Analyzing Patterns of Literature-Based Phenotyping Definitions for Text Mining Applications

Samar Binkheder, Ph.D. cand
School of Informatics and Computing, Department of BioHealth Informatics
Indiana University –Indianapolis
Indianapolis, Indiana
sbinkhed@iu.edu

Heng-Yi Wu, Ph.D.
College of Medicine, Department of Biomedical Informatics
The Ohio State University
Columbus, Ohio
Heng-Yi.Wu@osumc.edu

Sara Quinney, Ph.D.
School of Medicine, Department of Obstetrics and Gynecology
Indiana University –Indianapolis
Indianapolis, Indiana
squinney@iupui.edu

Lang Li, Ph.D.
College of Medicine, Department of Biomedical Informatics
The Ohio State University
Columbus, Ohio
Lang.Li@osumc.edu

Abstract—**Phenotyping definitions are widely used in observational studies that utilize population data from Electronic Health Records (EHRs). Biomedical text mining supports biomedical knowledge discovery. Therefore, we believe that mining phenotyping definitions from the literature can support EHR-based clinical research. However, information about these definitions presented in the literature is inconsistent, diverse, and unknown, especially for text mining usage. Therefore, we aim to analyze patterns of "phenotyping definitions" as a first step toward developing a text mining application to improve phenotype definition. A set random of observational studies was used for this analysis. Term frequency-inverse document frequency (TF-IDF) and Term Frequency (TF) were used to rank the terms in the 3958 sentences. Finally, we present preliminary results analyzing "phenotyping definitions" patterns.**

Keywords— Text Mining, Biomedical literature, Phenotyping, Electronic Health Records

I. INTRODUCTION

"Phenotyping definition" is the description of the criteria used to define a phenotype in observational studies to advance knowledge of a disease or adverse event in a population [1]. The nature of "phenotyping definitions" across different studies is highly diverse and inconsistent [2, 3]. There is no internationally agreed upon standard to assist in conducting and reporting "phenotyping definitions" in published studies [3]. This is problematic when developing research studies or comparing results across studies. Text mining methods are able to identify various phenotype definitions from the literature. There is evidence that "repeatable patterns within phenotyping definitions exist" [4]. Learning repeatable patterns in "phenotyping definitions" is a strong starting point for mining "phenotyping definitions" in the biomedical literature.

Biomedical literature mining has shown evidence in supporting biomedical knowledge [5]. Biomedical text mining offers several advantages, such as an accelerated knowledge discovery and cost reduction [6]. Literature-based knowledge discovery has shown evidence of successes in the biomedical domain, such as genome and gene expression, drug–target discovery, drug repositioning, and adverse events [7].

In this work, we provide some preliminary results of analyzing "phenotyping definitions" for text mining applications. There have been some studies in analyzing phenotypes in literature [8, 9]. However, to our knowledge, there has not been work on analyzing patterns of literature-based "phenotyping definitions" where how these phenotypes are defined for future text mining applications.

II. METHODS

The first goal of this work is to better understand the patterns of the "phenotyping definitions" in biomedical literature and to propose patterns that represent "phenotyping definitions". Therefore, we selected a random set of articles of observational studies that used electronic health records (EHRs) as a data source. Articles were reviewed to identify potential target sections that present information pertinent to "phenotyping definitions". After identifying the target section, all sentences from that section were extracted and tokenized.

Phenotype KnowledgeBase (PheKB) [10], which is a phenotype knowledgebase collaborative environment, inspired the data modalities for defining a phenotype. Here, we utilized the following modalities as our baseline features: "Standard codes", "Laboratories", "Medications", and "Natural Language Processing". We further added the "Biomedical and/or Procedure" as a feature. Furthermore, for each of these features, a set of handcrafted sub-patterns representing "phenotyping definitions" were identified. These patterns were term-based, phrase-based, and co-occurrence of terms. These sentences were analyzed using Unigrams and N-Grams techniques. In addition, we used term frequency-inverse document frequency (TF-IDF) and Term Frequency (TF) to rank the terms in the 3958 sentences. Finally, we provide the percentages of the pattern occurrences in our dataset of 3958 sentences.

III. RESULTS AND DISCUSSION

Our analysis revealed that information about phenotype definitions are mostly in "Method" sections of abstracts and full-text. In addition, we found that the major phenotype of the study are usually found in the title of the study. TABLE I

This is the author's manuscript of the article published in final edited form as:

Binkheder, S., Wu, H., Quinney, S., & Li, L. (2018). Analyzing Patterns of Literature-Based Phenotyping Definitions for Text Mining Applications. In 2018 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 374–376). https://doi.org/10.1109/ICHI.2018.00061

shows the percentages of these patterns in the analyzed sentences.

TABLE I. PERCENTAGES OF PHENOTYPING DEFINITIONS PATTERNS

Pattern	Number of sentences	%
Biomedical and procedure	1432 out of 3958	36.2%
Standard code	383 out of 3958	9.7%
Medication use	591 out of 3958	14.9%
Laboratories and quantitative values	247 out of 3958	6.2%
The use of Natural Language Processing (NLP)	49 out of 3958	1.2%

TABLE II provides the handcrafted criteria for each feature or pattern. In addition, we illustrate these criteria on some example sentences from our dataset.

TABLE II. PATTERNS' CRITERIA AND SENTENCES' EXAMPLES

Pattern	Pattern criteria	Examples
Biomedical or procedure	• [Biomedical/ Procedure term] AND [Definition term]	"[identification] of [syndromic conditions]" (PMID 17567225)
Standard code	• Terms indicating the use of standard codes, e.g. billing codes	"a primary or any secondary discharge diagnosis [ICD-9-CM code] of myoglobinuria (791.3)" (PMID15572716)
Medication use	• Terms describing medication use • [Drug term] AND [Definition term]	"Prior [antihypertensive therapy] was [defined as] [the use] of any [AHDs]" (PMID15323063)
Laboratories and quantitative values	• [Biomedical/Procedure term] AND [Measurable value]	"[triglyceride] [level less than 150 mg/dL], [HDL-C] [greater than 40 mg/dL]," (PMID16765240)
The use of Natural language Processing (NLP)	• [Biomedical/Procedure/Drug term] AND [NLP terms]	"[Rule-based] and machine learning techniques were applied to clinical narratives and [smoking status]" (PMID20819866)

Some of the top terms for each patterns ranked by TF-IDF scores for unigram, and TF scores for N-Grams, as the following:

1) *Biomedical or Procedure pattern terms (Top Unigrams and N-grams):* patients, with, diagnosis, patient, included, disease, 'patients with', 'defined as', 'at least', 'diagnosis of', 'was defined', 'to identify', 'based on', 'the following', and 'history of'.

2) *Standard Codes pattern terms (Top Unigrams and N-grams):* icd-9, code, diagnosis, codes, icd-9-cm, classification of diseases', 'icd-9 code', and 'at least'.

3) *Medications Use pattern terms (Top Unigrams and N-grams):* patient, medication, prescription, included, drug, prescribed, 'at least', 'defined as', 'at least one', 'category x', and 'date of'.

4) *Laboratories and quantitative values pattern terms (Top Unigrams and N-grams):* greater, equal, criteria, less, 'greater than', 'equal to', 'defined as', 'mm hg', 'at least', and 'mg/dl'.

5) *The use of Natural language Processing (NLP) pattern terms (Top Unigrams and N-grams:* nlp, language, natural, processing, algorithm, natural language processing', rule-based, 'nlp system', and 'clinical notes'.

We note that the results presented here are preliminary results of creating features for text mining application. The focus is not only on phenotypic data, but also on patterns that surround phenotypes in the literature.

IV. CONCLUSION

This study provides preliminary results of work in progress on analyzing patterns of literature-based "phenotyping definitions". We proposed five major patterns accompanied by examples. In addition, we illustrated some example terms ranked using text mining techniques (Unigrams, N-grams, TF-IDF, and TF) that characterizes each pattern. Finally, we believe that these results can assist in development of future text mining applications for "phenotyping definitions".

REFERENCES

[1] B. S. Glicksberg, R. Miotto, K. W. Johnson, K. Shameer, L. Li, R. Chen et al., "Automated disease cohort selection using word embeddings from Electronic Health Records," Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, vol. 23, pp. 145-156, 2018.

[2] Y. Christley, T. Duffy, and C. R. Martin, "A review of the definitional criteria for chronic fatigue syndrome," Journal of evaluation in clinical practice, vol. 18, no. 1, pp. 25-31, 2012.

[3] B. Rubbo, N. K. Fitzpatrick, S. Denaxas, M. Daskalopoulou, N. Yu, R. S. Patel et al., "Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations," International journal of cardiology, vol. 187, pp. 705-11, 2015.

[4] L. V. Rasmussen, W. K. Thompson, J. A. Pacheco, A. N. Kho, D. S. Carrell, J. Pathak et al., "Design patterns for the development of electronic health record-driven phenotype extraction algorithms," Journal of Biomedical Informatics, vol. 51, pp. 280-6, Oct, 2014.

[5] A. Kolchinsky, A. Lourenco, H. Y. Wu, L. Li, and L. M. Rocha, "Extraction of pharmacokinetic evidence of drug-drug interactions from the literature," PLoS ONE [Electronic Resource], vol. 10, no. 5, pp. e0122199, 2015.

[6] M. S. Simpson, and D. Demner-Fushman, "Biomedical Text Mining: A Survey of Recent Progress," Mining Text Data, C. C. Aggarwal and C. Zhai, eds., pp. 465-517, Boston, MA: Springer US, 2012.

[7] W. W. Fleuren, and W. Alkema, "Application of text mining in the biomedical domain," Methods, vol. 74, pp. 97-106, Mar, 2015.

[8] N. Collier, T. Groza, D. Smedley, P. N. Robinson, A. Oellrich, and D. Rebholz-Schuhmann, "PhenoMiner: from text to a database of phenotypes associated with OMIM diseases," Database (Oxford), vol. 2015, 2015.

[9] J. Henderson, R. Bridges, J. C. Ho, B. C. Wallace, and J. Ghosh, “PheKnow-Cloud: A Tool for Evaluating High-Throughput Phenotype Candidates using Online Medical Literature,” AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science, vol. 2017, pp. 149-157, 2017.

[10] J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford, O. Gottesman, P. L. Peissig et al., “PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability,” Journal of the American Medical Informatics Association, vol. 23, no. 6, pp. 1046-1052, 2016.