Automated Verification of Phenotypes using PubMed

Ryan Bridges
Epic Systems
Verona, WI 53593
rybridges90@gmail.com

Jette Henderson
The University of Texas at Austin
Austin, TX 78712
jette@ices.utexas.edu

Joyce C Ho
Emory University
Atlanta, GA 30322
joyce.c.ho@emory.edu

Byron C. Wallace
Northeastern University
Boston, MA, 02115
byron@ccs.neu.edu

Joydeep Ghosh
The University of Texas at Austin
Austin, TX 78712
jghosh@utexas.edu

ABSTRACT

In the realm of data driven clinical research, medical concepts, or phenotypes, are used to serve as indicators for patient clusters of interest. Often, studies will use groups of algorithmically generated phenotypes (feature groups) to predict the occurrence of heart disease, diabetes, and other conditions. When these groups are algorithmically generated, the most common method of verification is manual human annotation, which can be time consuming and sometimes inconsistent. In this paper, we propose a supervision-free method of verification that uses co-occurrence in PubMed articles to determine clinical significance. We demonstrate the efficacy of the method by 1) applying it to phenotypes generated through automatic machine learning methods and 2) showing it separates randomly generated groups of phenotypes from curated groups of known, clinical phenotypes.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; I.2.7 [**Natural Language Processing**]: Text analysis; J.3 [**Life and Medical Sciences**]: Health

General Terms

Phenotype verification, Pubmed, co-occurrence analysis

1. INTRODUCTION

Computational phenotyping is the practice of mapping the raw information contained in Electronic Health Records (EHRs) into sets of clinically relevant features, or phenotypes. Clinicians can use the EHR-based phenotypes to identify patients with specific characteristics or conditions of interest. Phenotypes also enable cohort identification to target patients for screening tests and interventions, support surveillance of infectious diseases, and aid in the conduct of pragmatic clinical trials and comparative effectiveness research [14]. An example is the type 2 diabetes mellitus phenotype (shown in Figure 1) [6]. The flowchart depicts a series of characteristics that must be present in a patient's EHR for that patient to be identified as a type 2 diabetes case patient.

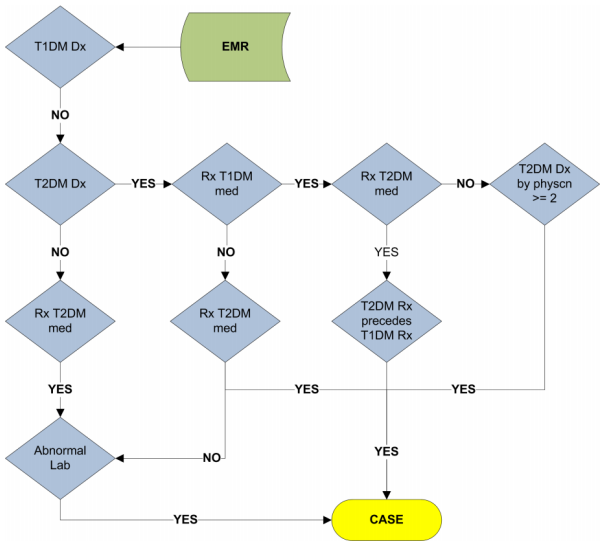

Figure 1: Type 2 diabetes phenotype from the Phenotype Knowledgebase [6], Source: https://phekb.org/phenotype/type-2-diabetes-mellitus

Constructing phenotypes can be a manual, iterative, and labor-intensive process requiring domain expertise [2, 3, 9]. Recent efforts have focused on machine learning developed methods to automatically generate candidate computational phenotypes in a high-throughput, unsupervised manner [7, 8, 10, 18, 20]. However, domain experts are still required to annotate these candidate phenotypes to verify the clinical significance, and several issues can arise during the annotation process beyond time–consumption. First, domain experts may disagree on the clinical relevance of a candidate phenotype. Second, unsupervised methods may generate

BCB'16, October 2–5, 2016, Seattle, WA, USA.

http://dx.doi.org/10.1145/2975167.2985844.

phenotypes that are unfamiliar to annotators, so they may incorrectly judge a phenotype as clinically insignificant when it is not. Additionally, given the diverse and different clusters of patients grouped by these methods, annotators may feel the objective or the phenotypes themselves are vague or undefined. Thus, there is a need to develop an automated, data–driven process to serve as an unbiased means of validation, leveraging all the medical expertise that has been collected to date.

In this paper, we present an automated framework that can provide third-party verifcation to facilitate and improve the annotation process and accuracy. Our method records co–occurrences of phenotypic terms in PubMed[1], a publicly available repository for medical literature that contains over 40GB of medical literature from over 6000 journals. From this data, we use lift, a metric that summarizes if two or more items co–occur more often than average while accounting for the commonality of the item, to determine clinical relevance. We analyze the lift patterns of 80 phenotypes produced by several of the high–throughput methods described above and demonstrate the correlation between lift with the significance as judged by domain experts. We also illustrate the upper and lower bounds of the lift metric by comparing phenotypes generated randomly to those curated to represent known medical concepts. We demonstrate the method is agnostic to the algorithm that generated the phenotypes by showing it can effectively determine the validity of candidate phenotypes produced by two different high-throughput algorithms, as well as curated phenotypes. We note however, that if an algorithm itself uses PubMed to generate phenotypes, this method of verification should not be used.

2. RELATED WORK

While many researchers have used PubMed to explore and discover issues in biology, medicine, and health informatics, few have used PubMeb as a validation tool. One such study by Boland et al. mined EHR records for patients who had disease-specific codes and then compared the association between birth month and the disease to a group of control patients who did not have the disease codes present in their EHRs [1]. They validated their results against papers queried from PubMed that had disease and birth month as topics. Neveol et al. did not use Pubmed as a validation tool but they did use it as a tool to generate candidate annotations for PubMed queries and then measured the inter-annotator agreement as well as annotation time between sets of queries with and without the candidate annotations [13]. While they were annotating PubMed in order to understand PubMed users' needs, their work shows that annotating tools can not only speed up annotating time, but increase inter-annotator agreement. However, annotating before annotators can examine the text can have the effect of biasing annotators, so it should be used carefully.

More commonly, researchers PubMed as an exploratory tool. Jensen et al. provide a thorough overview of how PubMed can be harnessed for information extraction and entity recognition [11]. Amongst the two methods they discuss for information extraction, natural language processing and co–occurrence analysis, co–occurrence is more prevalent due to its straightforward implementation and the intuitive interpretation of the results. While co–occurrence analysis does not give information about the type of relationship or any causal information, work done on bias towards publishing positive results allows for the assumption that when two phrases occur together the relationship exists [4, 5, 17]. Researchers have applied co-occurrence strategies to generate phenotypes. Some have used Pubmed to study links between diseases [16], which can be thought of as phenotype discovery, and to explore relationships between phenotypes and genotypes [15]. Having generated phenotypes through machine learning techniques, our work focuses on using co-occurrence analysis of Pubmed as a validation tool for annotations.

[1]http://www.ncbi.nlm.nih.gov/PubMed

3. METHOD

Annotators are often clinicians volunteering their time, and may or may not have computational backgrounds or annotation experience. Furthermore, medical perspectives can be drastically different amongst annotators as they are impacted by factors such as their medical expertise, patient population, and medical education (medical school and residency). In addition to these reasons, the vague and subjective nature of the annotation task can result in low inter–rater agreement amongst the different clinicians, with one high–throughput phenotyping method reporting an inter–rater agreement of 0.81 [18]. We propose to leverage the 26+ million biomedical literature citations found in PubMed as an objective third–party annotator by developing an automated method to capture co-occurrence of phenotypic items within the clinical narratives described throughout PubMed. Our framework utilizes the inherent publication biases associated with medical literature to observe that if a concept pairing is clinically significant, several papers will make mention of these concepts together in such a large and diverse corpus. This provides a reasonable objective baseline for determining significance that can be used as corroboration for existing annotation or as a tool to assist annotation efforts.

Although the idea is conceptually simple, there are several challenges that our automated framework must address. The representation of each element of the phenotype is important as it can drastically impact the number of articles returned during the PubMed query. Second, the co–occurrence search needs to account for encoding, form/tense, incorrect spellings, and also regularization. Finally, the co–occurrence metric should reflect the number of items contained in the phenotype elements as well as the commonality of the item itself in the PubMed literature. Thus, our automated verification process consists of several steps:

1. Feature (n-gram) generation from phenotypes
2. Counting co-occurrence in PubMed articles
3. Calculating and normalizing feature lifts within a particular group to determine significance

Figures 2 and 3 show the process for a phenotype from the feature generation to the calculation of the clinical significance.

3.1 Feature Extraction

Since medical terms can have multiple synonyms and representations across different articles, our framework first generates a suitable list of synonyms and related concepts for each element in the phenotype (each phenotypic item). The

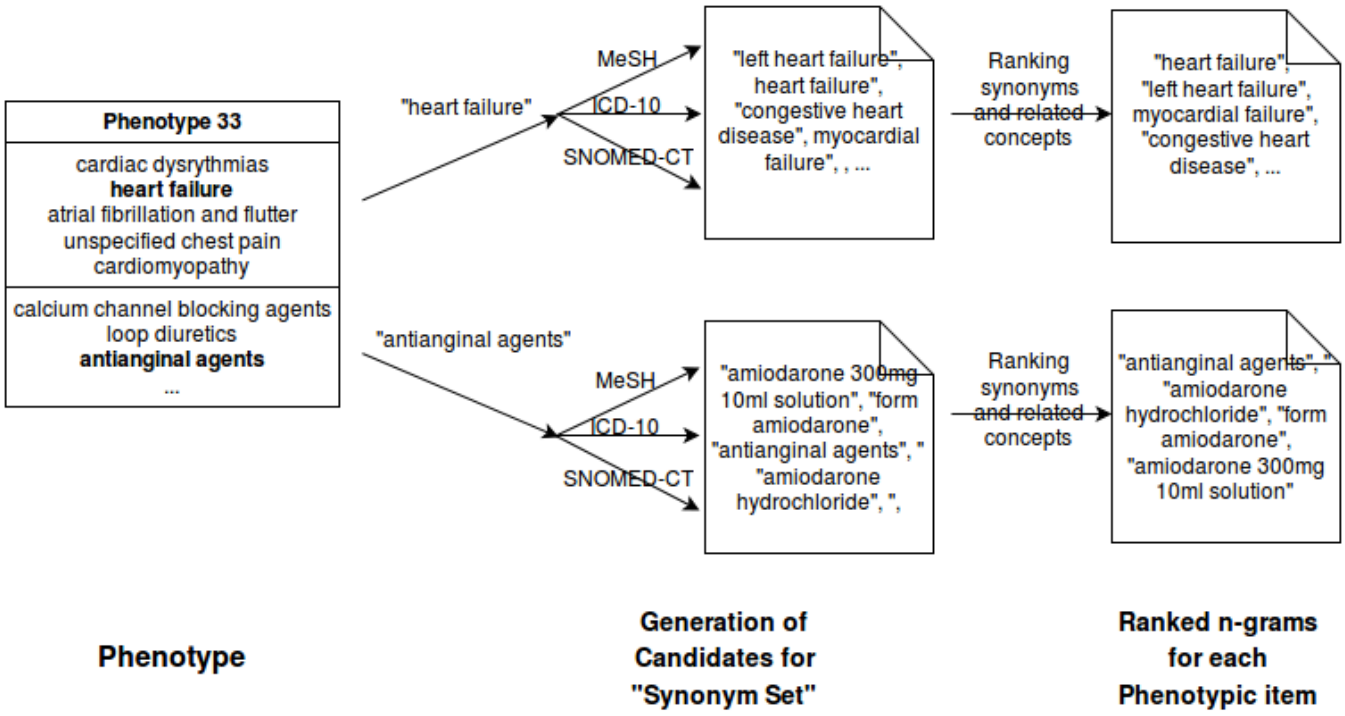

Figure 2: Feature extraction for Phenotypic Items

leftmost box of Figure 2 displays an example of a candidate phenotype, which was generated by an automatic method, that could be presented to the annotators. In this example, the term "heart failure" can also be referred to in various articles using other terms such "myocardial failure", "congestive heart disease", or may be even be referred to in the context of more general terms such as "cardiovascular health". Thus, the appropriate terms must be used so that heart failure as a concept may be discovered in a PubMed article reasonably often when it is being mentioned (recall); however, the terms may not be so general as to produce many false positives (precision). To produce a representation that has high recall and high precision, we first generate a large set of possible representative n-grams, and then filter all the candidate n-grams down to the most relevant n-grams (the filtration process is discussed in the next section).

A naïve approach is to use the item as it appears rather than to perform the candidate generation and filtration process. However, this can yield low recall as the text is often too specific or not phrased naturally. For example, note the phenotypic item "calcium channel blocking agents" in the phenotype in Figure 2. When searching through text, it may be more appropriate to shorten the phrase to "calcium channel blockers" or even alternatively use the phrase "hypertension medications."

Likewise, using a collection of individual words (unigrams) to represent phenotypes yields high recall, but low precision, as it is difficult to filter out enough of the words that lack specificity, but are important in some cases (e.g. "disease" or "results"). "Calcium", "channel", "blocking", and "agents" will obviously find all occurrences of "calcium channel blocking agents," but will also capture mentions having little to do with the subject, resulting in low precision.

In order to achieve high recall, a phenotypic item must be recognized when a conceptually equivalent or similar term is present and capture situations when an entirely dissimilar string is used to represent the same item (e.g. "heart attack" and "myocardial infarction"). In order to recognize such "aliases," we utilized several medical ontologies to collect a set of closely related concepts and synonyms to a given phenotypic item. One of the most complete and commonly used ontologies is the "Systemized Nomenclature of Medicine - Clinical Terms" (SNOMED-CT) ontology [19]. The first order connections on the SNOMED-CT ontology graph for a concept provided a reasonable number of aliases that we could then filter. We supplemented SNOMED-CT with two other common ontologies, ICD–10 and the NCBI MeSH terms. During implementation, we extracted the SNOMED-CT and ICD–10 ontologies with a python library called Pymedtermino.[2] Biopython,[3] a tool which also provides an interface to the Entrez tools, provided access to the MeSH terms. We queried the Pymedtermino and Entrez APIs to collect aliases from these three ontologies, and then placed the related terms from each ontology into a (potentially large) list of candidate concepts that may represent the phenotypic item. After assigning every phenotype a pooled set of related concepts, we removed stopwords, and then extracted the set of all unigrams, bi-grams, and tri-grams from the related concepts.

Figure 2 shows the feature generation process for the phenotypic items "heart failure" and "antianginal agents". For "heart failure", our framework generates the related concepts "congenital heart disease", "left ventricular structure", "myocardium", and "heart valve disorder" as indicated by the

[2]http://pythonhosted.org/PyMedTermino/
[3]http://biopython.org/

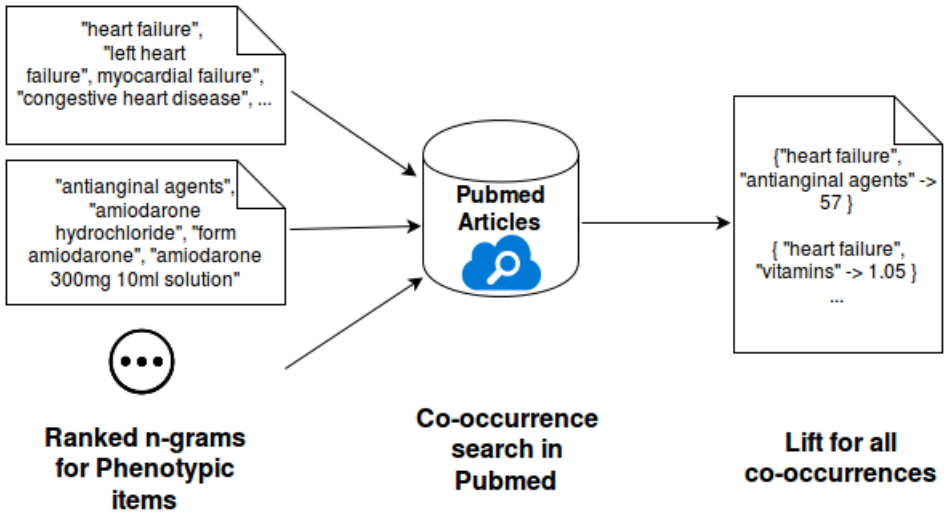

Figure 3: Significance calculation for phenotypic items

middle column of the figure. In the case of the term "antianginal agents", the algorithm generates "thyroid structure", "hydrochloride", "morphologic abnormality" and "penbutolol product" as potential n–grams. In order to ease computational burden further down the pipeline, in the scenario when a phenotypic item contains many (greater than 250) related n-grams, a subset of 250 were randomly selected. The choice of 250 n-grams allowed sufficient coverage of the related concepts, while allowing the computation time of the filtration step to remain feasible.

3.2 Selection and Filtration of candidate n-grams

After extracting all possible n–grams (unigrams, bigrams, and/or trigrams) relating to a phenotypic item, our framework determines the n–grams that are most related to the phenotypic item, which we refer to as the selection and filtration process. This process orders the n–grams generated from the previous process (e.g., the first and middle column of 2) by "relevance." We can then tune the trade-off between precision and recall by adjusting the number of relevant n–grams we use to represent each phenotypic item. We explore the trade-off between precision and recall in Section 4.

We define the "relevance" as the percentage overlap between the sets of PubMed query results from the original phenotypic item and each of its representative n-grams and calculate it by 1) recording the set of papers returned by each query and 2) finding the size of the intersection between the set of papers returned for the original item and each of the subsequent n–gram queries. We tried Word2vec [12] as a semantic similarity measure, but the empirical results generated more false positives than our PubMed querying method. Thus, the semantic similarity of the phenotypic item phrase and its n–grams are roughly measured by the PubMed search index, rather than a more complicated semantic measure.

Each phenotypic item is assigned a ranked list, based on the relevance score, of representative n–grams. We refer to the set of top ranked n-grams as the phenotypic item synonym set. Table 1 shows one example of the original phenotypic item and the ranked list with the eight highest ranked n-grams and their associated relevance scores. For example, the first synonym 'lacrimal apparatus diseases' of the second item 'Disorders of lacrimal system' has a relevance of .636, which means these phrases appeared together in 63.6% of Pubmed searches for each phrase separately. Based on our experimental results, we found selecting fifteen or even ten of the top ranked n–grams produced a suboptimal number of false positives. In Section 4 (Figure 6), we show using six of the top ranked n-grams gives a tolerable number of false positives. In addition to restricting each representation to six n-grams, we pared down the list of aliases even more by ordering the set of all n–grams by their sentence frequency in PubMed, as well as their interaction frequency with other phenotypes, and removing the most frequent 5% from the sentence frequency and interaction frequency lists. More work on consensus filtration, however, is merited, but these choices reflect the need to keep the framework as computationally efficient as possible.

3.3 Co-occurrence search in PubMed

NCBI has a publicly available download of PubMed. For computational reasons, we used a randomly selected subset of 25% of the articles available in PubMed for this analysis. In the future we plan to scale to more articles, select the subset based on when the articles were published, and select the subset based upon the journal's impact factor. In the random subset, we searched for occurrences of elements of the phenotypic item synonym sets (generated in the last section) for all items within each phenotype. For all articles in the subset, any sentence containing one or more of the n-grams from any phenotypic item was noted, and the set of n-grams appearing in the sentence was added to a master list of all co-occurrences. Each sentence was minimally processed, only regularizing capitalization and encoding (utf-8), taking out words included in NLTK's English stopword list, using a conservative regular expression to remove references (e.g. Smith, et al.), and removing special characters like quotes and parenthesis. The form/tense and spelling of words were left as written to be consistent with the n-grams derived from the ontology related phrases.

Table 1: One phenotypic item and its associated top eight most highly ranked representative n–grams. The score represents the percentage overlap of Pubmed searches between the term that appears in the phenotype and the "synonyms" extracted from various ontologies.

Original representation	Ranked list of n–grams
'Angiotensin-converting enzyme inhibitors'	('angiotensin-converting enzyme, inhibitor', 0.858) ('reaction ace inhibitor', 0.214) ('due ace', 0.207) ('hyperkalaemia due angiotensin-converting', 0.138) ('angiotensin-converting-enzy me inhibitor allergy', 0.082) ('inhibitor induced hyperkalemia', 0.071) ('antihypertensive drug disorder', 0.065) ('antihypertensive agent disorder', 0.065)
'Disorders of lacrimal system'	('lacrimal apparatus diseases', 0.636) ⋮

Any occurrence of an n-gram that is a part of the set of 3 to 6 n-grams (6 is the target, some phenotypic items have 3, 4, or 5 based upon filtration criteria) representing a phenotype counted as an appearance for its phenotypic item. This simplifying assumption eliminated the need to weigh n-grams by their "relevance." The assumption also ignores if a sentence contains more than one n-gram for an item. Using this measure of co-occurrence, the lift of every co-occurrence of phenotypic items was calculated. Recall that given A, B, and C in a sentence:

$$\text{lift}(A, B, C) = \frac{P(A \cap B \cap C)}{P(A) * P(B) * P(C)}$$

Probabilities are calculated as the number of sentences where the item occurs divided by the total number of sentences. While studying all possible combinations of phenotypic items may be interesting for identifying significant subsets of phenotype groups or connections between phenotypes, which we briefly discuss in 5, we primarily examine the average lift of phenotypic items within a given phenotype. This average lift of the phenotypic item co-occurrences serves as our primary metric, and allows for a simple classification of a phenotype as "clinically significant" or "not clinically significant"–this classification is discussed below.

For every co-occurrence, all possible subsets of co-occurring phenotypic items within phenotypes were also counted. For example, when A, B, C, and D co-occurred in a sentence, a co-occurrence for (A,B,C), (B,C,D), (A,B), (A,C), and so forth, were counted. In this way, the lifts for any combination in the power set of all phenotypic items that co-occurred were counted. This allows for convenient lookup of any co-occurrence of interest. While this complete set of co–occurrences is theoretically very large, not every combination of phenotypes returns a non-empty search.

We made the assumption that co-occurrences including more phenotypic items (we refer to this as higher "phenotype cardinality" for convenience; see the third column of Figure 4 for examples of this "cardinality") should be favored. That is, the more of the phenotype that is represented, the more can be said about the significance of the phenotype as a whole. To serve this preference, we counted co-occurrences with the largest phenotype cardinality first, and then ignored any co-occurrence that was a subset of any larger co-occurrence. This was achieved by simply ordering the combinations in descending cardinality order, and greedily inserting the combination into a set if it was not a subset of an already counted combination. This choice allows co-occurrences of any size to contribute to the average of a phenotype group but favors interactions including more phenotypes of the group (without double counting their subsets) assuming this is a better representation of the significance of the group as a whole. Note that we exclude empty co-occurrence sets (those with a lift of 0) from the lift average, and in future work, we will consider the tradeoffs of including the lift of these empty sets.

The lift metric divides the probability of co-occurrence by the product of each of the probabilities of the individual terms, which normalizes the rate of co-occurrence by the probability of random chance co-occurrence. If the probability of co-occurrence is higher than the rate of random co-occurrence (assuming independence), then the lift will be greater than 1 and indicates "statistical significance." However, since these co-occurrences are subject to grammatical rules, etc., lifts for co-occurrences are nearly always greater than 1. As we are interested in filtering out all but clinical significance (ignoring significance introduced by grammar and language convention), we randomly generated phenotypes from a set of phenotypic items, and measured the lift significance of these "phenotypes" to establish the level of lift significance introduced by the possible grammatical and language artifacts.

We began analysis by placing phenotypic item co-occurrences into groups by their phenotype cardinality (regardless of the clinical phenotype group membership). We found that, across all co-occurrences among phenotypic items (this includes those from multiple phenotypes), lift was strongly positively dependent on the number of items included in the co-occurrence. In fact, lift appears to be almost perfectly exponential as a function of the number of items included in the co-occurrence, which is illustrated in Figure 5. Thus, we divide each lift by the median of the lifts of that cardinality so that higher cardinality co-occurrences do not dominate the phenotype mean.

We note that the mean and standard deviation of each of these cardinality groups were skewed high, as the max lifts were significantly further from the median than the below-median lifts (Figure 5). Since the standard deviation is artificially increased by the largest lifts, the normalized above

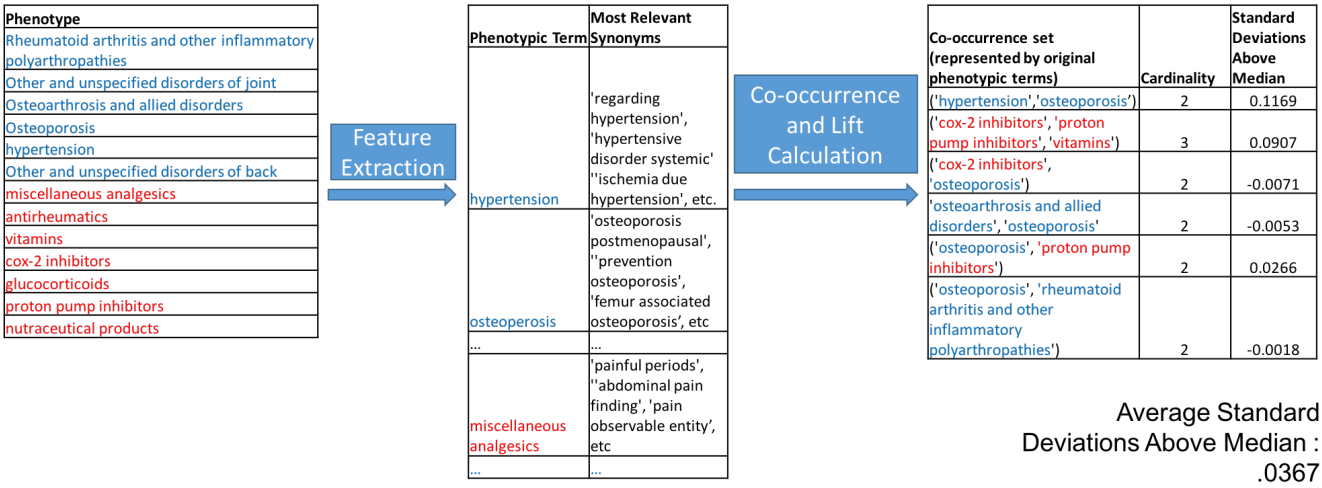

Figure 4: Example of lift calculation process for a phenotype that was found to be significant (i.e., above the threshold of 0.0284)

median lifts are still much greater than zero than the below median lifts are below. This fact makes it so that a lift threshold is likely to be closer on average to the lift of not significant phenotypes than to significant phenotypes. Figure 4 demonstrates the process of extracting the synonyms for the phenotypic terms, calculating the average standard deviations that a set of phenotypic terms is above the median, and then shows the overall average. While the middle column only contains a subset of the phenotypic item synonym set, the last column contains all the combinations of phenotypic terms that have non-zero standard deviations (i.e., the co-occurrence is non-zero). A phenotype is labeled clinically relevant if the average standard deviation of the median is above a chosen threshold.

4. EMPIRICAL STUDY

4.1 Dataset Description

Our study uses randomly generated phenotypes, phenotypes curated to represent known significant clinical narratives and the annotated results of candidate phenotypes generated by different unsupervised, high–throughput phenotype generation processes. The first automatic method, Rubik [18], generated phenotypes from a de–identified EHR dataset from Vanderbilt University Medical Center with 7,744 patients over a five year observation period. For more details about the pre–processing of the data and phenotype generation, please refer to their paper [18]. The authors graciously shared the file with 30 computational phenotypes as well as the annotations of the three domain experts. For each phenotype, each expert assigned one of the following three choices: 1) yes - the phenotype is clinically meaningful, 2) possible - the phenotype is possibly meaningful, and 3) not – the phenotype is not clinically meaningful. The second set of candidate phenotypes were generated by Marble [8] using the EHR data of a random subset of 10,000 patients from the *Centers for Medicare and Medicaid Services (CMS) Linkable 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)*, a publicly available dataset with claim records that span 3 years.[4] The 50 candidate phenotypes that Marble generated were then annotated by two domain experts in a manner identical to above. We combined the 30 Rubik-generated candidate phenotypes with the 50 Marble-generated candidate phenotypes and used the resulting set of 80 candidate phenotypes in the co-occurrence experiment. Of these 80 phenotypes, the annotators found that approximately 14% are clinically meaningful, 78% are possibly significant and 8% are not clinically meaningful.

In addition to the high–throughput generated phenotypes, we used randomly generated phenotypes and curated phenotypes to establish lower and upper bounds, respectively, for the lift score that measures phenotype significance. The random phenotypes are generated by randomly selecting phenotypic items from a set of 1000+ phenotypic items generated by Marble/Rubik phenotypes not used in this work. The curated phenotypes were constructed by representing clinical narratives described in Epocrates references[5] and the AHRQ national guidelines[6] using phenotypic items.

4.2 Results

To determine the optimal size of the phenotypic item synonym set (i.e., the number of "relevant" n-grams to use), we performed a grid search over the set sizes, the results of which are summarized in Figure 6. Figure 6 shows the precision, recall and F1 score for classifying phenotypes to their "significant" or "not significant" annotator labels when characterized by different numbers of n-grams. The choice of 6 n-grams resulted in classification with the best balance between precision and recall, achieving an F1 score of 0.87 (2 N-grams scored 0.88, but had lower precision).

We first examine the normalized lift averages of the randomly generated and curated phenotypes to establish a baseline for the difference between significant and not significant

[4]For more information see https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html
[5]http://www.epocrates.com/
[6]http://www.ahrq.gov/professionals/clinicians-providers/guidelines-recommendations/index.html

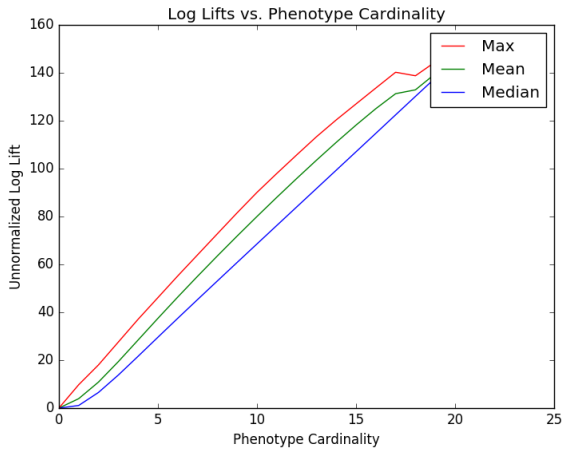

Figure 5: Log Lifts versus Phenotype Cardinality from all Combinations of Phenotypic Items from any Phenotype

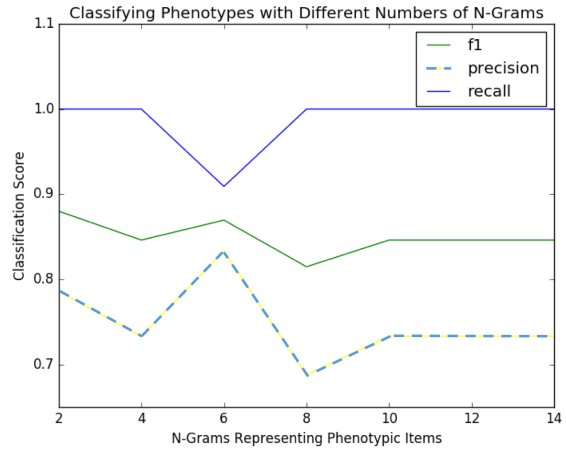

Figure 6: Classification Scores for Marble/Rubik Phenotypes versus size of Synonym Set

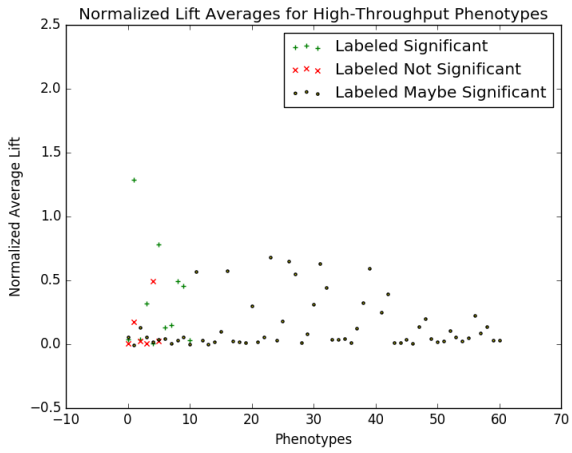

Figure 7: Normalized Average Lift of Marble/Rubik Phenotypes

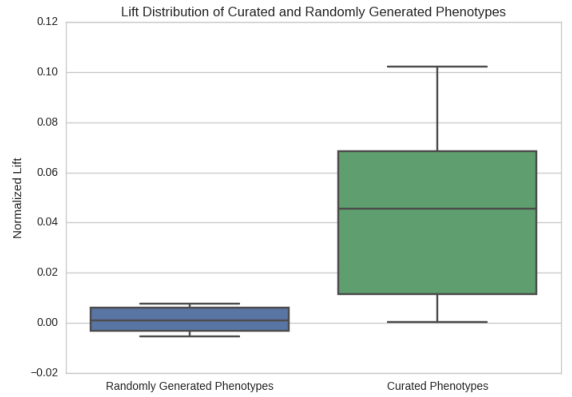

Figure 8: Normalized Average Lift of Curated Phenotypes

phenotypes. Figure 8 shows the distribution of average lifts for the two groups of phenotypes (represented, again, by a synonym set of 6 N-grams). In the majority of cases the normalized lift average of the curated phenotypes is above that of the randomly generated phenotypes. By choosing the optimal threshold, we are able to achieve 100% true negative classification, and 80% true positive classification, or an F1 score of 0.89.

Moving on from the separation of random and curated phenotypes, we applied this analysis to real world phenotypes generated by the high-throughput methods. Figure 7 shows the normalized lift average of the phenotypes generated by Marble and Rubik ([7, 8, 18]). Again, by determining the optimal lift threshold (determined to be 0.028 by exhaustion), we are able to classify "significant" and "not significant" phenotypes with an F1 score of 0.87.

While we do not perform any classification on the possibly significant groups, we plan to delve into this into the future. Annotators could use this analysis as evidence to give labels the phenotypes that may be significant. For example, if an annotator was not sure about the significance, he or she could use the average lift as another piece of information while making the decision. However, a more thorough analysis of the value of this and whether or not it would bias an annotator must be studied.

We note that while lift thresholding classifies phenotypes with relative success in both high-throughput and curated phenotypes, the method does not provide a universal threshold guaranteed for all phenotypes. In addition, the majority of phenotypes are very close to the optimal threshold. This suggests that further work is needed to improve the predictive value of lift thresholding.

5. CONCLUSION

We have presented an automated method for verifying the significance of phenotype groupings using co-occurrence of diagnoses/medications within the phenotype in a corpus of medical literature.

By representing phenotypes as a small set of relevant n-grams and calculating the lift of phenotypic item co-occurrences

in PubMed, we were able to classify a small set of curated phenotypes with an F1 score of 0.89, and a set of phenotypes generated from EHR tensor data with an F1 score of 0.87. While this ground truth set is small, the method shows promise to provide an objective and automated method of verification for arbitrary phenotype groups.

Further, since the item co-occurrences are found in natural language, a set of sentences describing the phenotypic item co-occurrence can be reported and synthesized into a human readable explanation for the significance of the phenotype. Previous work [13] has shown that annotators produce better annotations in less time when starting from pre-annotated results from automatic tools. This implies that in addition to corroborating human annotation, automatic labeling can be used to facilitate the annotation process. We wish to examine this more in the future.

Work to further verify and improve this method is merited, as a reasonably high level of classification accuracy was achieved without complex feature selection, or using co-occurrences from the remaining 75% of available PubMed articles. With these additions, the method could further help improve phenotype annotation quality.

6. REFERENCES

[1] M. R. Boland, Z. Shahn, D. Madigan, G. Hripcsak, and N. P. Tatonetti. Birth month affects lifetime disease risk: a phenome-wide method. *Journal of the American Medical Informatics Association*, page ocv046, 2015.

[2] R. J. Carroll, A. E. Eyler, and J. C. Denny. Naive electronic health record phenotype identification for rheumatoid arthritis. In *AMIA Annu Symp Proc*, volume 2011, pages 189–96, 2011.

[3] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):e253–e259, 2013.

[4] K. Dickersin. The existence of publication bias and risk factors for its occurrence. *Jama*, 263(10):1385–1389, 1990.

[5] P. J. Easterbrook, R. Gopalan, J. Berlin, and D. R. Matthews. Publication bias in clinical research. *The Lancet*, 337(8746):867–872, 1991.

[6] M. G. Hayes, L. Rasmussen-Torvik, J. A. Pacheco, W. K. Thompson, L. L. Armstrong, J. C. Denny, P. L. Peissig, A. W. Miller, W.-Q. Wei, S. J. Bielinski, C. G. Chute, C. L. Leibson, D. R. Crosslin, C. S. Carlson, K. M. Newton, W. A. Wolf, R. L. Chisholm, and W. L. Lowe. 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*, 19(2):212–218, Mar. 2012.

[7] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, Dec. 2014.

[8] J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 115–124, 2014.

[9] G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[10] C. Hu, P. Rai, C. Chen, M. Harding, and L. Carin. Scalable bayesian non-negative tensor factorization for massive count data. In *Machine Learning and Knowledge Discovery in Databases*, pages 53–70. Springer, 2015.

[11] L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv.org*, Jan. 2013.

[13] A. Névéol, R. I. Doğan, and Z. Lu. Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics*, 44(2):310–318, 2011.

[14] NIH Health Care Systems Research Collaboratory. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. July 2014.

[15] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015.

[16] D. K. Rajpal, X. A. Qu, J. M. Freudenberg, and V. D. Kumar. Mining emerging biomedical literature for understanding disease associations in drug discovery. *Biomedical Literature Mining*, pages 171–206, 2014.

[17] J. M. Stern and R. J. Simes. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *Bmj*, 315(7109):640–645, 1997.

[18] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274. ACM, 2015.

[19] H. Wasserman and J. Wang. An applied evaluation of snomed ct as a clinical vocabulary for the computerized diagnosis and problem list. In *AMIA Annual Symposium Proceedings*, volume 2003, page 699. American Medical Informatics Association, 2003.

[20] S. Yu, K. P. Liao, S. Y. Shaw, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane, and T. Cai. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5):993–1000, Apr. 2015.