Introduction: Understanding disease processes of complex neurodevelopmental diseases such as autism has been attempted for many years with relatively modest success. The mechanism of autism is still unknown, and in order to allow scientists to try and deduce new knowledge about this disease’s mechanism based on patient data, it is crucial to organize and semantically integrate patient data concerning genetic and environmental risk factors as well as phenotypic manifestations. Ontologies are important knowledge representations used in information systems and artificial intelligence research, which help in such data integration tasks as a way to standardize data and knowledge related to 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, the Autism Ontology [1], developed at Stanford University and represented in the Web Ontology Language (OWL) [3]. This ontology contains knowledge regarding autism assessment tools and phenotypes related to autism along with a set of rules which allows deduction of specific basic phenotypes based on patient data containing autism assessment tools’ results. However, this ontology does not allow deduction of more complex phenotypes as defined in the state-of-the-art Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and therefore cannot be used for diagnosis of autism based on patient data.

---

**Figure 1: Autism DSM-IV Criteria**

**299.00 Autistic Disorder**

A. A total of six (or more) items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3):

1. Qualitative impairment in social interaction, as manifested by at least two of the following:
   a. Marked impairment in the use of multiple nonverbal behaviors, such as eye-to-eye gaze, facial expression, body posture, and gestures to regulate social interaction
   b. Failure to develop peer relationships appropriate to developmental level
   c. A lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest)
   d. Lack of social or emotional reciprocity

2. Qualitative impairments in communication, as manifested by at least one of the following:
   a. Delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gesture or mime)
   b. In individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others
   c. Stereotyped and repetitive use of language or idiosyncratic language
   d. Lack of varied, spontaneous make-believe play or social imaginative play appropriate to developmental level

3. Restricted, repetitive, and stereotyped patterns of behavior, interests, and activities as manifested by at least one of the following:
   a. Encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus
   b. Apparently inflexible adherence to specific, nonfunctional routines or rituals
   c. Stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting or complex whole-body movements)
   d. Persistent preoccupation with parts of objects

B. Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years:

1. Social interaction
2. Language as used in social communication, or
3. Symbolic or imaginative play

C. The disturbance is not better accounted for by Rett’s disorder or childhood disintegrative disorder.
Objectives: Our goal is to enrich the existing autism ontology with additional knowledge regarding the diagnosis of autism and its associated phenotypic manifestations in order to support the semantic extraction of autism-related information from electronic health records and inference of more complex and abstract phenotypes related to autism.

Methods: We received data from the Simons Foundation Autism Research Initiative (SFARI). This data set includes results from structured 93-items interview used for diagnosing autism called the Autism Diagnostic Interview-Revised (ADI-R) [2]. We converted this data set into instances of the ADI-R class (one instance per patient) contained in the existing Stanford autism ontology.

The fourth version of DSM-IV, published by the American Psychiatric Association, contains a listing of psychiatric disorders and their corresponding diagnostic codes along with specific diagnostic criteria for each disorder. We would like to extend the existing autism ontology to represent all the DSM-IV autism diagnostic criteria. As can be seen in Figure 1, the DSM-IV diagnostic criteria for autism are hierarchical. The upper level includes 3 criteria: A, B, and C. The first criterion is the most complex one and includes 3 middle-level sub-criteria (1, 2, and 3), each containing 4 lower-level different patient states (a-d), referring to his/her basic phenotypes. A large portion of the basic autism phenotypes (such as "Delayed Word") are already represented in the existing autism ontology from Stanford.

The existing autism ontology contains classes of basic phenotypes (such as "Delayed Word"), dispositions (such as "Language Skills phenotype") and diagnostic tools (such as ADI-R). While the existing autism ontology includes basic phenotypes such as "Delayed Word" under the quality class hierarchy, we have added a hierarchy of classes referring to humans with findings (i.e., basic phenotypes) treated as a condition rather than a basic standalone phenotype. For example, Human_with_delayed_word is defined in OWL as Human has_finding has Delayed_Word. This allows us to infer which basic phenotypes are exhibited by patients (humans) based on their electronic health records (SFARI data set).

As done in [1], we used the Protégé knowledge modeling tool to define SWRL (Semantic Web Rule Language) rules [4] that operate on the OWL knowledge base. These rules are used to infer basic DSM phenotypes from ADI-R patient data found in the SFARI database that were migrated into the OWL knowledge base. For example, if a certain patient instance had the value 24 for the ADI-R item concerning the age (in months) of first spoken word, we would deduce that this patient has the phenotype of Human who has the finding: Delayed Word. The following SWRL rule describes that deduction. Note that at the beginning of the rule, we include an integrity check that the number specified is less than 900.
Based on the basic phenotypes, we built the DSM-IV diagnostic criteria using OWL class restrictions (definitions). As shown in Figure 2, we have added the concept of Human_with_DSM_IV_definition in order to be able to infer, from electronic medical records, those patients who meet the DSM diagnostic criteria of autism and parts of such diagnostic criterion (such as part 1_a). In order to represent the different sub-criteria of the middle and lower levels of DSM IV, we combine the basic patient phenotypes (e.g., Human_with_delayed_word) with logical operators. For example, to represent the criterion A-2(b) shown in Figure 1, we used the following OWL class definition:

\[
(has\_finding \text{ has Functional}\_use\_of\_spontaneous\_echoed\_or\_stereotyped\_language) \text{ or } (has\_finding \text{ has Fewer}\_than\_five\_words\_total\_or\_speech\_not\_used\_on\_a\_daily\_basis) \text{ and } ((has\_finding \text{ has Social}\_verbalization\_clear\_social\_quality) \text{ or } (has\_finding \text{ has Social}\_verbalization\_some\_social\_use)).
\]

The middle and upper level DSM-IV criteria A and B involve counting the number of sub-criteria from specific categories that hold, which necessitates support of k-of-N counting. OWL allows use of reasoners. A reasoner is a piece of software able to infer logical consequences from a set of asserted facts or axioms. For example, an OWL reasoner Pellet can deduce that a certain patient instance meets a certain class criteria based on its phenotypes. Since OWL does not have the capability to express axioms that involve counting, we will create a new plug-in which will perform the desired operation. Our new Plug-In will use Pellet's API to allow us to access the reasoner's capability to realize individuals (patient instances) who meet the different criteria and count them. After developing the plugin we will be able to infer whether patients meet DSM middle and upper-level criteria.

Results: 45 SWRL rules deducing different basic phenotypes from the SFARI data were implemented for 5 ADI-R items (about 5 SWRL rules for each ADI-R item). For example, regarding the basic phenotype age of the first spoken word, 5 different rules were implemented (delayed word, milestone not reached, no word, word not delayed, question not asked). Regarding the DSM-IV definitions, class restrictions were implemented for 2 lower level DSM-IV criteria concerning spoken language and social conversation. All restrictions and SWRL rules were successfully tested with actual SFARI data of 7 patients.

Conclusion: It seems feasible to use OWL definitions and reasoning to infer whether patients meet DSM-IV criteria and sub-criteria based on ADI-R assessment data. Our next steps include: (i) developing the new k-of-N counting plug-in for Protégé, and (ii) automating instance population with SFARI data.

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

References:

3. Web Ontology Language (OWL), http://www.w3.org/TR/owl-features/, last accessed 3.3.2013