Original Paper

Enter information for authors (including designations, affiliations, correspondence, contributions) in the online metadata form. Do not use periods after initials, and include degree designations and affiliations for all authors. Trial registration numbers are also filled in on the metadata forms online.

PIVET: A Scaled Phenotype Evidence Generation Framework using Online Medical Literature

Abstract

Background: Researchers are developing methods to automatically extract clinically relevant and useful patient characteristics from raw healthcare datasets. These characteristics, often capturing essential properties of patients with common medical conditions, are called *computational phenotypes.* After being generated by (semi)-automated, data-driven methods, such potential phenotypes need to be validated as clinically meaningful (or not) before they are acceptable for use in decision making.
Objective: We present Phenotype Instance Verification and Evaluation Tool (PIVET), a framework that uses co-occurrence analysis on an online corpus of publically available medical journal articles to build clinical relevance evidence sets for user-supplied phenotypes. PIVET adopts the same conceptual framework as the pioneering prototype tool, PheKnow-Cloud, which was developed for the phenotype validation task. PIVET completely refactors each part of the PheKnow-Cloud pipeline to deliver vast improvements in speed without sacrificing the quality of the insight PheKnow-Cloud achieved.
Methods: PIVET leverages indexing in NoSQL databases to efficiently generate evidence sets. Specifically, PIVET uses a succinct representation of the phenotypes that corresponds to the index on the corpus database and an optimized co-occurrence algorithm inspired by the Aho-Corasick algorithm. We compare PIVET's phenotype representation to that PheKnow-Cloud by using PheKnow-Cloud's experimental set-up. We also introduce a statistical model trained on domain-expert verified phenotypes to automatically classify phenotypes as clinically relevant or not. Additionally, we show how the classification model can be used to examine user-supplied phenotypes in an online, rather than batch, manner.
Results: PIVET maintains the discriminative power of PheKnow-Cloud in terms of identifying clinically relevant phenotypes for the same corpus with which PheKnow-Cloud was originally developed, but PIVET's analysis is an order of magnitude faster than that of PheKnow-Cloud. Not only is PIVET much faster, it can be scaled to a larger corpus and still retain speed. We trained multiple classification models on top of the PIVET framework and found ridge regression to perform best, realizing an average F1 score of 0.91 when predicting clinically relevant phenotypes.
Conclusions: We show PIVET improves on the most notable existing prototype tool in terms of speed and automation, and is comparable in terms of accuracy.

Keywords: text mining; computational phenotyping

Introduction

The rapidly expanding availability of electronic health records offers the promise to help clinicians better understand the populations they serve. The ability to efficiently characterize large volumes of healthcare data is essential to enabling clinicians to use this information effectively. Recently, machine learning and data mining researchers have attempted to address this need in several ways. One such line of work concerns developing methods to extract "computational phenotypes" from raw health data in an automated, high-throughput manner. Here we define a computational phenotype as a constellation of clinically interesting characteristics that delineates a cohesive group of patients. Such phenotypes can help clinicians reason about patient populations, identify patient cohorts, and identify and describe the progression of diseases within populations.

While being able to extract phenotypes in a high-throughput manner constitutes a potentially important step in helping clinicians reason about their patient populations on a larger scale, this potential will be realized only if the identified phenotypes are clinically meaningful. Therefore, to increase the utility of data-driven phenotypes, some measure quantifying the inferred clinical meaningfulness should be reported alongside the phenotypes to help practitioners sort signal from noise. To address this need, we present PIVET (Phenotype Instance Validation and Evaluation Tool), a tool that uses analysis of Open Access PubMed (a corpus of online medical articles) to generate evidence sets and clinical relevance scores for candidate phenotypes. These evidence sets can be used by: researchers when developing and tuning new computational phenotype methods; domain experts when they are validating candidate phenotypes; and, eventually, clinicians examining the phenotypes associated with their patient populations.

PIVET is an improvement on a recently introduced prototype tool called PheKnow-Cloud [1]. PheKnow-Cloud, which earned the Distinguished Paper Award at the 2017 AMIA Joint Summits, demonstrated the medical expertise contained in PubMed articles could be harnessed to build evidence sets for the clinical validity of candidate phenotypes. PIVET is built on the same conceptual framework PheKnow-Cloud, but in PIVET, we have optimized each piece of the PheKnow-Cloud's pipeline to deliver vast improvements in speed and interpretability without sacrificing the integrity of PheKnow-Cloud's phenotype evaluation.

The PheKnow-Cloud pipeline consists of three major steps: 1) representing each phenotype so occurrences of it and related terms in the corpus will be recognized (Phenotypic Representation); 2) analyzing the corpus using the Phenotype Representation (Corpus Analysis); and 3) calculating a clinical relevance score and designation (Clinical Validity Determination). In the Phenotype Representation step, PIVET uses succinct and possibly more interpretable representations of terms contained within each phenotype. In the Corpus Analysis step, PIVET migrates from a brute-force approach of analyzing the corpus to a NoSQL database to store and index the articles efficiently. PIVET then utilizes a variation of the Aho-Corasick to count appearances of the terms within each phenotype. Finally, in the Clinical Validity Calculation step, PIVET streamlines the

clinical relevance score analysis and uses a model, trained on domain-expert-verified phenotypes, to classify the clinical relevance of supplied phenotypes. Through a combination of these improvements, PIVET runs an order of magnitude faster than PheKnow-Cloud without sacrificing the discriminative power of the original tool.

PheKnow-Cloud was developed to function in high-throughput phenotyping situations where a researcher has a large set of phenotypes to validate. Consequently, PheKnow-Cloud was built to run only in a batch setting. However, in clinical settings and some research settings, a user may only have one phenotype to analyze, so we developed PIVET to run in either an online or batch environment. This improvement will allow clinicians to query PIVET with single phenotypes, which could possibly help in decision making processes. Additionally, it could help researchers to tune their phenotype extraction algorithms. Thus, while the prototype tool demonstrated the analysis of medical articles could be used to evaluate candidate phenotypes, the improvements in speed and automation realized by PIVET make it useful in both research and clinical settings.

The paper is organized as follows. We first present research related to PIVET, including a description of the original prototype tool (PheKnow-Cloud). Next, we describe the PIVET framework, noting the important differences between PheKnow-Cloud and the new system. We then report the performance of PIVET on automatically generated phenotypes as well as domain-expert-curated phenotypes and demonstrate how the framework can be used in an online setting. We conclude the paper a discussion of the limitations of this work and thoughts on future directions.

Related Work

PubMed

PubMed Central (PMC) is an online collection comprising over 3 million biomedical and biological journal articles gathered from thousands of journals [3]. PMC is maintained and curated by the National Library of Medicine (NLM) at the U.S. National Institute of Health [4].

In regard to phenotypes, researchers tend to use PubMed as an exploratory tool to *discover* new phenotypes rather than as a resource to *validate* candidate phenotypes. Boland et al. orchestrated one of the few studies that used PubMed as a validation tool. They mined electronic health records (EHRs) for patients with predefined disease codes and then compared the birth month and the disease of these patients to a group of control patients who did not have the disease codes present in their EHRs. They found a relationship between certain diseases and birth months in the case group [5]. They validated their results against papers retrieved from PubMed that mentioned disease and birth month.

More commonly, researchers use PubMed as tool to generate hypotheses and discover phenotypes and other biomedical issues [6] [7]. Multiple software packages like LitInspector [8], PubMed.mineR [9], ALIBABA [10] as well as python packages like Pymedtermino [11] and Biopython [12] have been developed to help researchers extract

and visualize PubMed. Other researchers have built tools to rank search results, discover topics and relationships within search results, visualize search results, and improve user interaction with PubMed [13].

Text Mining Pubmed

Jensen et al. give a thorough overview of how PubMed can be harnessed for information extraction and entity recognition [7]. Natural Language Processing (NLP) techniques form one approach to mining the literature. Some researchers have used NLP techniques on PubMed to discover disease-gene associations [14] and others have used PubMed in concert with other data sources to generate phenotypes [15]. Collier et al. used NLP techniques in conjunction with association rule mining to discover phenotypes using PubMed [16]. However, none of these approaches have sought to use PubMed as a validation tool for data-driven phenotypes.

Co-occurrence analysis, which is what PheKnow-Cloud and PIVET are built on, is more widely used because it is simple to implement and interpret. Researchers have applied co-occurrence strategies to generate phenotypes. Some have performed co-occurrence analysis on PubMed to study links between diseases [17] [18], which can be viewed as a simple type of phenotype discovery. Others have explored relationships between phenotypes and genotypes [19] [20]. In contrast to this work, our approach uses phenotypes as the starting point, and performs co-occurrence analysis over the PMC corpus as a means of assessing their validity. We assume these phenotypes were induced over other sources (e.g., EHRs) and not from PMC. Co-occurrence analysis has the drawback of not being able to explicitly model the type of relationship that exists between two or more terms (e.g., negative or positive). However, we require the terms within a phenotype are positively related to one another, which aligns with the findings of publication bias research.

Publication bias is the tendency for the academic publishing ecosystem (e.g., researchers, reviewers, editors, etc.) to submit and publish articles that show positive relationships between the entities being studied. The non-random omission of results that is not based on the quality of the methodology but on the direction of the results is a well-studied area of research and has been shown to have a negative effect on research in many cases [21] [22] [23] [24] [25] [26]. Publication bias introduces risks to researchers and to the general public to which research is applied (via policies and treatment decisions).

However, in PheKnow-Cloud and PIVET, this bias is a strength rather than a drawback. The current focus of PheKnow-Cloud and PIVET is on the presence of relationships within the user-supplied candidate phenotypes, so the publishing of positive results aligns with publication bias. Furthermore, since co-occurrence analysis does not attempt to infer information about the type of relationship or any causal information, the presence of publication bias allows for the assumption that when two phrases occur together, there is evidence that implies the relationship exists [22] [27] [28].

PheKnow-Cloud Prototype

Phenotype evaluation via co-occurrence analysis of online articles was first introduced by Bridges et al. [1]. Henderson and colleagues improved on the evaluation framework and

developed a prototype tool implementing the approach called PheKnow-Cloud, which provided a web interface for researchers and clinicians to interact with the technology [2]. We refer to the tool and framework introduced in those two works as PheKnow-Cloud. The input to the PheKnow-Cloud process is a set of potential phenotypes. Each phenotype consists of medical terms, which we refer to as phenotypic items, that are assumed to have been generated by an automatic high-throughput phenotyping process. PheKnow-Cloud generates evidence sets for batches of phenotypes based on co-occurrence analysis of the PubMed corpus. We refer the reader to [1] and [2].

PheKnow-Cloud was developed as a proof-of-concept tool, and while it showed the PubMed corpus could be used to determine whether a phenotype was clinically valid, it had several drawbacks that PIVET addresses. One is the length of time the prototype method required to complete analyses; Table 1 compares the time that each method takes to perform each step. The computational bottlenecks for the prototype method are the co-occurrence generation and clinical relevance score analysis steps. The synonym generation step speed is determined by the number of requests that can be made to the NLM Medical Subject Headings (MeSH) database, which is an off-site system that places limits on the number of requests users can make in a given window of time. PIVET speeds up this process considerably. Another drawback of PheKnow-Cloud is that the clinical relevance scores for phenotypes are calculated only relative to all other phenotypes and must be used in a batch setting. In contrast, PIVET can analyze a single phenotype at a time, which makes it more flexible than PheKnow-Cloud. Finally, designating whether a candidate phenotype is clinically relevant or not is a manual process in PheKnow-Cloud. For PIVET, we built a classifier trained on a validated set of phenotypes. This classifier can be ported to other environments and can be used to automatically classify new, individual phenotypes.

Table 1. The time in seconds and (hours: minutes: seconds) each method used to complete task in phenotype generation process All experiments were run on a machine with 3 AMD A6-5200 APU with Radeon(TM) HD Graphics processors, 8 GB of memory, 1 TB hard drive, running Ubuntu 14.04.5 LTS.

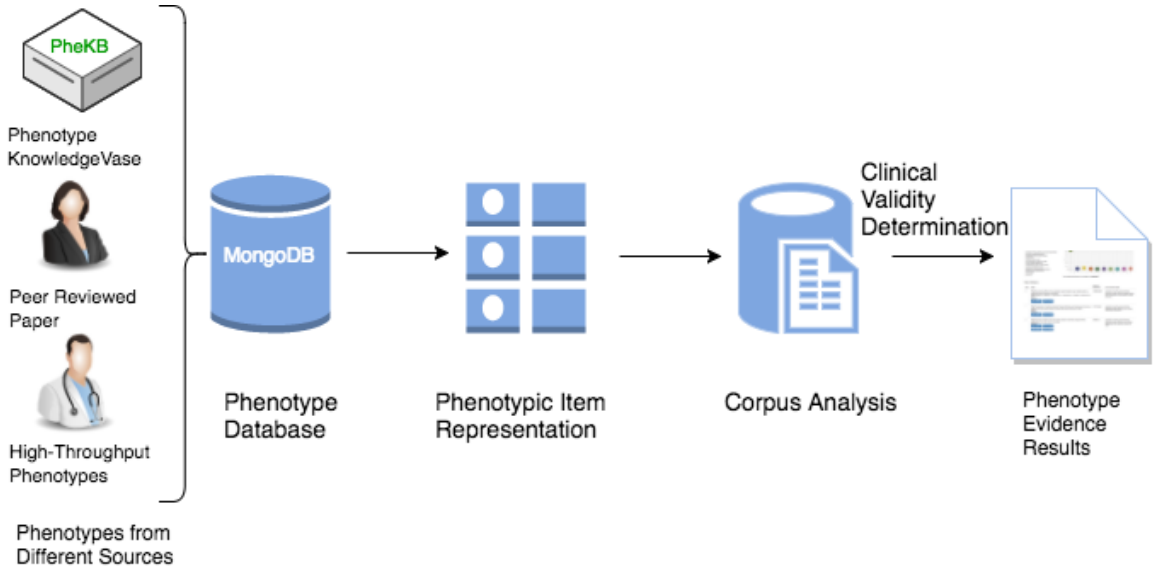
	PheKnow-Cloud	PIVET
Synonym Generation	7,809 (02:10:09)	5,948 (01:39:08)
Co-Occurrence Analysis	50,822 (14:07:02)	289 (00:04:59)
Lift Analysis	2,092 (00:34:52)	2 (00:00:02)
Total	**60,723 (16:52:03)**	**6,239 (01:43:59)**

Methods

In this section, we describe how PIVET performs co-occurrence analysis on an online corpus of publicly available journal articles to build evidence sets for phenotypes. This involves five components: (1) a database of phenotypes to analyze; (2) a database of Pubmed article corpus indexed by medical terms the articles contain; (3) an algorithm to generate and rank synonyms for the phenotypic items (Phenotypic Item Representation); (4) a co-occurrence analysis module (Corpus Analysis); and (5) a clinical relevance scoring system (Clinical Validity Determination). Figure 1 captures the PIVET workflow

and the different components of the system. Both MongoDB (an open-source document-based NoSQL database system) and MySQL (an open-source relational database management system) are used to ensure consistency, durability, and efficiency.

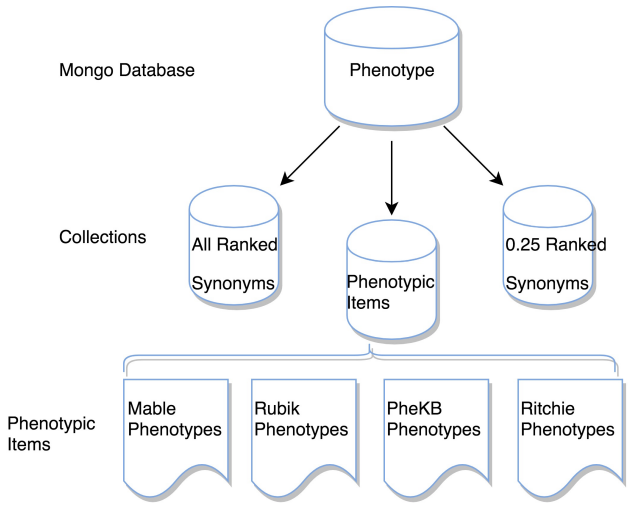
Figure 1. PIVET analysis process. Phenotypes are collected in standardized format in a MongoDB (i.e., "Phenotype Database"). For a single phenotype, synonyms for each phenotypic item in a phenotype are generated using the NLM MeSH database and ranked based on their similarity to the phenotypic item (i.e., "Phenotypic Item Representation"). Co-occurrence analysis is performed on PubMed using the synonyms generated in the previous step (i.e., "Corpus Analysis"). Lift analysis is performed, clinical relevance scores are calculated, and a classifier classifies the phenotype as clinically relevant or not (i.e., "Clinical Validity Determination"). The results of the analysis of the phenotype are presented to the view (i.e., "Phenotype Evidence Results").



Phenotype Storage

PIVET can be used to analyze phenotypes generated using a variety of methods. Every phenotype analyzed by PIVET is stored in a MongoDB using a standardized representation to ensure consistency. We also created a simple parser to ingest new phenotypes that are stored in Javascript Object Notation (JSON). The choice of JSON also facilitates the eventual integration of a web platform where users can provide new phenotypes. We populate the phenotype database with phenotypes from different sources (Figure 2).

Figure 2. Database schema for storing phenotype information



Specifically, we collect phenotypes from two high-throughput phenotyping algorithms, a catalog of algorithms from a collaborative database, and a peer-reviewed paper. The phenotype database contains 80 phenotypes generated using two unsupervised, nonnegative tensor factorization models to perform automated phenotyping [32] [33]. These were subsequently annotated by domain experts, and they were the phenotypes used to validate PheKnow-Cloud. The two automatic methods, Rubik [33] and Marble [32], extracted 30 and 50 candidate phenotypes, respectively, from the diagnoses and medications of 7,744 de-identified patients from Vanderbilt University Medical Center recorded over a five-year observation period. Each member of the panel assigned all phenotypes one of the following three labels: (1) *yes*, the candidate phenotype is clinically meaningful and therefore a phenotype; (2) *no*, the candidate phenotype is not clinically meaningful and therefore not a phenotype, or (3) *maybe*, the candidate phenotype is possibly clinically meaningful. Of the 80 Marble and Rubik combined phenotypes, the domain experts labelled approximately 14% as clinically meaningful, 78% as possibly significant, and 8% as not clinically meaningful. For the handful of phenotypes where the domain experts disagreed on the clinical relevance, the label that awarded the least amount of clinical significance was assigned. These annotated phenotypes were graciously shared by the authors of Rubik.

Additionally, two groups of domain-expert generated phenotypes are included in the phenotype database. The first set, which we will refer to as the "gold standard" phenotypes, are from the Phenotype KnowledgeBase (PheKB), an online phenotype knowledgebase that stores researchers' collaborations of electronic algorithms of phenotypes [34]. Gold standard phenotypes are developed by panels of domain experts across multiple sites. We manually extracted 13 phenotypes that have been reviewed and finalized by the Electronic Medical Records and Genomics (eMerge) Phenotype Working Group. The second set of domain-expert-derived phenotypes, which we will refer to as "silver standard" phenotypes, are the group of validated phenotype algorithms published by Ritchie et al. [35]. Silver standard phenotypes are developed by a panel of domain

experts at a single site. Nine phenotypes were manually extracted from the article. This peer-reviewed paper is not part of the article corpus.

Pubmed Open Access Corpus

PIVET works by analyzing co-occurrences of phenotypic items within the PMC Open Access (OA) subset, an openly available online repository of medical articles which constitutes roughly one-third of the total collection of articles in the PMC (over 1 million articles). The articles within the OA subset are copyright protected but have a flexible license concerning reuse. Trimmed down versions of the articles are stored in a MongoDB. We use the NoSQL database MongoDB because it is a document-based database without restrictive schema ideal for storing articles that vary in content. Furthermore, MongoDB has been shown to outperform SQL-based databases in terms of read, write, and delete operations and scaling to larger datasets [33] [34] [35].

We limit the corpus in the database to those articles with attached Medical Subject Headings (MeSH) terms; this amounts to 379,766 articles. MeSH is a hierarchical vocabulary curated by the National Library of Medicine (NLM) to index and catalog biomedical information [36]. There are 26,000 biomedical concepts or headings and over 200,000 supplementary concepts that form qualifiers for the headings. MeSH has two major benefits over the other existing ontologies. First, a large portion of the PubMed corpus has been manually annotated with MeSH labels. Expert indexers at the NLM assign MeSH terms to each article that best summarize the text. These terms are periodically reviewed and updated. We index the PMC database with the MeSH terms each article contains, and we represent each item in a phenotype with a set of MeSH terms, which is discussed in the next section. The index and phenotypic item representation combined with search optimization techniques described in the subsequent section speed up the co-occurrence analysis process considerably.

Phenotypic Item Representation: Constructing MeSH Synonym Sets

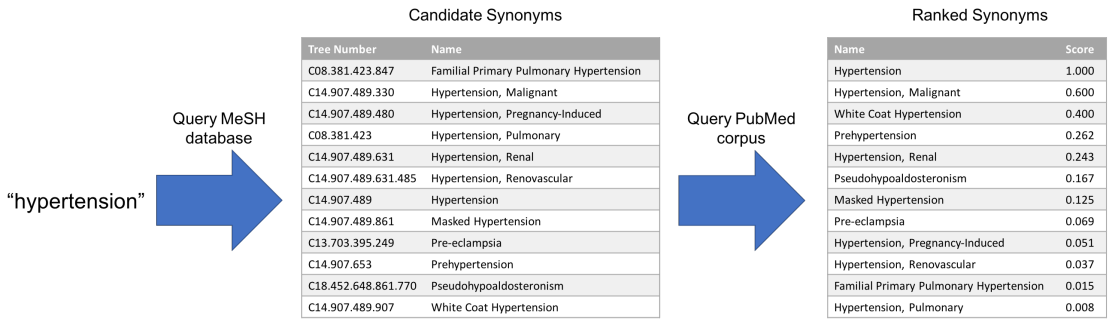
Once the phenotypes are stored in the database, the next step is to build representations for the terms within each phenotype, which refer to as “phenotypic items.” Medical terms can have various synonyms (representations) across different articles. For example, the term “heart attack” can also be referred to as “cardiovascular stroke”, “myocardial infraction”, and “cardiogenic shock”. Thus, it is important to generate a list of synonyms for each phenotypic item to achieve high recall within the PubMed corpus. PheKnow-Cloud built representations for each phenotypic item from related terms and concepts found in the following medical ontologies: MeSH, Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), and International Classification of Diseases (ICD-9 or ICD-10). Further experiments indicated this approach can introduce noise into the representation. Instead, PIVET uses only MeSH terms to generate a phenotypic item representation for each phenotypic item with the following two-step process: (1) assign the most relevant MeSH term and (2) generate a ranked list of closely related MeSH terms.

To generate a candidate set of representations for a phenotypic item, PIVET first queries the NLM MeSH database using Biopython [12] with a cleaned version of the phenotypic item. The search returns a set of MeSH tree numbers. MeSH terms are formed into a hierarchical tree, where each MeSH term is assigned a node in the tree and labeled by a number. This number designates the MeSH term's place in the hierarchy. For example, the tree number of "Hypertension" is C14.907.489, which indicates that it is a child of the node C14.907 ("Vascular Diseases"). Vascular Diseases is in turn a child of node C14 ("Cardiovascular Diseases"). Gathering nodes with the prefix C14.907.489 gives a set of possible synonyms for the original phenotypic item "Hypertension." Generally, this hierarchy gives a relatively straightforward method for finding synonyms and relevant concepts.

Since the query does not rank the results (i.e., it does not designate which tree number is most relevant to the search), it is necessary to identify the MeSH term that most closely matches the phenotypic item. For example, querying the phenotypic item "hypertension" returns the tree numbers that map to the natural language headings: 'Hypertension, Malignant', 'Hypertension, Portal', 'Hypertension, Pulmonary', 'Hypertension, Renal', 'Hypertension', 'Masked Hypertension', 'Prehypertension', etc. (shown in Figure 3). PIVET designates the "most relevant synonym" for the original phenotypic item by finding the natural language heading associated with each of the tree numbers that most closely matches the original phenotypic item. Specifically, for each natural language heading or synonym, PIVET forms a set where each element is a word of the synonym, and then finds size of the intersection between the set and the original cleaned item, which has also been turned into a set. It also records the size difference between the two sets. For example, the phenotypic item "hypertension" and candidate synonym "Hypertension, Malignant" have an intersection of length one (i.e., "hypertension") and a size difference of 1. However, PIVET would assign "Hypertension" as the most relevant synonym because it has an intersection of size one and a set size difference of 0 with the original phenotypic item. In the event of a tie, the algorithm designates the tied candidate synonyms as the most relevant synonyms and builds the synonym sets for each.

The remaining synonyms are then ranked based on the overlap between each candidate synonym and the most relevant synonym in our PubMed Open Access corpus. The percentage overlap, calculated as the number of times the candidate synonym appears with the most relevant synonym divided by the number of times the candidate synonym appears overall, serves as the relevance score to rank each synonym. The ranked list is then used to adjust the number of synonyms. An example of a ranked synonym set can be seen in Figure 3.

Figure 3: Synonym generation process for the term "hypertension". First the NLM MeSH database is queried with the term "hypertension," which returns a list of candidate MeSH terms. From this query result, the "most relevant synonym" is determined through a process of string matching between the original queried term and the candidate synonyms. In this case the most relevant synonym is "Hypertension." The candidate synonyms are then ranked based on the percentage overlap between PubMed articles that contain the MeSH term associated with the candidate synonym and the MeSH term of the most relevant synonym.



Tree Number	Name
C08.381.423.847	Familial Primary Pulmonary Hypertension
C14.907.489.330	Hypertension, Malignant
C14.907.489.480	Hypertension, Pregnancy-Induced
C08.381.423	Hypertension, Pulmonary
C14.907.489.631	Hypertension, Renal
C14.907.489.631.485	Hypertension, Renovascular
C14.907.489	Hypertension
C14.907.489.861	Masked Hypertension
C13.703.395.249	Pre-eclampsia
C14.907.653	Prehypertension
C18.452.648.861.770	Pseudohypoaldosteronism
C14.907.489.907	White Coat Hypertension

Name	Score
Hypertension	1.000
Hypertension, Malignant	0.600
White Coat Hypertension	0.400
Prehypertension	0.262
Hypertension, Renal	0.243
Pseudohypoaldosteronism	0.167
Masked Hypertension	0.125
Pre-eclampsia	0.069
Hypertension, Pregnancy-Induced	0.051
Hypertension, Renovascular	0.037
Familial Primary Pulmonary Hypertension	0.015
Hypertension, Pulmonary	0.008

Corpus Analysis

The aim of the corpus analysis step is to gauge the strength of the relationship between items in a phenotype. However, it is unlikely all items in a phenotype will appear together, so instead PIVET searches the corpus for occurrences of subsets of the phenotypic items (as represented by their phenotypic item MeSH synonym sets as described in the last section). Through experimentation we found only a small fraction of subsets of any phenotype occur in the article corpus. This means it is inefficient as well as computationally infeasible for even moderately sized phenotypes to look for all possible subsets (i.e., the power set, which in this case has $2^{|S|} \cdot n_1 \cdot n_2 \cdot \cdots \cdot n_{|S|}$ elements, where |S| is the cardinality of the phenotype and n_i is the synonym set size for phenotypic item i).

Moreover, as the size of the subset increases, the likelihood of all the terms appearing in any given article diminishes. Therefore, it is not necessary to enumerate all the possible subsets. Using this observation, we use an algorithm inspired by the string-matching Aho-Corasick algorithm to search the space effectively [37]. We sketch the algorithm with a set comprised of terms A, B, C, and D that we assume all occur in the corpus. We observe that if terms A and B, comprising a tuple (A,B), do not co-occur in any article together, then any larger subset also containing these two terms will necessarily have zero counts (e.g., (A, B, C), (A, B, D), and (A, B, C, D)). As a result, only non-zero (feasible) co-occurrence subsets need to be expanded. A key insight for efficient expansion of an existing co-occurrence subset with non-zero counts is to join it with the associated tuple pairs with one overlapping term that have non-zero counts. For example, if the only non-zero tuple pairs are (A, C), (A, D), (B, C), (B, D), and (C, D), then the possible tuples with cardinality 3 are (A, C, D) and (B, C, D). As increasing the cardinality size of the tuple is equivalent to a join operation in a SQL database, PIVET uses MySQL to implement this portion of the analysis. After constructing the query tuples of MeSH terms in MySQL, PIVET then counts the number of articles where each tuple appears.

Additionally, we set a few more restrictions on the subset queries to make them even more efficient. For one, each subset is constructed using "different" phenotypic items to avoid arbitrary inflation of counts. If two or more phenotypic items contain identical MeSH synonym sets, a "super" phenotypic item is formed (e.g., "Tuberculosis of adrenal glands" and "Tuberculosis of adrenal glands, bacteriological or histological examination

not done" are merged together). In addition, terms for the same phenotypic item (e.g., all MeSH terms associated with "myocardial infraction") are never paired with one other.

Given these tuple co-occurrence counts, the next step is to map the co-occurring subsets of phenotypic synonyms back to their phenotypic items. For example, if the synonym set for the phenotypic item "Attention deficit disorder" contains two synonym terms "Attention Deficit and Disruptive Behavior Disorders" and "Attention Deficit Disorder with Hyperactivity", then any tuple of cardinality 1 with either of these terms are collected and the sum of the co-occurrences are then designated as the number of times the phenotypic item, "Attention deficient disorder", occurred. The aggregated co-occurrence count for all the non-zero subsets of the phenotypic items are then used to calculate the clinical relevance scores for the phenotype.

Clinical Validity Determination

PIVET uses a two-step process to calculate the clinical relevance score: (1) obtain the *lift* (see below) for each co-occurring subset of phenotypic items; and (2) classify the relevance of the phenotype based on features derived from the previous step. As in PheKnow-Cloud, lift is used because it measures the strength of the relationship between a set of items. Specifically, given items $I_1, I_2, \dots, I_N$,

$$lift(I_1, I_2, \dots, I_N) = \frac{P(I_1 \cap I_2 \cap \cdots \cap I_{N-1} \cap I_N)}{P(I_1)P(I_2) \cdots P(I_{N-1})P(I_N)}$$

A lift of greater than 1 suggests a non-random relationship. In PIVET, the lift calculation entails dividing the percentage of times items appear together in the corpus by the product of percentages of times each item appears individually in the corpus, which can be rewritten as:

$$lift(I_1, I_2, \dots, I_N) = \frac{count(I_1, I_2, \dots, I_{N-1}, I_N)}{count(I_1)count(I_2) \cdots count(I_N)} D^{N-1},$$

where count(A) is the number of articles in the corpus that contain the set A, and D is the number of documents in the corpus.

It was observed in PheKnow-Cloud that the lift increases exponentially with the size of the co-occurrence set [1]. This is consistent with the above equation, for example, if a set of six items appear together then the fraction of counts will be multiplied by the size of the corpus to the fifth power. These lifts of larger co-occurring subsets drown out the lifts of smaller-sized subsets, which is not necessarily desirable. Thus, we must "normalize" the cardinality of co-occurrence sets. To this end, PheKnow-Cloud calculated the lift for any subset that occurred in the corpus without regard to whether the subset occurred in a phenotype, separated the lifts by the cardinality of the subset, computed the standard deviations above the median within that cardinality, aggregated all the standard deviations above the median values back into the respective phenotypes, and averaged the standard deviation values for each phenotype. This average served as the "clinical relevance score" for that phenotype. This implies that the relevance score will vary

depending on the phenotype corpus, as phenotype scores are relative to other candidate phenotypes.

PIVET mitigates this issue inherent to PheKnow-Cloud normalization by including the number of tuples with no co-occurrences. The number of subsets that had zero occurrences in the corpus is calculated using a simple combinatorial formula:

$$Size(zeros\ for\ phenotype\ j)\ = \sum_{i=1}^{S^j} \binom{S^j}{i} - size(cooccurrences\ of\ cardinality\ i\ ),$$

where S^j is the number of phenotypic items in phenotype j. Including the zero occurrence counts for each cardinality pulls down the overall lift of the larger items (since it's improbable that large subsets of the phenotype will occur) and thus mitigates the impact of larger co-occurring subsets. Consequently, PIVET avoids the need to pool the phenotypic items across all the phenotypes and avoids unnecessary co-occurrence queries for tuples that do not occur in a phenotype. Perhaps more importantly, this implies that the relevance score is decoupled from the phenotype corpus and can be computed independently for a given phenotype.

The final step in the process is to classify the relevance of the phenotype. We compared four separate classification models: logistic regression, logistic regression with Lasso (Least Absolute Shrinkage and Selection Operator), Ridge logistic regression, and K-nearest neighbors, on the entire phenotype corpus to predict clinically significant versus not clinically significant. Gold and silver standard phenotypes are denoted as clinically significant due to their relatively small numbers. The features we use are lift mean, lift median, and lift standard deviation for each individual cardinality from 1, 2, 3, and 4 (12 features). We also include the overall lift mean, median, and standard deviation (3 features), and the average cardinality of subsets of the phenotype with non-zero co-occurrences (16 features in total). Model-specific parameters (i.e., K for K-nearest neighbors and the regularization parameter for Ridge and Lasso) are chosen based on the best area under receiver operating characteristic (AUC) via 5-fold cross validation.

In sum, the PIVET lift analysis differs from that performed by PheKnow-Cloud in two key ways. First, we eliminate the need to pool the lifts across the entire phenotype corpus, which means that phenotypes can be analyzed on an individual basis. Second, we introduce classification models to determine relevance based on lift-based features, removing the need to perform an exhaustive search to determine the clinical relevancy threshold.

Results

PIVET is evaluated using two different methods. The first compares the new framework with its predecessor, PheKnow-Cloud, on the set of phenotypes PheKnow-Cloud examined. Differences in computation time, synonym generation, and clinical relevance scores are quantitatively and qualitatively examined. The second method incorporates the

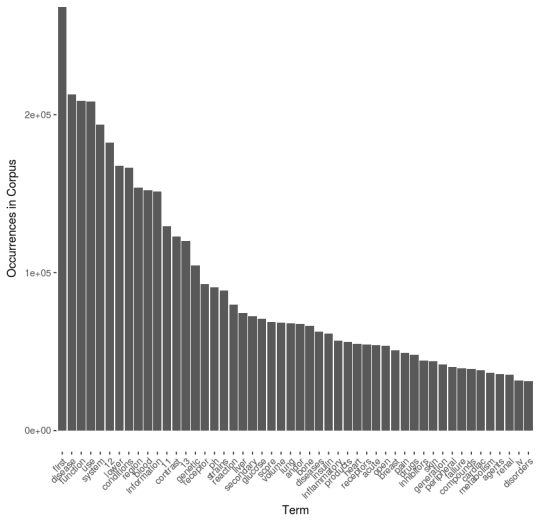
gold and silver standard phenotypes and shows that our classification method can be used to identify clinically relevant phenotypes from the pool of possibly clinically relevant phenotypes.

PheKnowCloud vs PIVET Comparison

A subset comprising one quarter of the PMC OA corpus is used to compare our framework to PheKnow-Cloud. This subset is identical to the one used in the original evaluation of Pheknow-Cloud (see [1] for more details regarding the construction of the dataset). We limit this subset to articles with MeSH terms, which results in a corpus that encompasses approximately 8% of the PMC OA subset. We restrict the phenotypes in question to the Marble and Rubik phenotypes used in the original papers [1] [2].

PIVET takes less than 2 hours to evaluate 80 phenotypes on the 25% PMC OA subset; PheKnow-Cloud required 17 hours for the same phenotypes. The breakdown of the computation time for the major components of the two frameworks is shown in Table 1. The phenotypic item representation process time is roughly the same for both PIVET and PheKnow-Cloud, because querying the NLM MeSH database remains the bottleneck. However, PIVET is 170 and 35 times faster for the corpus analysis and clinical relevance determination steps, respectively. Not only does PIVET provide an overall speedup of 10 times on the same article corpus, but the entire process does not need to be repeated to analyze new phenotypes.

Figure 4. Most common synonyms found in corpus using PheKnow-Cloud synonym generation process.



As discussed in an earlier section, the phenotypic item representation is different between the two frameworks. PIVET uses sets of MeSH terms to represent each phenotypic item while PheKnow-Cloud's representative synonym sets are built from several ontologies that include the MeSH terms. Qualitatively, we found PheKnow-Cloud's synonym sets produced a sizeable number of less descriptive words. Overall, PIVET seems to find more descriptive, discriminative, and possibly more interpretable representations of phenotypic items. Figure 4 shows the top 50 PheKnow-Cloud generated synonyms that

were found in the corpus. While PheKnow-Cloud excludes the first 30 most common terms from its co-occurrence analysis, the remaining 20 words are not discriminative. For example, the word "diseases" is associated with many of the phenotypic items but is too generic to be a meaningful representation of the items.

Further qualitative evidence of the non-specific nature of the synonym sets produced by PheKnow-Cloud can be found by consideration of examples. Table 2 show the synonyms for the phenotypic item "unspecified chest pain". Under the PheKnow-Cloud framework, while discriminative terms like "unspecified chest pain" and "chest pain" are present in the synonym set, the terms "pain," "chest," and "unspecified" are words that will be present in many articles that do not actually refer to "unspecified chest pain". In contrast, under the PIVET framework, the MeSH term for "unspecified chest pain" is "Chest Pain," which while less specific than the original term, has the advantage that it will only be found in articles that mention chest pain.

Table 2. Comparison of representation of the phenotypic item "unspecified chest pain" generated by PheKnow-Cloud (left column) and PIVET (right column).

PheKnow-Cloud (Synonyms)	PIVET (MeSH Terms)
unspecified chest pain	Chest Pain
chest pain	
unspecified chest	
pain	
chest	
unspecified	

In some cases, the synonym sets are reasonable representations of the item and similar for both frameworks. For example, PIVET and PheKnow-Cloud can capture the meaning of the phenotypic item "laxatives" (shown in Table 3). PheKnow-Cloud extracts synonyms that are close literal matches to the phenotypic item or specific kinds of laxatives. Similarly, PIVET finds a MeSH term that is an exact match to the phenotypic item and a specific example of the phenotypic item. When looking through the corpus for occurrences of the original term "laxatives", both frameworks should recover mentions of the original term.

Table 3. Comparison of representation of the phenotypic item "laxatives" generated by PheKnow-Cloud (left column) and PIVET (right column).

PheKnow-Cloud (Synonyms)	PIVET (MeSH Terms)
laxatives	Laxatives
laxatives pharmacological action	Senna Extract
psyllium	
senna	
senna extract	

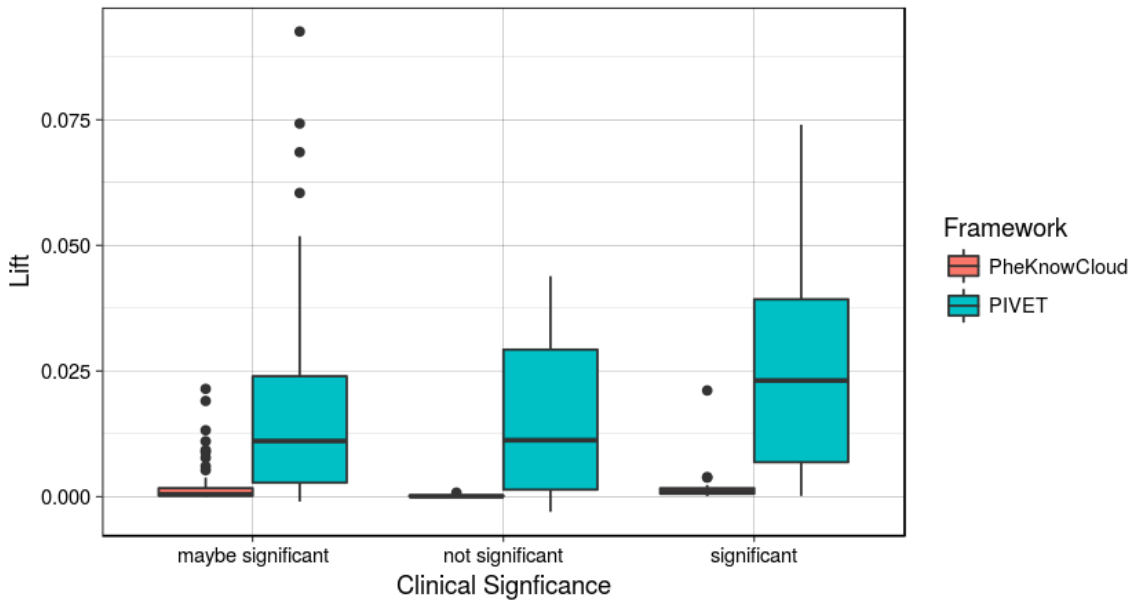
We instrumented PIVET to record co-occurrences in the same manner as PheKnow-Cloud. Table 4 summarizes the number of articles that are found under each framework. Although the PIVET MeSH representation identifies significantly fewer articles from the corpus, the articles have an 85% overlap with PheKnow-Cloud articles. In conjunction with Figure 4 and Table 2, the results suggest that not all of the PheKnow-Cloud articles are relevant or directly related to the phenotypic item. Thus, PIVET synonym sets may result in higher precision.

Table 4. Number of articles that each framework's synonym generation process found.

	Number of Articles
PIVET Synonyms	28,068
PheKnow-Cloud Synonyms	79,786
PIVET and PheKnow-Cloud Synonyms	23,901

Next, we replaced the PIVET lift analysis with the normalized lift analysis from PheKnow-Cloud to compare differences in the clinical relevance scores between the ways of generating synonym sets. Figure 5 plots the pooled normalized lift values for the 80 phenotypes based on the annotated significance level. Under the PIVET representation, there is a larger difference in the normalized lift between significant and not significant phenotypes compared to the PheKnow-Cloud synonym representation. PIVET retains most of the discriminative power of the original framework PheKnow-Cloud but at a fraction of the computation time.

Figure 5. Lift comparison between PIVET and PheKnow-Cloud



PIVET's Classification Score Evaluation

We evaluated the ability of the PIVET classification system to identify clinically significant phenotypes. The entire phenotype corpus, including the gold and silver

standard phenotypes, are analyzed against the entire PMC OA corpus. There is ambiguity regarding the possibly significant Marble and Rubik phenotypes, and they were therefore excluded from the training set. Thus, a total of 45 phenotypes were used to build the classifier, with 7 annotated as not significant.

The diversity of the phenotypes in our corpus yielded phenotypes that contained anywhere from 3 to over 63 phenotypic items. The size of the phenotype sets impacted the cardinality of the non-zero co-occurrence tuples, thus we limited the lift summary features to only include tuples up to 4 (the average across the phenotype corpus). Figure 6 illustrates the differences in the mean lift values between the various categories, with the gold and silver standard separated from the clinically significant group. The results show that the phenotypes that are clinically significant exhibited a higher (more positive) distribution in lift mean compared to the non-significant phenotypes. Moreover, for co-occurrence cardinality less than 5, gold standard phenotypes generally had a higher lift. The figure suggests the suitability of using the mean lift of tuples of cardinalities 2, 3, and 4 as individual features to distinguish the clinical significance of a phenotype.

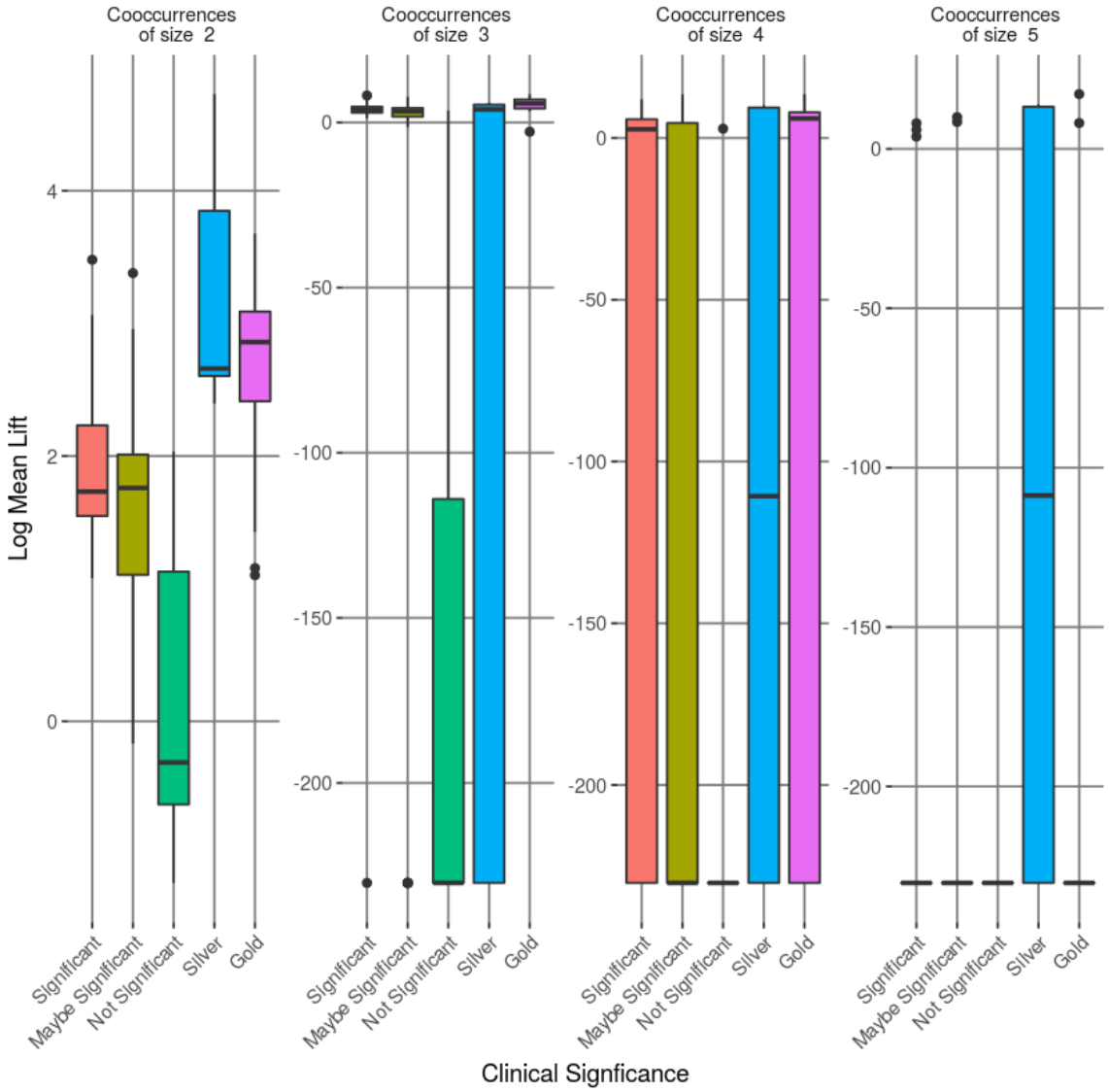


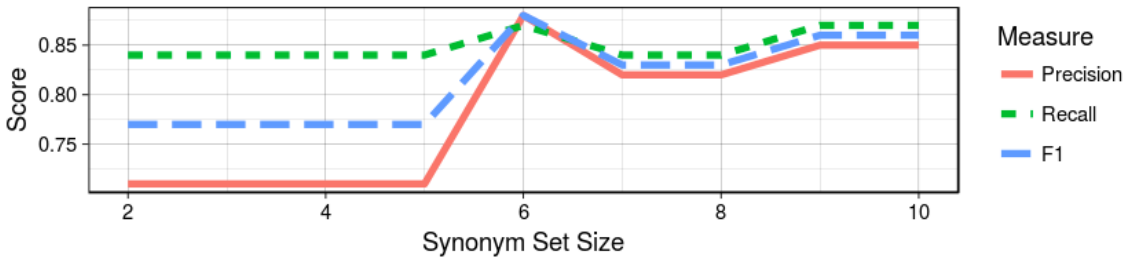
Figure 6. Log mean lift for co-occurrences of sizes 2, 3, 4, and 5 for each type of phenotype

Next, we analyzed the effect of synonym set using a logistic regression model. For each synonym set size ranging from 2 to 10, we used 5-fold cross-validation to examine how the size of the synonym set generalizes to an unseen dataset for different metrics. Figure 7 plots the average precision, recall, and F1 score as a function of the synonym set size where F1 is defined as

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The figure shows significant increases for all three metrics at synonym size 6, at which point an F1 score of 0.89, recall rate of 0.89, and a precision score of 0.88 are achieved. Based on these results, we used synonym set size of 6 for the remaining analysis.

Figure 7. Classification scores for different sizes of synonyms using the PIVET framework.



	Logistic Regression	K-Nearest Neighbor	Lasso	Ridge Regression
AUROC	0.79	0.72	0.33	0.60
F-1	0.87	0.90	0.77	0.91

We repeated the classification process using four models (logistic regression, k-nearest neighbor, logistic regression with LASSO, and logistic regression with a Ridge term) with 6 MeSH term synonyms for each phenotypic item. Of the four classification models, Ridge regression achieved the highest F1 score of 0.91 and an area under the Receiver Operating Curve (AUROC) score of 0.61. Based on these results, we use Ridge regression as our classification model for the remaining results. Incorporating a classification model into the framework is an improvement over PheKnow-Cloud, which depended on an exhaustive search to obtain a boundary between clinically relevant and not clinically relevant phenotypes.

PIVET Analysis of Possibly Clinically Significant Phenotypes

We demonstrate the potential of using PIVET to annotate phenotypes by examining the 57 "possibly clinically significant" phenotypes in our phenotype dataset. Using the PIVET classification Ridge model, we predicted the clinical relevance scores of these ambiguous phenotypes. Table 5 shows the two extremes based on the averaged prediction score: phenotypes with the highest probability of being "clinically significant" (top two rows) and phenotypes with the lowest probability of being "clinically significant" (bottom two rows) as well as the annotator's comment on the phenotype and the average lift calculated by PIVET. The prediction scores seem to reflect the annotator's certainty, as the lowest prediction score is associated with a question mark, while the top two scoring phenotypes seem to capture a relevant concept. The results underscore the potential of PIVET system to help resolve uncertainties.

Table 5: Diagnoses and medications for candidate phenotypes along with domain expert annotations, classification score, and lift for two possibly significant phenotypes with high (top two rows) and low (bottom two rows) classification scores.

Diagnoses	Medications	Comment	Score	Lift
hypotension, heart failure, cardiac dysrhythmias, unspecified chest pain, ischemic heart disease, hypertension,	statins, proton pump inhibitors, gabapentin, non-cardioselective beta blockers, sodium, group v antiarrhythmics, potassium-sparing	The arrythmic heart patient	1	317.380

cardiomyopathy	diuretics			
disorders of fluid, electrolyte, and acid-base balance, other and unspecified anemias, hypertensive chronic kidney disease, hypertension, diabetes mellitus, type 2, other disorders of kidney and ureter, chronic kidney disease (ckd)	antiadrenergic agents, centrally acting, angiotensin receptor blockers, angiotensin converting enzyme inhibitors, selective immunosuppressants, loop diuretics, gabapentin	Heading towards dialysis	0.999	24683.383
Volume depletion; dehydration, Nausea and/or vomiting, Hypopotassemia, Abdominal pain	heparins, antihistamines, 5HT3 receptor antagonists, minerals and electrolytes, narcotic analgesic combinations, proton pump inhibitors	Gastroenteritis	0.418	0.270
disorders of fluid, electrolyte, and acid-base balance, other diseases of lung, hypotension, pleurisy, atelectasis and pulmonary collapse, unspecified chest pain, other disorders of kidney and ureter	anticholinergic bronchodilators, loop diuretics	Lung diseases?	0.417	0.509

Discussion

Automated, high-throughput phenotype methods have been proposed to help clinicians quickly characterize and understand vast amounts of healthcare data. However, the potential for these computational phenotypes to help physicians reason about patient populations will only be realized if the identified phenotypes are clinically meaningful. To increase the utility of such data-driven phenotype discovery, some measure of inferred clinical meaningfulness should be reported to help clinicians sort the signal from the noise. We developed PIVET to meet this need. PIVET generates evidence sets and

clinical relevance scores for data-driven candidate phenotypes using the literature available in PubMed, a large online repository of biomedical articles.

We compared our framework with PheKnow-Cloud, its predecessor, and showed that the predictive performance was similar but that PIVET improves the run-time dramatically. In addition to scaling up to the entire PMC OA corpus, PIVET can analyze phenotypes individually and automatically assign clinical relevance scores that are independent of the other phenotypes in the corpus. Furthermore, there was anecdotal evidence that the PIVET synonym generation process was more discriminative and meaningful than its PheKnow-Cloud counterpart. In the future, one goal is to make PIVET available to researchers and clinicians. To this end we plan to deploy a live version of the phenotype parser that users can interact with via a REST API and receive phenotype JSON files in return. We are currently investigating the best way to release PIVET for general use.

One possible way to improve PIVET is to include more phenotypes when training the classifier. We continue to gather domain expert annotated phenotypes to include in the framework. One limitation of the current analysis was that all the gold and silver standard phenotypes were combined with the domain-expert labelled for classification purposes. As we continue to gather more gold and silver phenotypes, we plan to refine the classification process by incorporating this information. We also plan to test new sets of features that incorporate interaction between the lift statistics as well as examine different metrics for evaluating the clinical significance of candidate phenotypes.

Acknowledgements
Please include all authors' contributions, funding information, financial disclosure, role of sponsors, and other acknowledgements here. This description should include the involvement, if any, in review and approval of the manuscript for publication and the role of sponsors. Omit if not applicable.

Conflicts of Interest
None declared.

Abbreviations
PIVET: Phenotype Instance Verification and Evaluation Tool
MeSH: Medical Subject Heading
NLM: National Library of Medicine
NLP: Natural Language Processing
PMC OA: Pubmed Central Open Access

References

1. Bridges R, Henderson J, Ho JC, Wallace BC, Ghosh J. Automated Verification of Phenotypes Using PubMed. In 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2016; Seattle, WA. p. 595--602, DOI: 10.1145/2975167.2985844.

2. Henderson J, Bridges R, Ho JC, Wallace BC, Ghosh J. PheKnow-Cloud: A Tool for Evaluating High-Throughput Phenotype Candidates using Online Medical Literature. In AMIA Joint Summits on Translational Sciences; 2017; San Francisco, CA.
3. (NCBI) NCfBI. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2016 January;(44): p. D7-D19, DOI: 10.1093/nar/gkv1290.
4. National Center for Biotechnology Information. PMC FAQs. [Online]. [cited 2017 9 11. Available from: https://www.ncbi.nlm.nih.gov/pmc/about/faq/, http://www.webcitation.org/6tPHmk6mp.
5. Boland MR, Shahn Z, Madigan D, Hripcsak G, Tatonetti NP. Birth month affects lifetime disease risk: a phenome-wide method. Journal of the American Medical Informatics Association. 2015 June; 22(5): p. 1042--1053, DOI: https://doi.org/10.1093/jamia/ocv046.
6. Ananiadou S, Kell DB, Tsuji Ji. Text mining and its potentialapplications in systems biology. TRENDS in Biotechnology. 2006 December; 24(12): p. 571--579.
7. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews. Genetics. 2006 February; 7(2): p. 119--129. DOI: 10.1038/nrg1768.
8. Frisch M, Klocke B, Haltmeier M, Frech K. LitInspector: literature and signal transduction pathway mining in PubMed abstracts. Nucleic acids research. 2009; 37(Supp 2): p. W135-W140.
9. Rani J, Shah AR, Ramachandran S. pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. Journal of Biosciences. 2015 October; 40(4).
10. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. ALIBABA: PubMed as a graph. Bioinformatics. 2006 July; 22(19): p. 2444-2445, DOI: https://doi.org/10.1093/bioinformatics/btl408.
11. Lamy JB, Venot A, Duclos C. PyMedTermino: an open-source generic API for advanced terminology services. In MIE; 2015; Madrid. p. 924--928.
12. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 March; 11(11): p. 1422-1423, DOI: https://doi.org/10.1093/bioinformatics/btp163.
13. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford). 2011 January; 2011(baq036).
14. Kim J, Kim Jj, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Scientific Reports. 2017; 7: p. DOI: 10.1038/srep40154.
15. Alnazzawi N, Thompson P, Batista-Navarro R, Ananiadou1 S. Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. BMC Med Inform Decis Mak. 2015 June; 15(Suppl 2): p. S3, doi: 10.1186/1472-6947-15-S2-S3.

16. Collier N, Groza T, Smedley D, Robinson PN, Oellrich A, Rebholz-Schuhmann D. PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. Database (Oxford). 2015; 2015: p. bav104, DOI: 10.1093/database/bav104.
17. Kho A, Hayes M, Rasmussen-Torvik L, Pacheco J, Thompson W, Armstrong L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association. 2011; 19(2): p. 212-218.
18. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics. 2004; 20(3): p. 389-398, doi: 10.1093/bioinformatics/btg421.
19. Pletscher-Frankild S, Palleja A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease--gene associations. Methods. 2015 March; 74: p. 83--89, DOI: 10.1016/j.ymeth.2014.11.020.
20. Xu D, Zhang M, Xie Y, Wang F, Chen M, Zhu KQ, et al. DTMiner: identification of potential disease targets through biomedical literature mining. Bioinformatics. 2016 August; 32(23): p. 3619-3626, DOI: https://doi.org/10.1093/bioinformatics/btw503.
21. Hopewell SaLK, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. The Cochrane Library; 2009.
22. Dickersin K. The Existence of Publication Bias and Risk Factors for Its Occurrence. Journal of the American Medical Association. 1990; 263(10): p. 1385-1389.
23. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technology Assessment. 2010 February; 14(8): p. 1--193, DOI: 10.3310/hta14080.
24. Song F, Hooper L, Loke Y. Publication bias: what is it? How do we measure it? How do we avoid it? Open Access Journal of Clinical Trials. 2013 July; 2013(5): p. 71--81, DOI https://doi.org/10.2147/OAJCT.S34419.
25. Ekmekci PE. An increasing problem in publication ethics: Publication bias and editors' role in avoiding it. Medicine, Health Care, and Philosophy. 2017 March; 20(2): p. 171–178, DOI: 10.1007/s11019-017-9767-0.
26. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS one. 2008 August; 3(8): p. e3081. DOI: 10.1371/journal.pone.0003081.
27. Easterbrook PJ, Gopalan R, Berlin J, Matthews DR. Publication bias in clinical research. The Lancet. 1991; 337(8746): p. 867--872.
28. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ. 1197; 315(7109): p. 640--645.
29. Li Y, Manoharan S. A performance comparison of SQL and NoSQL databases. In

Communications, computers and signal processing (PACRIM), 2013 IEEE Pacific Rim; 2013. p. 15--19, Electronic ISBN: 978-1-4799-1501-9 ISSN Information: INSPEC Accession Number: 13838793 DOI: 10.1109/PACRIM.2013.6625441.

30. Boicea A, Radulescu F, Agapin LI. MongoDB vs Oracle--database comparison. In Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on; 2012. p. 330--335, ISBN Information: INSPEC Accession Number: 13169567 DOI: 10.1109/EIDWT.2012.32.
31. Indrawan-Santiago M. Database research: Are we at a crossroad? Reflection on NoSQL. In Network-Based Information Systems (NBiS), 2012 15th International Conference on; 2012. p. 45--51, DOI: 10.1109/NBiS.2012.95.
32. Ho JC, Ghosh J, Sun J. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014; New York, NY. p. 115--124, DOI: 10.1145/2623330.2623658.
33. Wang Y, Chen R, Ghosh J, Denny JC, Kho A, Chen Y, et al. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. In 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015; Sydney, NSW, Australia. p. 1265-1274, DOI:10.1145/2783258.2783395.
34. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association. 2016 March 28.
35. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. The American Journal of Human Genetics. 2010 Apr; 86(4): p. 560-572.
36. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000; 88(3): p. 265-266.
37. Aho AV, Corasick MJ. Efficient string matching: an aid to bibliographic search. Communications of the ACM. 1975; 18(6): p. 333-340, DOI: 10.1145/360825.360855.
38. Névéol A, Islamaj-Doğan R, Lu Z. Semi-automatic semantic annotation of PubMed Queries: a study on quality, efficiency, satisfaction. Journal of Bioinformatics. 2011 April; 44(2): p. 310--318, DOI: 10.1016/j.jbi.2010.11.001.