined as the presence of ICD-9-CM diagnosis codes 430. X X -438. X X" (PMID20819866)
	history	"Cerebrovascular disease was defined as the presence of ICD-9-CM diagnosis codes 430. X X -438. X X or a history of carotid stenting or endarterectomy (ICD-9-CM procedure codes 00.61, 00.63, 38.10)." (PMID20819866)
Phrases	Evidence of	Controls were patients without evidence of PAD. (PMID20819866)
	Medical records	"We retrospectively reviewed the medical records to collect the following data: patient age, sex, smoking history, previous and current antihypertensive medications, history of intolerance to antihypertensive agents, comorbid conditions, and BP." (PMID11388131)
Medical NER presence – (Medical entities features)	ADE	"patients for hyperkalemia: (1) potassium value &gt;5.5 mmol/L; or (2) diagnosis code for hyperkalemia." (PMID17712071)
	CLINICAL	"Patient has heart disease diagnosis at any time." (PMID23449283)
	Procedure	"The primary endpoint was an asthma-related event (ARE), which was defined as (1) an emergency department (ED) visit or (2) hospital admission with a primary asthma diagnosis code ICD-9-CM code 493.xx." (PMID17269833)
Compound features (c-features) examples		
Two words co-occurrence	Definition terms + Medical NER Presence	"target BP was defined as systolic BP (CLINICAL)" (PMID11388131)
	Inclusion terms + Exclusion terms	"the inclusion and exclusion criteria of the clinical definition were mapped to a list of ICD9CM codes." (PMID27940627)

	Inclusion terms + Patient terms	"Patients with the presence of at least one diagnosis of major depression determined . . . from the EMR for inclusion in a data set (referred to as a data mart)." (PMID22466034)
Two words co-occurrence followed by abbreviation (with order)	Patient terms + with + abbreviation	"Patients with CAD (as defined by a history of myocardial infarction . . ." (PMID16765240)
Co-occurrence of terms preceded with abbreviation (with order)	Abbreviation + [diagno | event | disease | identif]	"We used the confirmed SCD diagnosis from Michigan NBS administrative records as the gold standard." (PMID24882379)
	Abbreviation + [positive | negative]	"HF positive vs HF negative." (PMID17567225)

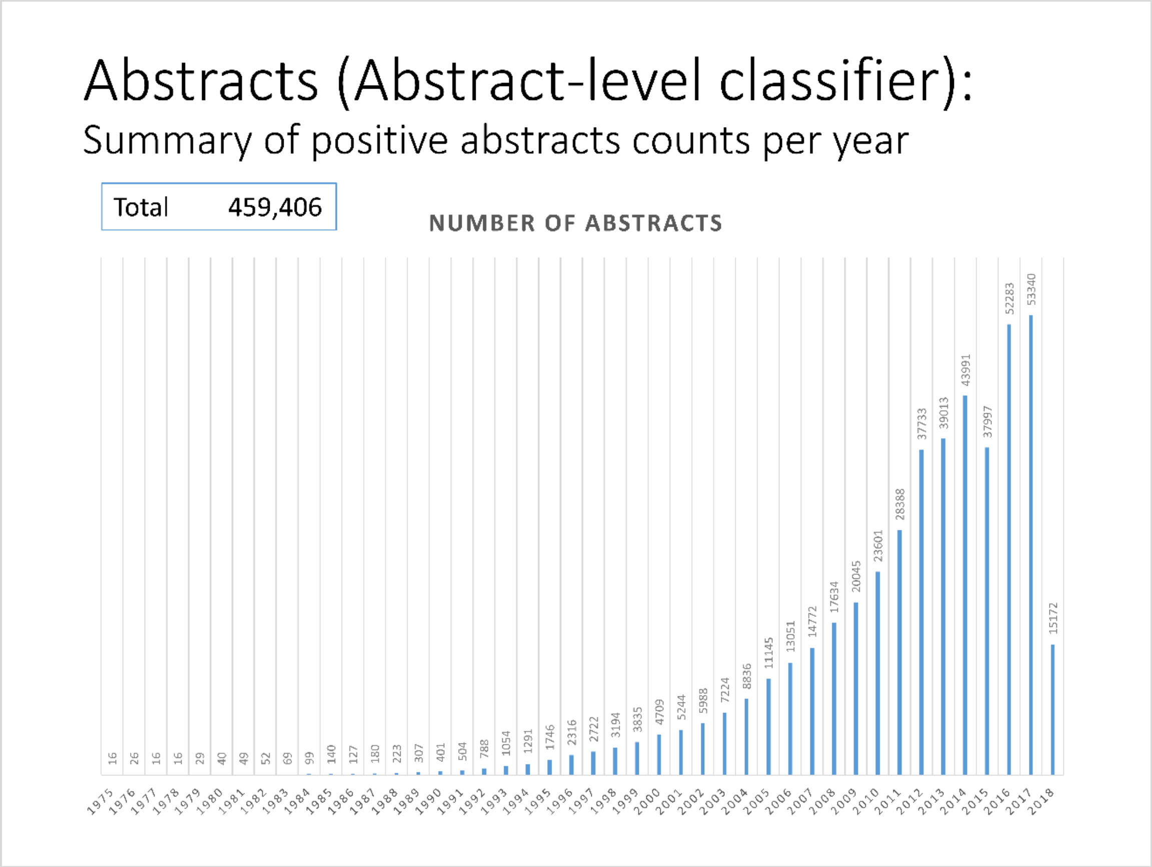
Appendix 8

IMRAD keywords used and rules

Keywords used to identify methods sections	
Section	Keywords
Relevant (Method section)	"Methodology", "METHODS", "Methods", "Method", "METHOD","DESIGN, SETTING, AND PATIENTS", "Design, Setting, and Participants", "Design", "DESIGN", "SETTING", "Setting", "SUBJECTS", "Materials and methods", "Materials and Methods", "Material and methods", "Patients and methods", "Participants and methods", "Experiment", "EXPERIMENT", "Subjects and methods", "Data source", "Research design and methods", u"Materials and methods", "Methods and Procedures", "Methods and Materials“
Not relevant	"Discussion", "DISCUSSION", "Findings", "Finding", "Result", "RESULT", "Results", "FINDING", "BACKGROUND", "Background", "Introduction", "INTRODUCTION", "IMPORTANCE", "Keywords", "Key Words:", "In conclusion", "Conclusion", "CONCLUSION", "REFERENCES", "COMMENT"]
Examples of rules used	
• A sentence starts with a keyword • "part 1: CHECK IF IT ENDS WITH S" • "part 2: CHECK IF IT FOLLWOED BY : OR ." • "part 3: CHECK IF IT FOLLWOED BY SPACE" • "part 4: CHECK IF IT ENDS WITH —" • Check if the following word is upper case or number • The position of the sentence by sent_index	

Appendix 9

Distribution of predicted positive abstracts between 1975 and 2018



Abstracts (Abstract-level classifier):

Summary of positive abstracts counts per year

Total 459,406

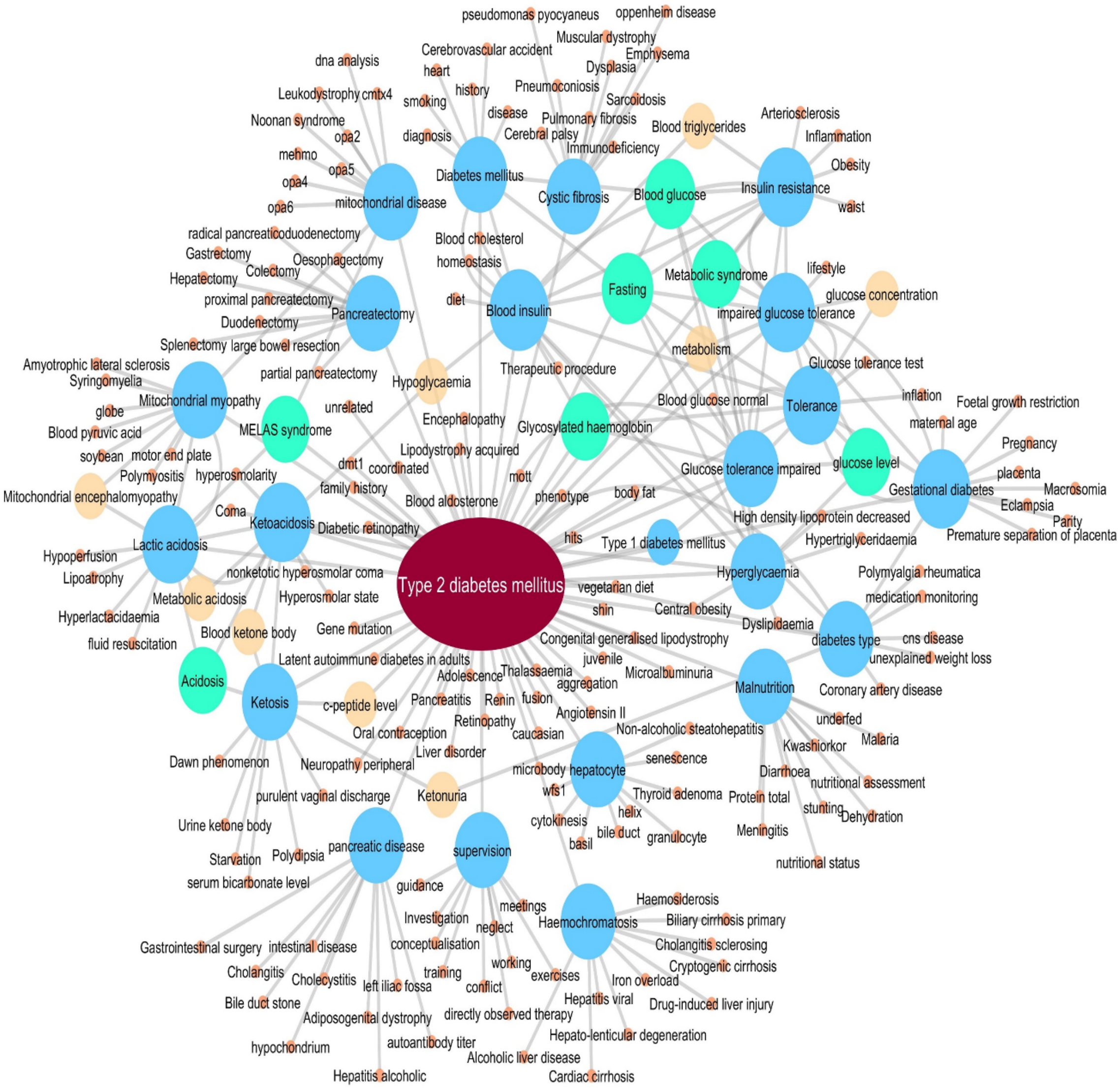
Appendix 10

Top 20 terms for type 2 diabetes mellitus (T2DM)

Rank	Type 2 diabetes mellitus (T2DM)		
	Concept	DICE	Frequency
1	Type 1 diabetes mellitus	0.2326	94
2	Glucose tolerance impaired	0.0171	17
3	Hepatocyte	0.0152	5
4	Diabetes type	0.0151	8
5	Mitochondrial disease	0.0148	5
6	Blood insulin	0.0137	113
7	Cystic fibrosis	0.0134	9
8	Ketoacidosis	0.0130	8
9	Impaired glucose tolerance	0.0115	9
10	Gestational diabetes	0.0109	12
11	Haemochromatosis	0.0105	4
12	Malnutrition	0.0103	10
13	Mitochondrial myopathy	0.0097	3
14	Shin	0.0096	4
15	Hyperglycaemia	0.0094	15
16	Pancreatic disease	0.0087	3
17	Insulin resistance	0.0083	10
18	Ketosis	0.0083	3
19	Diabetes mellitus	0.0081	419
20	Lactic acidosis	0.0079	3

Appendix 11

Type 2 diabetes mellitus graph



Appendix 12

Examples of risk factors or complications that appeared only in literature-based definitions

Category	literature-based concepts	Example of sentences with definitions	DICE for PMID
Diagnosis and/or symptom	Infectious disease	"Patients with inflammatory or infectious diseases, autoimmune and rheumatic diseases, cancer, haematological diseases and severe renal or liver failure, as well as those who were under treatment with anti-inflammatory drugs, were excluded." (PMID20836881)	8.278
Diagnosis and/or symptom	Rhinitis	"Vasomotor and allergic rhinitis…ICD-10 codes(J30.0-J30.4 R97)…Related ICPC-2E codes (R97)" (PMID27560181)	2.737
Diagnosis and/or symptom	Pancreatic disease	"we excluded patients with other kidney diseases such as … pancreatic disease, and psychopathy; with malignant tumors and secondary DM." (PMID23691167)	16.300
Procedure	Splenectomy	"Patients older than 20 years with splenic injury who underwent splenectomy (ICD-9-OP 41.5)…" (PMID25738485).	2.502
Laboratory	Microalbuminuria	"Persistent microalbuminuria was defined as a urinary albumin excretion of 30-300 mg/24 in at least two of three consecutive samples." (PMID24146865)	11.089
Laboratory	Blood bicarbonate	"Diabetic ketoacidosis (DKA) at diagnosis was reported for incident cases only and is based on having at least one of the following criteria noted in the medical record: 1) blood bicarbonate 15 mmol/l or pH 7.25 (venous) or 7.30 (arterial or capillary), 2) ICD-9 code 250.1 at discharge, or 3) diagnosis of DKA mentioned in the medical records." (PMID19246578)	10.606
Criteria	Clinical findings	"Patients with normal C-peptide levels, those who were considered to have maturity-onset diabetes of the young (MODY) based on the family history and clinical findings, those with T2DM, and those with a chronic disease (such as thalassemia, cystic fibrosis, drug-induced types) were excluded from the study."(PMID23419424)	11.916
Risk factor and/or complication	Renal failure	"For type 2 diabetes and HNF1Aand HNF4A-MODY, we tracked microvascular-related complications of blindness, renal failure, and amputation and macrovascular complications of angina, myocardial infarction, congestive heart failure, and stroke." (PMID24026547)	28.169
Risk factor and/or complication	Adolescence	"Little is known about the use of the A1C test for the diagnosis of type 2 diabetes and prediabetes in childhood and adolescence." (PMID21515842)	4.627

Risk factor and/or complication	Encephalopathy	“Subjects who tested positive for anti-glutamic acid decarboxylase (GAD) antibodies and those diagnosed with mitochondrial disease (mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes [MELAS]) or maturity-onset diabetes of the young (MODY) were not included.” (PMID23342076)	13.902
Risk factor and/or complication	Dermatophytosis	“This classification was used to create medical lists that enabled us to identify cases of three bacterial (i.e., septicemia, lower respiratory tract infection [LRTI], cutaneous cellulitis), two viral (i.e., herpes zoster, varicella), one parasitic (i.e., scabies), and two fungal (i.e., local candidiasis, dermatophytosis) infections recorded in the database (code lists available in S1 Table).” (PMID27218256)	0.328
Risk factor and/or complication	Dietary fibre intake	“Role of TCF7L2 risk variant and dietary fibre intake on incident type 2 diabetes.” (PMID22782288)	1.128
Risk factor and/or complication	Gout	“Gout was diagnosed according to the American College of Rheumatology 1977 criteria C. (PMID25031188)	5.078

REFERENCES

Abacha, A. B., & Zweigenbaum, P. (2011). *Medical entity recognition: a comparison of semantic and statistical methods*. Paper presented at the Proceedings of BioNLP 2011 Workshop, Portland, Oregon.

Agarwal, S., & Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics (Oxford, England), 25*(23), 3174-3180.

Agarwal, V., Podchiyska, T., Banda, J. M., Goel, V., Leung, T. I., Minty, E. P., . . . Shah, N. H. (2016). Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association, 23*(6), 1166-1173.

Alnazzawi, N., Thompson, P., & Ananiadou, S. (2014). *Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature.* Paper presented at the Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis, Gothenburg, Sweden.

Alnazzawi, N., Thompson, P., Batista-Navarro, R., & Ananiadou, S. (2015). Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Medical Informatics & Decision Making, 15 Suppl 2*, S3. doi:http://dx.doi.org/10.1186/1472-6947-15-S2-S3

American National Standards Institute., & Council of National Library and Information Associations (U.S.). (1979). *American national standard for writing abstracts*. New York: The Institute.

Ananiadou, S., Kell, D. B., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology, 24*(12), 571-579. doi:https://dx.doi.org/10.1016/j.tibtech.2006.10.002

Artstein, R. (2017). Inter-annotator Agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 297-313). Dordrecht: Springer Netherlands.

Banda, J. M., Callahan, A., Winnenburg, R., Strasberg, H. R., Cami, A., Reis, B. Y., . . . Shah, N. H. J. D. S. (2016). Feasibility of Prioritizing Drug–Drug-Event Associations Found in Electronic Health Records. *Drug Safety, 39*(1), 45-57. doi:10.1007/s40264-015-0352-2

Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B., & Shah, N. H. (2016). A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific data, 3*, 160026.

Banda, J. M., Seneviratne, M., Hernandez-Boussard, T., & Shah, N. H. (2018). Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annual Review of Biomedical Data Science, 1*(1), null. doi:10.1146/annurev-biodatasci-080917-013315

Barber, C., Lacaille, D., & Fortin, P. R. (2013). Systematic Review of Validation Studies of the Use of Administrative Data to Identify Serious Infections. *Arthritis Care Res, 65*(8), 1343-1357. doi:10.1002/acr.21959

Barber, C., Lacaille, D., & Fortin, P. R. (2013). Systematic review of validation studies of the use of administrative data to identify serious infections. *Arthritis Care Res (Hoboken), 65*(8), 1343-1357. doi:10.1002/acr.21959

Binkheder, S., Wu, H., Quinney, S., & Li, L. (2018, 4-7 June). *Analyzing Patterns of Literature-Based Phenotyping Definitions for Text Mining Applications.* Paper presented at the IEEE International Conference on Healthcare Informatics (ICHI).

Botsis, T., & Ball, R. (2013). Automating case definitions using literature-based reasoning. *Applied Clinical Informatics, 4*(4), 515-527.

Brown, E. G., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Safety, 20*(2), 109-117.

Carroll, R. J., Eyler, A. E., & Denny, J. C. (2011). Naive Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA ..* 189-196.

Casqueiro, J., Casqueiro, J., & Alves, C. (2012). Infections in patients with diabetes mellitus: A review of pathogenesis. *Indian J Endocrinol Metab, 16 Suppl 1*, S27-36. doi:10.4103/2230-8210.94253

Castro, V. M., Apperson, W. K., Gainer, V. S., Ananthakrishnan, A. N., Goodson, A. P., Wang, T. D., . . . Murphy, S. N. (2014). Evaluation of matched control algorithms in EHR-based phenotyping studies: a case study of inflammatory bowel disease comorbidities. *Journal of Biomedical Informatics, 52*, 105-111. doi:http://dx.doi.org/10.1016/j.jbi.2014.08.012

Chen, Y., Ghosh, J., Bejan, C. A., Gunter, C. A., Gupta, S., Kho, A., . . . Malin, B. (2015). Building bridges across electronic health record systems through inferred phenotypic topics. *Journal of Biomedical Informatics, 55*, 82-93. doi:http://dx.doi.org/10.1016/j.jbi.2015.03.011

Chiang, C.-W., Zhang, P., Wang, X., Wang, L., Zhang, S., Ning, X., . . . Li, L. (2018). Translational High-Dimensional Drug Interaction Discovery and Validation Using Health Record Databases and Pharmacokinetics Models. *Clinical pharmacology and therapeutics, 103*(2), 287-295.

Chiu, P. H., & Hripcsak, G. (2017). EHR-based phenotyping: Bulk learning and evaluation. *Journal of Biomedical Informatics, 70*, 35-51.

Christley, Y., Duffy, T., & Martin, C. R. (2012). A review of the definitional criteria for chronic fatigue syndrome. *Journal of evaluation in clinical practice, 18*(1), 25-31.

Chung, Y. M., & Lee, J. Y. (2001). A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology, 52*(4), 283-296. doi:Doi 10.1002/1532-2890(2000)9999:9999<::Aid-Asi1073>3.3.Co;2-X

Chute, C. G., Pathak, J., Savova, G. K., Bailey, K. R., Schor, M. I., Hart, L. A., . . . Huff, S. M. (2011). The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. *AMIA .. 2011*, 248-256.

Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., . . . Smalheiser, N. R. (2010). *Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools*. Paper presented at the Proceedings of the 1st ACM International Health Informatics Symposium, Arlington, VA.

Collier, N., Groza, T., Smedley, D., Robinson, P. N., Oellrich, A., & Rebholz-Schuhmann, D. (2015). PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. *Database (Oxford), 2015*. doi:10.1093/database/bav104

Conway, M., Berg, R. L., Carrell, D., Denny, J. C., Kho, A. N., Kullo, I. J., . . . Pathak, J. (2011). Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annual Symposium Proceedings, 2011*, 274-283.

Czaja, A. S., Ross, M. E., Liu, W., Fiks, A. G., Localio, R., Wasserman, R. C., . . . Adams, W. G. Electronic health record (EHR) based postmarketing surveillance of adverse events associated with pediatric off-label medication use: A case study of short-acting beta-2 agonists and arrhythmias. *Pharmacoepidemiology and Drug Safety, 0*(0). doi:doi:10.1002/pds.4562

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. *Int J Comput Appl, 28*(2), 37-40.

Daniel, C., & Choquet, R. (2014). Information technology for clinical, translational and comparative effectiveness research. Findings from the section clinical research informatics. *Yearb Med Inform, 9*, 224-227. doi:http://dx.doi.org/10.15265/IY-2014-0040

Davazdahemami, B., & Delen, D. (2018). A chronological pharmacovigilance network analytics approach for predicting adverse drug events. *Journal of the American Medical Informatics Association, 25*(10), 1311-1321. doi:10.1093/jamia/ocy097

Davis, A. P., Wiegers, T. C., Rosenstein, M. C., & Mattingly, C. J. (2012). MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database: The Journal of Biological Databases and Curation, 2012*, bar065.

de Bie, S., Coloma, P. M., Ferrajolo, C., Verhamme, K. M. C., Trifiro, G., Schuemie, M. J., . . . consortium, E.-A. (2015). The role of electronic healthcare record databases in paediatric drug safety surveillance: a retrospective cohort study. *British journal of clinical pharmacology, 80*(2), 304-314.

de Vries, C., & de Jong-van den Berg, L. (2001). 22 Pharmacovigilance and Pharmacoepidemiology In *Pharmacy practice* (pp. 367).

Declerck, G., Hussain, S., Daniel, C., Yuksel, M., Laleci, G. B., Twagirumukiza, M., & Jaulent, M. C. (2015). Bridging data models and terminologies to support adverse drug event reporting using EHR data. *Methods of information in medicine, 54*(1), 24-31.

Denny, J. C. (2012). Chapter 13: Mining electronic health records in the genomics era. *PLoS Computational Biology, 8*(12), e1002823. doi:http://dx.doi.org/10.1371/journal.pcbi.1002823

Dogan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics, 47*, 1-10. doi:https://dx.doi.org/10.1016/j.jbi.2013.12.006

Duke, J. D., Han, X., Wang, Z., Subhadarshini, A., Karnik, S. D., Li, X., . . . Li, L. (2012). Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Computational Biology, 8*(8), e1002614. doi:10.1371/journal.pcbi.1002614

Eriksson, R., Werge, T., Jensen, L. J., & Brunak, S. (2014). Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population. *Drug Saf, 37*(4), 237-247. doi:10.1007/s40264-014-0145-z

Evert, S. (2005). The statistics of word cooccurrences: word pairs and collocations.

Fiest, K. M., Jette, N., Quan, H., St Germaine-Smith, C., Metcalfe, A., Patten, S. B., & Beck, C. A. (2014). Systematic review and assessment of validated case definitions for depression in administrative data. *BMC psychiatry, 14*, 289.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr, W. (2011). Word co-occurrence features for text classification. *Information Systems, 36*(5), 843-858.

Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods (Duluth), 74*, 97-106. doi:10.1016/j.ymeth.2015.01.015

Fort, D., Wilcox, A. B., & Weng, C. (2014). Could Patient Self-reported Health Data Complement EHR for Phenotyping? *AMIA Annual Symposium Proceedings, 2014*, 1738-1747.

Fox, B. I., Hollingsworth, J. C., Gray, M. D., Hollingsworth, M. L., Gao, J., & Hansen, R. A. (2013). Developing an expert panel process to refine health outcome definitions in observational data. *Journal of biomedical informatics, 46*(5), 795-804.

Frey, L. J., Lenert, L., & Lopez-Campos, G. (2014). EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group. *Yearbook of Medical Informatics, 9*, 206-211. doi:http://dx.doi.org/10.15265/IY-2014-0006

Glicksberg, B. S., Miotto, R., Johnson, K. W., Shameer, K., Li, L., Chen, R., & Dudley, J. T. (2018). Automated disease cohort selection using word embeddings from Electronic Health Records. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 23*, 145-156.

Glueck, M., Hamilton, P., Chevalier, F., Breslav, S., Khan, A., Wigdor, D., & Brudno, M. (2016). PhenoBlocks: Phenotype Comparison Visualizations. *IEEE Transactions on Visualization & Computer Graphics, 22*(1), 101-110. doi:http://dx.doi.org/10.1109/TVCG.2015.2467733

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., . . . e, M. N. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Medicine, 15*(10), 761-771. doi:10.1038/gim.2013.72

Greenhalgh, T. How to implement evidence-based healthcare. In.

Gurulingappa, H., Klinger, R., Hofmann-Apitius, M., & Fluck, J. (2010). *An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature.* Paper presented at the 2nd Workshop on Building and evaluating resources for biomedical text mining (7th edition of the Language Resources and Evaluation Conference).

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics, 45*(5), 885-892. doi:10.1016/j.jbi.2012.04.008

Gurwitz, D., & Pirmohamed, M. (2010). Pharmacogenomics: the importance of accurate phenotypes. *Pharmacogenomics, 11*(4), 469-470. doi:10.2217/pgs.10.41

Halpern, Y., Choi, Y., Horng, S., & Sontag, D. (2014). Using Anchors to Estimate Clinical State without Labeled Data. *AMIA Annual Symposium Proceedings, 2014*, 606-615.

Hansen, R. A., Gray, M. D., Fox, B. I., Hollingsworth, J. C., Gao, J., & Zeng, P. (2013). How well do various health outcome definitions identify appropriate cases in observational studies? *Drug safety, 36 Suppl 1*, S27-32.

Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther, 91*(6), 1010-1021. doi:10.1038/clpt.2012.50

Harpaz, R., Haerian, K., Chase, H. S., & Friedman, C. (2010). *Mining electronic health records for adverse drug effects using regression based methods*. Paper presented at the Proceedings of the 1st ACM International Health Informatics Symposium, Arlington, Virginia, USA.

Harpaz, R., Vilar, S., Dumouchel, W., Salmasian, H., Haerian, K., Shah, N. H., . . . Friedman, C. (2013). Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association, 20*(3), 413-419. doi:10.1136/amiajnl-2012-000930

Henderson, J., Bridges, R., Ho, J. C., Wallace, B. C., & Ghosh, J. (2017). PheKnow-Cloud: A Tool for Evaluating High-Throughput Phenotype Candidates using Online Medical Literature. *AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science, 2017*, 149-157.

Henry, S., McQuilkin, A., & McInnes, B. T. (2018). Association measures for estimating semantic similarity and relatedness between biomedical concepts. *Artificial Intelligence in Medicine*. doi:10.1016/j.artmed.2018.08.006

Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., & Sun, J. (2014). Limestone: high-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics, 52*, 199-211. doi:http://dx.doi.org/10.1016/j.jbi.2014.07.001

Homsted, L. (2000). Institute of Medicine report: to err is human: building a safer health care system. *The Florida nurse, 48*(1), 6.

Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association, 20*(1), 117-121. doi:http://dx.doi.org/10.1136/amiajnl-2012-001145

Hripcsak, G., & Albers, D. J. (2017). High-fidelity phenotyping: richness and freedom from bias. *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocx110

Hripcsak, G., Levine, M. E., Shang, N., & Ryan, P. B. (2018). Effect of vocabulary mapping for conditions on phenotype cohorts. *Journal of the American Medical Informatics Association, 25*(12), 1618-1625. doi:10.1093/jamia/ocy124

Hsu, J., Pacheco, J. A., Stevens, W. W., Smith, M. E., & Avila, P. C. (2014). Accuracy of phenotyping chronic rhinosinusitis in the electronic health record. *American Journal of Rhinology & Allergy, 28*(2), 140-144. doi:http://dx.doi.org/10.2500/ajra.2014.28.4012

Iyer, S. V., Harpaz, R., LePendu, P., Bauer-Mehren, A., & Shah, N. H. (2014). Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association, 21*(2), 353-362. doi:10.1136/amiajnl-2013-001612

John, G. H., & Langley, P. (1995). *Estimating continuous distributions in Bayesian classifiers*. Paper presented at the Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Canada.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology, 1*(1), 4-20.

Kilicoglu, H., Rosemblat, G., Malicki, M., & ter Riet, G. (2018). Automatic recognition of self-acknowledged limitations in clinical research literature. *Journal of the American Medical Informatics Association, 25*(7), 855-861. doi:10.1093/jamia/ocy038

Kim, J.-D., Ohta, T., & Tsujii, J. i. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics, 9*, 10.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics, 19 Suppl 1*, i180-182.

Kim, J. J., Zhang, Z., Park, J. C., & Ng, S. K. (2006). BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. *Bioinformatics, 22*(5), 597-605. doi:https://dx.doi.org/10.1093/bioinformatics/btk016

Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics, 12 Suppl 2*, S5. doi:10.1186/1471-2105-12-S2-S5

Kirby, J. C., Speltz, P., Rasmussen, L. V., Basford, M., Gottesman, O., Peissig, P. L., . . . Denny, J. C. (2016). PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association, 23*(6), 1046-1052. doi:10.1093/jamia/ocv202

Kolesnikova, O. (2016). Survey of Word Co-occurrence Measures for Collocation Detection. *Computacion Y Sistemas, 20*(3), 327-344. doi:10.13053/CyS-20-3-2456

Kotfila, C., & Uzuner, O. (2015). A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. *Journal of Biomedical Informatics, 58 Suppl*, S92-S102. doi:http://dx.doi.org/10.1016/j.jbi.2015.07.016

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., . . . Valencia, A. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics, 7*(Suppl 1 Text mining for chemistry and the CHEMDNER track), S2. doi:10.1186/1758-2946-7-S1-S2

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology, 9 Suppl 2*, S8. doi:10.1186/gb-2008-9-s2-s8

Krenn, B. (2000). *The usual suspects: Data-oriented models for identification and representation of lexical collocations.* DFKI & Universität des Saarlandes, Saarbrücken, Germany.

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research, 44*(D1), D1075-1079. doi:10.1093/nar/gkv1075

Lasko, T. A., Denny, J. C., & Levy, M. A. (2013). Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data.

PLoS ONE [Electronic Resource], 8(6), e66341. doi:http://dx.doi.org/10.1371/journal.pone.0066341

Lecessie, S., & Vanhouwelingen, J. C. (1992). Ridge Estimators in Logistic-Regression. *Applied Statistics-Journal of the Royal Statistical Society Series C, 41*(1), 191-201.

Leong, A., Dasgupta, K., Bernatsky, S., Lacaille, D., Avina-Zubieta, A., & Rahme, E. (2013). Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records. *PloS one, 8*(10), e75256.

Li, D., Endle, C. M., Murthy, S., Stancl, C., Suesse, D., Sottara, D., . . . Pathak, J. (2012). Modeling and executing electronic health records driven phenotyping algorithms using the NQF Quality Data Model and JBoss Drools Engine. *AMIA* .. 532-541.

Li, L. (2015). The potential of translational bioinformatics approaches for pharmacology research. *Br J Clin Pharmacol, 80*(4), 862-867. doi:10.1111/bcp.12622

Li, Q., Melton, K., Lingren, T., Kirkendall, E. S., Hall, E., Zhai, H., . . . Solti, I. (2014). Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. *Journal of the American Medical Informatics Association : JAMIA, 21*(5), 776-784.

Liao, K. P., Ananthakrishnan, A. N., Kumar, V., Xia, Z., Cagan, A., Gainer, V. S., . . . Cai, T. (2015). Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts. *PLoS ONE [Electronic Resource], 10*(8), e0136651. doi:http://dx.doi.org/10.1371/journal.pone.0136651

Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., . . . Kohane, I. (2015). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ, 350*, h1885. doi:http://dx.doi.org/10.1136/bmj.h1885

Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bull Med Libr Assoc, 88*(3), 265-266.

Lui, J. T., & Rudmik, L. (2015). Case definitions for chronic rhinosinusitis in administrative data: A systematic review. *American journal of rhinology & allergy, 29*(5), e146-151.

Macdonald, K. I., Kilty, S. J., & van Walraven, C. (2016). Chronic rhinosinusitis identification in administrative databases and health surveys: A systematic review. *The Laryngoscope, 126*(6), 1303-1310.

Malinowski, J., Farber-Eger, E., & Crawford, D. C. (2014). Development of a data-mining algorithm to identify ages at reproductive milestones in electronic medical records. *Pacific Symposium on Biocomputing*, 376-387.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb), 22*(3), 276-282.

Mo, H., Thompson, W. K., Rasmussen, L. V., Pacheco, J. A., Jiang, G., Kiefer, R., . . . Harris, P. A. (2015). Desiderata for computable representations of electronic health records-driven phenotype algorithms. *Journal of the American Medical Informatics Association, 22*(6), 1220-1230. doi:http://dx.doi.org/10.1093/jamia/ocv112

Nair, P. R., & Nair, V. D. (2014). Organization of a Research Paper: The IMRAD Format. In *Scientific Writing and Communication in Agriculture and Natural Resources* (pp. 13-25): Springer.

Newton, K. M., Peissig, P. L., Kho, A. N., Bielinski, S. J., Berg, R. L., Choudhary, V., . . . Denny, J. C. (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association, 20*(e1), e147-154. doi:http://dx.doi.org/10.1136/amiajnl-2012-000896

Ogu, C. C., & Maxa, J. L. (2000). Drug interactions due to cytochrome P450. *Proceedings (Baylor University Medical Center), 13*(4), 421-423.

Organization, W. H. (2002). The importance of pharmacovigilance.

Ouyang, L., Apley, D. W., & Mehrotra, S. (2016). A design of experiments approach to validation sampling for logistic regression modeling with error-prone medical records. *Journal of the American Medical Informatics Association, 23*(e1), e71-78. doi:http://dx.doi.org/10.1093/jamia/ocv132

Overby, C. L., Pathak, J., Gottesman, O., Haerian, K., Perotte, A., Murphy, S., . . . Weng, C. (2013). A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. *Journal of the American Medical Informatics Association, 20*(e2), e243-252. doi:http://dx.doi.org/10.1136/amiajnl-2013-001930

Ozair, F. F., Jamshed, N., Sharma, A., & Aggarwal, P. (2015). Ethical issues in electronic health records: A general overview. *Perspectives in clinical research, 6*(2), 73-76. doi:10.4103/2229-3485.153997

Pace, R., Peters, T., Rahme, E., & Dasgupta, K. (2017). Validity of Health Administrative Database Definitions for Hypertension: A Systematic Review. *The Canadian journal of cardiology, 33*(8), 1052-1059.

Park, H., & Choi, J. (2014). V-Model: a new perspective for EHR-based phenotyping. *BMC Medical Informatics & Decision Making, 14*, 90. doi:http://dx.doi.org/10.1186/1472-6947-14-90

Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association, 20*(e2), e206-211. doi:http://dx.doi.org/10.1136/amiajnl-2013-002428

Peterson, K. J., & Pathak, J. (2014). Scalable and High-Throughput Execution of Clinical Quality Measures from Electronic Health Records using MapReduce and the JBoss Drools Engine. *AMIA ... Annual Symposium Proceedings/AMIA Symposium, 2014*, 1864-1873.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, 185-208.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*: Elsevier.

Rasmussen, L. V., Thompson, W. K., Pacheco, J. A., Kho, A. N., Carrell, D. S., Pathak, J., . . . Starren, J. B. (2014). Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *Journal of Biomedical Informatics, 51*, 280-286. doi:http://dx.doi.org/10.1016/j.jbi.2014.06.007

Rebholz-Schuhmann, D., Oellrich, A., & Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature reviews. Genetics, 13*(12), 829.

Reich, C., Ryan, P. B., Stang, P. E., & Rocca, M. (2012). Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of Biomedical Informatics, 45*(4), 689-696. doi:https://doi.org/10.1016/j.jbi.2012.05.002

Richesson, R., Smerek, M., Rusincovitch, S., Zozus, M. N., Chaudhuri, P. S., Hammond, W. E., . . . Uhlenbrauck, G. In Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. Bethesda, MD: NIH Health Care Systems Research Collaboratory. Retrieved from http://rethinkingclinicaltrials.org/resources/ehr-phenotyping/.

Richesson, R. L., Hammond, W. E., Nahm, M., Wixted, D., Simon, G. E., Robinson, J. G., . . . Califf, R. M. (2013). Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association, 20*(e2), e226-231. doi:10.1136/amiajnl-2013-001926

Richesson, R. L., Rusincovitch, S. A., Wixted, D., Batch, B. C., Feinglos, M. N., Miranda, M. L., . . . Spratt, S. E. (2013). A comparison of phenotype definitions for diabetes mellitus. *Journal of the American Medical Informatics Association, 20*(e2), e319-326. doi:http://dx.doi.org/10.1136/amiajnl-2013-001952

Richesson, R. L., Sun, J., Pathak, J., Kho, A. N., & Denny, J. C. (2016). Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med, 71*, 57-61. doi:10.1016/j.artmed.2016.05.005

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics, 42*(5), 950-966. doi:10.1016/j.jbi.2008.12.013

Roden, D. M., & Denny, J. C. (2016). Integrating electronic health record genotype and phenotype datasets to transform patient care. *Clinical Pharmacology & Therapeutics, 99*(3), 298-305. doi:http://dx.doi.org/10.1002/cpt.321

Rodriguez-Esteban, R. (2009). Biomedical text mining and its applications. *PLoS Computational Biology, 5*(12), e1000597. doi:10.1371/journal.pcbi.1000597

Rosenman, M., He, J., Martin, J., Nutakki, K., Eckert, G., Lane, K., . . . Hui, S. L. (2014). Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory. *Journal of the American Medical Informatics Association, 21*(2), 345-352. doi:http://dx.doi.org/10.1136/amiajnl-2013-001942

Rubbo, B., Fitzpatrick, N. K., Denaxas, S., Daskalopoulou, M., Yu, N., Patel, R. S., . . . Flaig, R. (2015). Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *International journal of cardiology, 187*, 705-711.

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196-207. doi:10.1016/j.jbi.2014.11.002

Shah, B. R., & Hux, J. E. (2003). Quantifying the risk of infectious diseases for people with diabetes. *Diabetes Care, 26*(2), 510-513.

Shankar-Hari, M., Phillips, G. S., Levy, M. L., Seymour, C. W., Liu, V. X., Deutschman, C. S., . . . Sepsis Definitions Task, F. (2016). Developing a New Definition and Assessing New Clinical Criteria for Septic Shock: For the Third International

Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA, 315*(8), 775-787. doi:10.1001/jama.2016.0289

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504. doi:10.1101/gr.1239303

Shatkay, H., & Craven, M. (2012). *Mining the biomedical literature*: MIT Press.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association, 21*(2), 221-230. doi:http://dx.doi.org/10.1136/amiajnl-2013-001935

Simonett, J. M., Sohrab, M. A., Pacheco, J., Armstrong, L. L., Rzhetskaya, M., Smith, M., . . . Fawzi, A. A. (2015). A Validated Phenotyping Algorithm for Genetic Association Studies in Age-related Macular Degeneration. *Scientific Reports, 5*, 12875. doi:http://dx.doi.org/10.1038/srep12875

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach %J Comput. Linguist. *22*(1), 1-38.

Somnath Pal, B. (2017). Reporting and Consequences of Adverse Events. *U.S. Pharmacist, 42*(10), 12.

Souri, S., Symonds, N. E., Rouhi, A., Lethebe, B. C., Garies, S., Ronksley, P. E., . . . McBrien, K. A. (2017). Identification of validated case definitions for chronic disease using electronic medical records: a systematic review protocol. *Systematic reviews, 6*(1), 38.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics, 6*(3), 239-251.

Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. *Proceedings / AMIA* .. 662-666.

Sultana, J., Cutroneo, P., & Trifiro, G. (2013). Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother, 4*(Suppl 1), S73-77. doi:10.4103/0976-500X.120957

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspect Biol Med, 31*(4), 526-557.

Tache, S. V., Sonnichsen, A., & Ashcroft, D. M. (2011). Prevalence of adverse drug events in ambulatory care: a systematic review. *Ann Pharmacother, 45*(7-8), 977-989. doi:10.1345/aph.1P627

Tafti, A. P., Badger, J., LaRose, E., Shirzadi, E., Mahnke, A., Mayer, J., . . . Peissig, P. (2017). Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure. *Jmir Medical Informatics, 5*(4). doi:ARTN e51
10.2196/medinform.9170

Tatonetti, N. P., Fernald, G. H., & Altman, R. B. (2012). A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *Journal of the American Medical Informatics Association, 19*(1), 79-85. doi:10.1136/amiajnl-2011-000214

Thiese, M. S. J. B. m. B. m. (2014). Observational and interventional study design types; an overview. *24*(2), 199-210.

Thompson, W. K., Rasmussen, L. V., Pacheco, J. A., Peissig, P. L., Denny, J. C., Kho, A. N., . . . Pathak, J. (2012). An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. *AMIA ... Annual Symposium Proceedings/AMIA Symposium, 2012*, 911-920.

U.S. Food and drug administration. Preventable Adverse Drug Reactions: A Focus on Drug Interactions. (2018, 03/06/2018). Retrieved from https://www.fda.gov/drugs/developmentapprovalprocess/developmentresources/druginteractionslabeling/ucm110632.htm

U.S. Food and drug administration. What is a Serious Adverse Event? (02/01/2016). Retrieved from https://www.fda.gov/safety/medwatch/howtoreport/ucm053087.htm

Uzuner, O. (2009). Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association, 16*(4), 561-570. doi:10.1197/jamia.M3115

Verspoor, K., Jimeno Yepes, A., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., & Plazzer, J. P. (2013). Annotating the biomedical literature for the human variome. *Database: The Journal of Biological Databases and Curation, 2013*, bat019. doi:https://dx.doi.org/10.1093/database/bat019

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine, 37*(5), 360-363.

Wang, J.-F., & Chou, K.-C. (2010). Molecular Modeling of Cytochrome P450 and Drug Metabolism. *Current Drug Metabolism, 11*(4), 342-346. doi:10.2174/138920010791514180

Wang, X., Hripcsak, G., Markatou, M., & Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association : JAMIA, 16*(3), 328-337.

Wei, C.-H., Kao, H.-Y., & Lu, Z. (2012, 2012/04/05). *PubTator: A PubMed-like interactive curation system for document triage and literature curation*.

Wei, W. Q., & Denny, J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med, 7*(1), 41. doi:10.1186/s13073-015-0166-y

Wei, W. Q., Teixeira, P. L., Mo, H., Cronin, R. M., Warner, J. L., & Denny, J. C. (2016). Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association, 23*(e1), e20-27. doi:http://dx.doi.org/10.1093/jamia/ocv130

Wiedemann, G., & Niekler, A. *Hands-On: A Five Day Text Mining Course for Humanists and Social Scientists in R.* Paper presented at the Proceedings of the 1st Workshop Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin.

Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics, 7*, 356. doi:10.1186/1471-2105-7-356

Wiley, L. K., Moretz, J. D., Denny, J. C., Peterson, J. F., & Bush, W. S. (2015). Phenotyping Adverse Drug Reactions: Statin-Related Myotoxicity. *AMIA Joint*

Summits on Translational Science proceedings AMIA Joint Summits on Translational Science, 2015, 466-470.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., . . . Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research, 34*(Database issue), D668-672. doi:10.1093/nar/gkj067

Wu, H. Y., Zhang, S., Desta, Z., Quinney, S., & Li, L. (2017). Translational Drug Interaction Evidence Gap Discovery Using Text Mining. *Clinical Pharmacology & Therapeutics, 101*(S1), S91-S92.

Wu, S. C., Fu, C. Y., Muo, C. H., & Chang, Y. J. (2014). Splenectomy in trauma patients is associated with an increased risk of postoperative type II diabetes: a nationwide population-based study. *American Journal of Surgery, 208*(5), 811-816. doi:10.1016/j.amjsurg.2014.03.003

Wu, S. T., Liu, H., Li, D., Tao, C., Musen, M. A., Chute, C. G., & Shah, N. H. (2012). Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *Journal of the American Medical Informatics Association, 19*(e1), e149-156. doi:10.1136/amiajnl-2011-000744

Xu, J., Rasmussen, L. V., Shaw, P. L., Jiang, G., Kiefer, R. C., Mo, H., . . . Montague, E. (2015). Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. *Journal of the American Medical Informatics Association, 22*(6), 1251-1260. doi:http://dx.doi.org/10.1093/jamia/ocv070

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining Electronic Health Records (EHRs): A Survey. *ACM Comput. Surv., 50*(6), 1-40. doi:10.1145/3127881

Yao, L., Zhang, Y., Li, Y., Sanseau, P., & Agarwal, P. (2011). Electronic health records: Implications for drug discovery. *Drug discovery today, 16*(13), 594-599.

Yeleswarapu, S., Rao, A., Joseph, T., Saipradeep, V. G., & Srinivasan, R. (2014). A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC Med Inform Decis Mak, 14*, 13. doi:10.1186/1472-6947-14-13

Yetisgen-Yildiz, M., & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics, 39*(6), 600-611. doi:10.1016/j.jbi.2005.11.010

Yu, S., Liao, K. P., Shaw, S. Y., Gainer, V. S., Churchill, S. E., Szolovits, P., . . . Cai, T. (2015). Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association, 22*(5), 993-1000. doi:http://dx.doi.org/10.1093/jamia/ocv034

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*: Cambridge University Press.

Zeng, Z., Deng, Y., Li, X., Naumann, T., & Luo, Y. (2018). Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM transactions on computational biology and bioinformatics*.

Zhang, P., Du, L., Wang, L., Liu, M., Cheng, L., Chiang, C. W., . . . Li, L. (2015). A Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction

Effects on Myopathy Using Electronic Medical Record Databases. *CPT: pharmacometrics & systems pharmacology, 4*(8), 474-480.

Zhang, P., Wu, H. Y., Chiang, C. W., Wang, L., Binkheder, S., Wang, X., . . . Li, L. (2018). Translational Biomedical Informatics and Pharmacometrics Approaches in the Drug Interactions Research. *CPT: pharmacometrics & systems pharmacology, 7*(2), 90-102. doi:10.1002/psp4.12267

Zhao, D., & Weng, C. (2011). Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of Biomedical Informatics, 44*(5), 859-868. doi:10.1016/j.jbi.2011.05.004

Zhao, J., Henriksson, A., Asker, L., & Bostrom, H. (2015). Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Med Inform Decis Mak, 15 Suppl 4*, S1. doi:10.1186/1472-6947-15-S4-S1

CURRICULUM VITAE

Samar Hussein Binkheder

Education

Indiana University, Indianapolis, USA

- Ph.D. Health Informatics (Minor: Bioinformatics) July 2019
- Master of Science in Bioinformatics December 2013

Umm Al-Qura University, Makkah, KSA

- Bachelor Degree of Applied Medical Sciences in Laboratory Medicine June 2007

Academic Appointments and Professional Experience

King Saud University, Riyadh, KSA

- Lecturer (Health Informatics) at Medical Informatics & E-learning Unit, College of Medicine March 2018 - Current
- Teaching Assistant (Health Informatics) at Medical Informatics & E-learning Unit, College of Medicine January 2015 - March 2018

Indiana University, Indianapolis, USA

- Adjunct lecturer (Health Informatics): Teaching an undergraduate course INFO-B481-Health Information Standards & Terminology Spring 2018 & Spring 2019
- Research Assistant at School of Informatics and Computing January 2014 - December 2014

King Abdulaziz University Hospital, Jeddah, KSA

- Medical Laboratory Technology Volunteer at Histopathology Laboratory February 2009 - April 2009

- Medical Laboratory Technology Intern — July 2007 - July 2008

King Faisal Specialist Hospital & Research Center, Jeddah, KSA

- Medical Laboratory Technology Volunteer at Department of Pathology & Laboratory Medicine — June 2006 - July 2006

Professional Development, Conferences, and Certification

- Attended IEEE International Conference on Healthcare Informatics (ICHI) and presented a poster, New York City, USA — July 2018
- Attended The American Medical Informatics Association (AMIA) Annual Symposium and presented poster, Chicago, USA — November 2016
- The Sixth Annual Indiana Clinical and Translational Sciences Institute (CTSI) Symposium on Disease and Therapeutic Response Modeling, Indianapolis, USA — November 2016
- Academic Intensive English Program (AIEP) with TOFEL iBT preparation (Three levels), UCLA Extension, Los Angeles, USA — December 2011
- Cal America Education Institute for English as a second language (4 months), Los Angeles, California, USA — December 2010

Technical Skills

- R statistical language
- Python
- Extensible Markup Language (XML)
- UNIX
- Access Database
- MySQL Database
- WEKA
- Cytoscape

Peer-Review for Journals and Conferences

- Applied Clinical Informatics Journal — 2018-Current
- American Medical Informatics Association Symposium — 2015 and 2019

Scholarships

- Full tuition for PhD degree in the USA, King Saud University, Riyadh, KSA, January 2015-Now
- Academic scholarship and Graduate assistantship for earning PhD in Health Informatics, School of Informatics and Computing, Indiana University, Indianapolis, USA, January 2014-December 2014
- Full tuition for Master's and PhD degree in the USA, Saudi Arabian Cultural Mission (SACM), KSA, August 2010-January 2015

Publications and Posters

1. **Binkheder, S**., Wu, H. Y., Quinney, S., Zhang, S., Zitu, M., Chiang, C., Wang, L., Wu, H., Jones, J., & Li, L. A large-scale biomedical literature mining of phenotyping definition sentences. *Journal of Biomedical Informatics*. (Submitted on April 01, 2019)
2. **Binkheder, S**., Wu, H. Y., Quinney, S., Zhang, S…, Jones, J., & Li, L. A Corpus for Annotating Sentences with Information of Phenotyping Definitions in Biomedical Literature. *Journal of Biomedical Semantics*. (In-progress for submission).
3. **Binkheder, S**., Wu, H. Y., Quinney, S., & Li, L. (2018, June). Analyzing Patterns of Literature-Based Phenotyping Definitions for Text Mining Applications. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 374-376).
4. Zhang, P., Wu, H. Y., Chiang, C. W., Wang, L., **Binkheder, S.**, Wang, X., ... & Li, L. (2018). Translational biomedical informatics and pharmacometrics approaches in the drug interactions research. *CPT: pharmacometrics & systems pharmacology*, *7*(2), 90-102.
5. Holden, R. J., **Binkheder, S.**, Patel, J., & Viernes, S. H. P. (2018). Best Practices for Health Informatician Involvement in Interprofessional Health Care Teams. *Applied clinical informatics*, *9*(01), 141-148.
6. **Binkheder, S.**, Dixon, B. E., & Grannis, S. J. (2016). Transmission of ELR messages To Improve Public Health Reporting. In *AMIA*. (Poster)
7. Dixon, B. E., Henderson, M., McFarlane, T., Kirbiyik, U., **Binkheder, S.**, Albertin, C., Saurabh Rahurkar, C. (2016). Early Findings from a Public Health Informatics Laboratory for Interprofessional Doctoral Students. In *AMIA*. (Poster)

8. Jones, J. F., Zolnoori, M., **Binkheder, S.**, Schilling, K., Pondugala, L. R., & Lenox, M. (2014). The Extent to which US Hospitals Promote Their Patient Engagement Activities and Outcomes: Preliminary Results of Quantitative Content Analysis Research. In *AMIA*. (Poster)