Classification of genetic developmental disorders in the peer-reviewed literature using biomedical ontologies - Abstract

T. Michael Yates[1,*], T. Ian Simpson[1,2]

[1]*Institute for Adaptive and Neural Computation, Informatics Forum, The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK*

[2]*Simons Initiative for the Developing Brain, The University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XF, UK*

Abstract

Diagnosis in genetic developmental disorders (GDD) requires analysis of phenotypic information from thousands of conditions described in the peer-reviewed literature. Automated search and classification of these manuscripts is needed to enable phenotype data extraction at scale. Here, we describe a classifier developed for this purpose, using features derived from several biomedical ontologies. PubMed search strategies were evaluated to identify filters which enriched for GDD-relevant manuscripts. This was used to inform the creation of a corpus of 13,000 papers describing 230 exemplar GDD. The title and abstract were manually annotated to papers containing phenotypic data for GDD or not. A random forest binary classification algorithm was trained on this dataset, using features derived from manuscript metadata, MeSH (Medical Subject Headings) ontology, Human Phenotype Ontology and Mondo Disease Ontology. The PubMed gene symbol[title] filter was selected as the best balance between minimizing large numbers of results (>100,000 for some genes) and enriching for GDD-relevant manuscripts. 36 features were used, with the top 5 by feature importance being: disease name in abstract, case report in publication type, disease name in title, published in GDD journal, and 'mutation' in MeSH terms. The precision was 0.81 and recall 0.83. We demonstrate effective classification of the GDD peer-reviewed literature using supervised machine learning. Application of this system, in combination with named entity recognition, will allow for creation of phenotypic data models across the spectrum of GDD.

Keywords

Ontology, genetic developmental disorders, machine learning, phenotype analysis

15th International Conference on Biomedical Ontologies 2024, July 17-19, 2024, Enschede, The Netherlands

*Corresponding author.

tyates@ed.ac.uk (T. M. Yates); ian.simpson@ed.ac.uk (T. I. Simpson)

0000-0002-0118-5946 (T. M. Yates); 0000-0003-0495-7187 (T. I. Simpson)