Expanding the Autism Ontology to DSM-IV Criteria

Omri Mugzach\textsuperscript{1}, BA; Mor Peleg\textsuperscript{1}, PhD; Steven C. Bagley\textsuperscript{2}, MD, MS; Russ B. Altman\textsuperscript{2}, MD, PhD

\textsuperscript{1}Department of Information Systems, University of Haifa, Israel, 31905
\textsuperscript{2}Department of Genetics, Stanford University, Stanford, CA, 94305

Abstract: In order to understand the environmental and genetic factors contributing to autism, we are extending an existing Autism Ontology with DSM-IV vocabulary definitions, risk factors, phenotypic manifestations and their prevalence. We use logical rules along with class restrictions for deducing phenotypes from patients’ ADI-R data and the DSM-IV criteria.

Background: The mechanism of autism is unknown, and it is critical to organize patient data concerning genetic and environmental risk factors as well as phenotypic manifestations. Ontologies help in such data integration task as a way to standardize data and knowledge about the disease and to create a knowledge infrastructure for studying how genetic and environmental factors impact the disease development. There already exists an ontology for autism, represented in the OWL formalism [1]. It contains knowledge regarding assessment tools and phenotypes relevant to autism and a set of rules which allows deduction of specific phenotypes based on autism assessment tools’ results. In order to support the extraction of information from electronic health records, our goal is to enrich this ontology with knowledge regarding the diagnosis of autism, its risk factors, and phenotypic manifestations.

Methods: We extend the current Autism ontology with DSM-IV vocabulary definitions, risk factors, phenotypic manifestations and their prevalence, using controlled vocabulary terms and synonyms for these concepts. The DSM-IV definitions for autism are hierarchical. The upper level includes 3 criteria. The first criterion is the most complex one, and includes 3 sub-criteria, each containing 4 different patient phenotypes. We represent relevant concepts and definitions using a hierarchical relations structure provided by the Web Ontology Language (OWL) as implemented in the Protégé tool. All concepts (risk factors and manifestations) and their definitions are represented as a condition of a certain human (e.g., human with delayed spoken words, human meeting DSM definition A2, etc.) rather than a standalone class (or concept). We define OWL classes corresponding to DSM-IV sub-criteria by combining patient phenotypes with logical operators. The upper level DSM-IV criteria necessitate counting the number of sub-criteria from specific categories that hold, which requires support of k-of-N counting. Since OWL reasoners cannot typically perform k-of-N counting, we are creating a plug-in to perform the desired operation. To deduce a diagnosis for a patient we take the following steps: New patient instances are created and populated manually with data from the Simons Foundation Autism Research Initiative (SFARI). This data set includes for each patient results from structured interviews used for diagnosing autism (Autism Diagnostic Interview-Revised (ADI-R) [2]). We define SWRL (Semantic Web Rule Language) rules based on ADI-R to abstract data from the populated patient instances and deduce which phenotypes each patient displays. In this way, the Pellet OWL reasoner can deduce all the DSM-IV sub-criteria that a specific patient instance meets. After developing the plugin we will be able to infer whether patients meet DSM criteria.

Results: 45 SWRL rules deducing different phenotypes from the SFARI data were implemented for 5 ADI-R items (e.g., 5 different rules were implemented for age of first spoken word: delayed word, milestone not reached, no word, word not delayed, question not asked). Class restrictions were implemented for 2 DSM-IV criteria concerning spoken language and social conversation. All restrictions and SWRL rules were tested with actual SFARI data of 7 patients.

Conclusion: We can use OWL definitions and reasoning to infer patients meeting DSM-IV criteria based on ADI-R assessments. Next steps include: (i) developing the new k-of-N counting plug-in for Protégé; (ii) automating instance population with SFARI data; (iii) adding prevalence and frequency information for phenotypes; (iv) adding vocabulary codes and synonyms.

Acknowledgement: This work was partly funded by the Conte Center for Computational Neuropsychiatric Genomics (NIH P50MH94267)

References: