Published online: 2020-08-21

Ontologies, Knowledge Representation, and Machine Learning for Translational Research: Recent Contributions

Peter N. Robinson[1,2], **Melissa A. Haendel**[3,4]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[2] Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA
[3] Oregon Clinical & Translational Research Institute, Oregon Health & Science University, Portland, OR, USA
[4] Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR, USA

Summary

Objectives: To select, present, and summarize the most relevant papers published in 2018 and 2019 in the field of Ontologies and Knowledge Representation, with a particular focus on the intersection between Ontologies and Machine Learning.

Methods: A comprehensive review of the medical informatics literature was performed to select the most interesting papers published in 2018 and 2019 and that document the utility of ontologies for computational analysis, including machine learning.

Results: Fifteen articles were selected for inclusion in this survey paper. The chosen articles belong to three major themes: (i) the identification of phenotypic abnormalities in electronic health record (EHR) data using the Human Phenotype Ontology ; (ii) word and node embedding algorithms to supplement natural language processing (NLP) of EHRs and other medical texts; and (iii) hybrid ontology and NLP-based approaches to extracting structured and unstructured components of EHRs.

Conclusion: Unprecedented amounts of clinically relevant data are now available for clinical and research use. Machine learning is increasingly being applied to these data sources for predictive analytics, precision medicine, and differential diagnosis. Ontologies have become an essential component of software pipelines designed to extract, code, and analyze clinical information by machine learning algorithms. The intersection of machine learning and semantics is proving to be an innovative space in clinical research.

Keywords

Artificial intelligence; machine learning; ontology; natural language processing

Yearb Med Inform 2020:159-62
http://dx.doi.org/10.1055/s-0040-1701991

1 Introduction

Unprecedented amounts of clinically relevant data are now available for clinical and research use, including Electronic Health Records (EHRs), laboratory reports, imaging, clinical instrument outputs, drugs and drug doses, genomic investigations, and dynamic data from wearable devices [1]. Machine Learning (ML) is increasingly being applied to these data sources for predictive analytics, precision medicine, and differential diagnosis. Ontologies can be used to encode clinical data for ML and other forms of computational analysis [1]. This survey paper will cover selected important advances at the intersection of ML and clinical ontologies in the last several years.

The terms Artificial Intelligence (AI) and ML are occasionally used interchangeably and have been defined in many different ways. AI is defined by the Food and Drug Administration (FDA) as "the science and engineering of making intelligent machines" [2], but it often invokes fears of machines taking over the world as in Arnold Schwarzenegger's Terminator movies, with Elon Musk, CEO of Tesla, recently calling AI the biggest existential threat to humanity [3]. In contrast, ML refers to a category of AI algorithms that learn from data to perform tasks such as classification and clustering. In practice, ML (still) usually requires humans to define the categories of interest, to provide and at least partially label large amounts of relevant data, and to specify a particular ML algorithm that is suited to the ML task at hand [3].

Many ML algorithms expect data to be encoded in numerical form, and a plethora of methods have emerged for encoding data prior to ML. For instance, one common and simple scheme is "one-hot encoding" where a table would be used to represent a cohort of patients and one column is created for each patient feature. If feature *j* is present in a patient *i*, 1 is entered in cell (i,j) of the table, otherwise 0 is entered. For a cohort of breast cancer patients, one might see one column with cells describing age bins (20-29 years, 30-39 years, ...etc.), tumor size (0-4mm, 5-9 mm, ..., etc.), menopause status, presence of lymph node metastases, history of irradiation, ..., etc. While this type of encoding is easy to perform and enables most ML algorithms to use the data, it ignores the semantic structure of the data, *e.g.*, the relationships between or within features (for instance, that the age groups are ordered or that certain biomarkers are associated with specific subtypes of breast cancer).

Ontologies represent a way of computationally encoding data to reflect our knowledge of a domain. It requires substantially more effort to encode data with ontologies than with simple schemes such as one-hot encoding, but in some cases this may improve the results of ML classification. In contrast to terminologies, which can be defined as systematic nomenclatures that provide a set of preferred or official terms in a domain (*e.g.*, Medical Subject Headings used for information retrieval in PubMed), ontologies define relationships between concepts

in a way that allows computational logical reasoning to infer new knowledge from related assertions. For example, if an ontology defines "bacterium" as an infectious agent, and "infectious meningitis" as a type of meningitis due to an infectious agent, then it would conclude that "bacterial meningitis" is a subclass of "infectious meningitis" [1].

In this survey paper, we summarize a selection of recent outstanding publications in the use of ontologies for translational research, with a focus on the use of ontologies to advance ML technologies.

2 About the Paper Selection

A comprehensive literature review was conducted to identify articles with a focus on ML, ontologies, and knowledge representation for translational research and medical informatics. The selection was performed by querying PubMed/Medline (from NCBI, National Center for Biotechnology Information) with the keywords: "Ontology", "Machine Learning", and "Artificial Intelligence". Databases were searched on December 14th, 2019 for papers published since January 1st, 2018. Results were manually filtered for relevance to translational research and medical informatics. In some cases, additional works are cited to provide context. A total of 15 articles were finally selected for inclusion [4–18].

3 Identifying Phenotypic Abnormalities in EHR Data

Textual data within Electronic Health Records (EHRs) is often a rich narrative describing a variety of patient features, including phenotypic features that are simply not encoded within the EHR's structured data. However, these unstructured data are extremely difficult to use in their raw form for ML. Natural language processing (NLP), specifically Named Entity Recognition (NER), transforms the data into structured information annotated to terminological entities for analytics.

The Human Phenotype Ontology (HPO) is a widely used resource for capturing human disease phenotypes for computational analysis to support differential diagnoses [19]. A number of publications have appeared in the last two years that apply NLP supplemented by ontologies and ML to extract Human Phenotype Ontology (HPO) terms from medical texts. Integration of detailed phenotype information with genetic data can facilitate accurate diagnosis of Mendelian diseases by exome or genome sequencing [20]. Computational tools for HPO-driven prioritization of variants identified by exome or genome sequencing typically require manual entry of the proband's clinical signs and symptoms and other phenotypic abnormalities [21]. While this is common practice in research settings, for clinical care, it would be desirable to extract a comprehensive and accurate set of HPO terms directly from EHR data. The heterogeneity and noise in EHR narratives make this challenging. As a first step towards this goal, Son and colleagues [5] developed an NLP pipeline to process genetic notes from EHRs and extract and normalize phenotype concepts by using HPO using MedLEE [22] and MetaMap [23], two well-regarded NLP tools. Both tools extracted more HPO terms than human experts (on average 18-19 for NLP tools, and 11 for human experts). In a retrospective study, the authors show that prioritization results using the NLP-extracted HPO terms had a performance comparable to that with expert curation of phenotype terms from EHR narratives [5]. This is critically important because it demonstrates that there are significant efficiency gains that can potentially increase the use of deep phenotyping in settings without substantial curation resources.

More recent efforts have demonstrated a high accuracy in identifying HPO terms in EHRs. ClinPhen [6] breaks EHR-derived free text into sentences and words, and uses heuristics to identify HPO terms in commonly encountered clinical phraseology. For instance, the phrase "Hands are large" will match the HPO term "*Large hands* (HP:0001176)", and excluded phenotypic abnormalities or those that were found in other family members are recognized as such. The authors showed that ClinPhen had a similar performance when compared to two commonly used, general purpose NLP tools. Importantly, HPO terms derived by ClinPhen displayed better performance in gene prioritization tasks.

Bastarache and colleagues [7] present a phenotype risk score for rare disease based on a mapping to HPO ontology terms to consolidate billing codes from the EHR. For instance, the HPO term "*Hydroureter* (HP:0000072)" was mapped to the consolidated billing code for "*Stricture/obstruction of ureter"*. The mappings, termed "phecodes", are used to define a Phenotype Risk Score (PheRS) for Mendelian diseases as the sum of clinical features observed in a given subject weighted by the log inverse prevalence of the feature. PheRS was shown to be effective in identifying patients with diagnosed Mendelian disease using only phecodes derived from the EHR. In a study on 21,701 genotyped adults, an increased burden of phenotypes among individuals with rare variants in Mendelian disease genes was found, and subsets of patients with likely genetic causes for common diseases were identified [7]. This work was innovative in leveraging integration of billing data to support identification of individuals with rare genetic disease from EHR data.

Ontology-based encoding of clinical data can, to some extent, mitigate the fact that many NLP tools were designed for English-language texts. A recent study used the Chinese translation of the HPO to extract ontology terms from Chinese EHR data in order to develop a venous thromboembolism risk assessment model [8]. This work illustrates how the use of a standardized ontology may even be used across clinical systems implemented in different languages in a standardized manner.

NLP extraction from EHRs is now being used to support clinical care. In a study performed in a neonatal intensive care unit (ICU), NLP was applied to 101 children with genetic diseases [9]. A mean of 4.3 NLP-extracted phenotypic features matched the expected phenotypic features of those diseases, compared with a match of 0.9 phenotypic features used in manual interpretation. The accuracy and speed of NLP extraction of HPO terms is essential in the setting of a neonatal ICU, where speed is of essence in order to avoid serious and potentially irreversible complications [9]. A number of approaches have been attempted to further improve accuracy. For example, Doc2HPO is a Web-based tool that allows users to vet and improve the results of HPO terms that can be extracted from clinical notes using one of five NLP engines [10].

Taken together, these results document that NLP of phenotypic data is becoming a mature field that can be used to improve clinical care.

4 Ontologies and Machine Learning for Medical NLP

Deep learning (DL) methods are extremely powerful in a wide range of applications. However, deep learning has been most successful on data with an underlying Euclidean structure, in which data points can be represented as numeric vectors [24]. In many settings, clinical knowledge can be represented as a knowledge graph (KG) that represents data as an heterogeneous graph with nodes and edges of many different types. However, additional algorithmic techniques are required to apply deep learning to these graphs. In essence, these methods transform the KG into numeric vectors, a process referred to as graph embedding. Many of these algorithms extend the word2vec family of algorithms, which produce vector embeddings of words by training two-layer neural networks on the contexts of words in some corpus of texts. Word2vec predicts the probability of surrounding words with a radius of *m* words given a word in the center. Vectors with 50-200 dimensions are randomly initialized to start the algorithm, and a loss (error) function is defined that compares the predictions of the vectors representing the words with words actually observed in texts. Then, gradient descent learning is performed using standard deep learning approaches, which adjusts the values of the vectors representing each word to reduce the error rate [25]. Thus, the vector embeddings are produced as a by-product of backpropagation in deep learning.

A number of methods have emerged for graph embedding, whereby the nodes of the graph represent "words" and walks across the network of the graph generate "texts" that can be passed to the word2vec algorithm, which then performs the actual embedding. A random walk on a graph begins at a start node and selects a node at random, and moves to it. Then, a neighbor of this node is selected at random and the walk moves to it. In this way, a random sequence of nodes (a path through the graph) is selected [26]. Deep learning approaches based on random walks sample from many random walks from a node in order to generate a graph embedding that preserves graph properties. The DeepWalk algorithm treats series of nodes on a path analogously to a series of words in the word2vec algorithm and tries to predict the probability of context nodes given a node of interest [27]. DeepWalk and extensions of DeepWalk such as node-2vec provide class or multiclass predictions for nodes, in which every node in the graph is assigned one or more labels representing a finite set of categories [28,29]. Graph embedding techniques have been used for a number of prediction tasks in the biomedical domain including the prediction of polypharmacy side effects [30].

Vector embedding approaches are being adopted in the translational research community. For instance, HPO2Vec+ embeds the HPO with disease phenotype associations and was used to stratify rare disease patients in EHR data at the Mayo Clinic [11]. Another approach to ontology-guided graph embedding uses a convolutional neural network to encode input phrases and then rank medical concepts based on the similarity in that space. It uses the hierarchical structure provided by the HPO and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) terms as an implicit prior embedding to better learn embedding of various terms [4]. Other new research shows the importance of wisely choosing text corpora for training. Vector embeddings trained on different sources can show dramatically different accuracies on medically relevant prediction tasks [12, 18]. Recently, an embedding of the Unified Medical Language System (UMLS) concept identifiers (CUIs) was generated from an insurance claims database of 60 million members, a collection of 20 million clinical notes, and 1.7 million full text biomedical journal articles, resulting in the largest ever set of embeddings for 108,477 medical concepts. The resulting approach, terms cui2vec, was shown to be superior to word2vec and some related algorithms on tasks including prediction of comorbidity and UMLS Semantic Type [13]. All of these methods demonstrate the intersection of the use of ontologies for NER with deep learning to further maximize extraction of knowledge from clinical data.

5 Ontology-based Extraction of Structured Data from EHRs

Limited EHR interoperability between institutions makes secondary research use of EHR data challenging, especially in collaborative projects involving more than a single institution. The HL7 Fast Healthcare Interoperability Resources (FHIR) is one of several models that intend to provide a standardized data representation for EHR data. Hong and coworkers [14] presented a pilot study on a generic integration approach for modeling EHR data with the FHIR data model. They developed a rule-based approach to assign NLP output types as transformations of structured EHR data where appropriate. They showed that their system achieved acceptable accuracy in extracting information about medication statements in EHR data [14]. The same group then used an extension of that framework, called NLP2FHIR, to convert discharge summaries into corresponding FHIR resources, which were then passed to ML algorithms to predict obesity and comorbidities, achieving effective prediction accuracy [15]. SemEHR is a similar approach to extracting structured and unstructured components of EHRs [16]. Finally, our own approach to converting laboratory encoded data into HPO transforms the outcomes of commonly used laboratory tests (encoded using the Logical Observation Identifiers Names & Codes (LOINC)) with HPO terms, thereby providing a means of interpreting the outcomes of laboratory tests using an ontology of phenotypic abnormalities. In a study on 15,681 patients with respiratory complaints, our approach allowed us to convert the majority of the laboratory tests into HPO terms and assign an average of 57.7 unique phenotypes to each patient. A number of previously described asthma biomarkers were found to have statistically significant overrepresentation in individuals diagnosed with this disease [17]. Approaches like these are likely to be influential as means of improving portability and interoperability of ML-based phenotyping algorithms across different institutions and EHR systems.

6 Conclusion

The last several years have witnessed a growth in the importance of ML in many areas of medical informatics. In this survey, we have presented a selection of recent articles that document the utility of ontologies in extracting and coding clinical information for ML and other computational analysis approaches. In the initial phases of the HPO project (2008-2015), HPO terms were captured by bespoke tools or entered by hand prior to use in tools for differential diagnostic support or research [31]. The articles summarized in this piece describe a diverse set of tools to extract HPO from EHR data using different approaches to perform research or accelerate clinical diagnostics in ways that would have been unimaginable a decade ago. One of the hottest new topics in the intersection between ontologies and ML in the last two years has been the application of node embedding algorithms to clinical data. Finally, several works have emphasized the benefit of using ontologies as part of pipelines to extract clinical profiles from EHR for phenotyping, research, or decision support. As such technologies evolve, there will likely be increasing use of different ontologies in different ways to perform EHR-based analytics — supporting not only improved but also more standardized clinical decision support and clinical research.

Acknowledgements

The authors were supported by a grant from the National Institutes of Health's National Center for Advancing Translational Sciences, Grant Number U24TR00230.

References

1. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. N Engl J Med 2018;379:1452–62.
2. Toh TS, Dondelinger F, Wang D. Looking beyond the hype: Applied AI and machine learning in translational medicine. EBioMedicine 2019;47 607–15.
3. National Academies of Sciences, Engineering, and Medicine. Artificial Intelligence and Machine Learning to Accelerate Translational Research: Proceedings of a Workshop. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513721/pdf/Bookshelf_NBK513721.pdf
4. Arbabi A, Adams DR, Fidler S, Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. JMIR Med Inform 2019;7:e12596.
5. Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, et al. Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. Am J Hum Genet 2018;103:58–73.
6. Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet Med 2019;21:1585–93.
7. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 2018;359:1233–9.
8. Yang Y, Wang X, Huang Y, Chen N, Shi J, Chen T. Ontology-based venous thromboembolism risk assessment model developing from medical records. BMC Med Inform Decis Mak 2019;19:151.
9. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med 2019;11(489):eaat6177. Available from: http://dx.doi.org/10.1126/scitranslmed.aat6177
10. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res 2019;47:W566–W570.
11. Shen F, Peng S, Fan Y, Wen A, Liu S, Wang Y, et al. HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. J Biomed Inform 2019;96:103246.
12. Lin C, Lou Y-S, Tsai D-J, Lee C-C, Hsu C-J, Wu D-C, et al. Projection Word Embedding Model With Hybrid Sampling Training for Classifying ICD-10-CM Codes: Longitudinal Observational Study. JMIR Med Inform 2019;7:e14499.
13. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer NP, et al. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. Pac Symp Biocomput 2020;25:295-306. Available from: http://arxiv.org/abs/1804.01486.
14. Hong N, Wen A, Shen F, Sohn S, Liu S, Liu H, et al. Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. AMIA Jt Summits Transl Sci Proc 2018;2017:74–83.
15. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen L, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform 2019;99:103310.
16. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018;25:530–7.
17. Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, Carmody LC, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med 2019;2:32. Available from: http://dx.doi.org/10.1038/s41746-019-0110-4.
18. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018;87:12–20.
19. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2019;47(D1):D1018-D1027. Available from: http://dx.doi.org/10.1093/nar/gky1105.
20. Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, Weiss ME, Köster J, Marais A, et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. Eur J Hum Genet 2017;25:176–82.
21. Köhler S, Øien NC, Buske OJ, Groza T, Jacobsen JOB, McNamara C, et al. Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics. Curr Protoc Hum Genet 2019;103:e92.
22. Sevenster M, van Ommering R, Qian Y. Automatically correlating clinical findings and body locations in radiology reports using MedLEE. J Digit Imaging 2012;25:240–9.
23. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229-36.
24. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: going beyond Euclidean data. IEEE Signal Processing Magazine 2017;34(4):18-42. Available from: http://arxiv.org/abs/1611.08097
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. Available from: http://arxiv.org/abs/1301.3781
26. Lovász L. Random walks on graphs: A survey. Combinatorics, Paul erdos is eighty 1993;2:1–46.
27. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM 2014:701–10.
28. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. KDD 2016;2016:855–64.
29. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. IEEE Trans Knowl Data Eng 2018;30(9):1616-37. Available from: http://arxiv.org/abs/1709.07604
30. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 2018;34:i457–i466.
31. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet 2009;85:457–64.

Correspondence to:
Peter Robinson
The Jackson Laboratory for Genomic Medicine
10 Discovery Drive
Farmington CT 06032, USA
E-mail: peter.robinson@jax.org