

ADO: A formal ontology for understanding structure of information and domain knowledge of Alzheimer's disease

Ashutosh Malhotra, Michaela Gündel and Martin Hofmann-Apitius

Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin
53754 Germany

Formal biomedical ontologies offer the capability to retrieve domain-specific knowledge. Particularly for grievous diseases like Alzheimer's where latency of onset to clinical symptoms takes many years, automated structured mining approaches can generate great value. No single system is currently capable of covering a complete domain of Alzheimer's disease (AD) by itself and manual effort will be insufficient to retrieve all relevant information. This makes it necessary to develop focused automated applications that can deal with individual aspects of the disease in a reliable manner. A formal Alzheimer's disease ontology (ADO) with potential to engineer knowledge can provide a good way to describe and organize the AD-specific knowledge sphere. We have developed an ADO representing knowledge about impact, treatment, risk factors, diagnostic, effects and phases of AD. This is the first example of an ontology covering different aspects of AD.

ADO is constructed in accordance to the ontology building life cycle and using the Basic Formal Ontology (BFO) as its formal upper ontology. Protégé OWL editor was used as a tool for building ADO in Ontology Web Language (OWL) format. Collection of terms related to AD was performed by scanning various knowledge sources which includes review articles, content of online books, standard knowledge bases, encyclopedias, glossaries, and informative online sources and websites. ADO was developed around four main concepts, namely clinical things, etiological things, molecular and cellular mechanisms and Non clinical things; these concepts and their hierarchies represent what we call "contextual views" on the ontology. Concepts under each of these views are inferred using reasoning and are a collection of classes that can be either continuants or occurrents; thus, BFO upper class of these views is "entity". This solution that aims at creating a contextual collection of classes can aid in domain-specific retrieval of documents, producing desired search results. Further, ADO was enriched by adding synonyms, useful annotations and references.

ADO has been manually curated by an expert clinician, who added certain more clinically relevant concepts to ADO increasing its pragmatic usability. Also, the structural features of ontologies reflecting topological and logical properties were measured by means of context-free metrics

Including depth, breadth, tangledness, and fan-outness. Using our state-of-the-art text mining tool, ProMiner, which takes the ADO dictionary as input, functional evaluation of the ontology was done over a test set of 200 abstracts. Results of the evaluation produced an *F-score* of 0.72 showing that our designed ontology accurately captures relative good range of AD concepts in scientific text.

The ADO dictionary was also used to mine 650 AD Electronic patient health record's (EHR) for comorbidity analysis, systematically looking for other diseases or disorders that may exist simultaneously but independently in patients suffering from AD. Term frequency count calculation shows that hypertension, diabetes and stroke are the top three disorders which occur frequently in patients already suffering from AD. When integrated into our information retrieval system SCAIView, ADO leveraged the efficiency of semantic information retrieval and knowledge representation by providing the possibility to perform cross sectional searches with other dictionaries. Particularly its intersection with 'HypothesisFinder' facilitated the retrieval and aggregation of domain-specific speculations (hypotheses). Some of the hypotheses proposed for a particular stage of AD retrieved by using this combination methodology are:

We suggest that M-CSF, thus generated, contributes to the pathogenesis of AD, and that M-CSF in cerebrospinal fluid might provide a means for monitoring neuronal perturbation at an early stage in AD. (PMID: 9144231)

Our results show that modification of tubulin function may contribute to intermediate or late stages in the pathogenesis of sporadic and inherited AD as well as FTDP-17. (PMID:16155344)

In a benchmark experiment, the combination methodology using ADO and HypothesisFinder for extracting AD stage specific speculative sentences was compared to the human curated Alzswan knowledge base. Our approach displays a much higher stage specific coverage of speculative statements about the etiology of AD than the human (expert) curated database AlzSWAN.