# Creating Consistent Diagnoses List for Developmental Disorders Using UMLS

Nuaman Asbeh[1], Mor Peleg[2], Mitchell Schertz[3], and Tsvi Kuflik[2]

[1]Department of Statistics, University of Haifa, Haifa, Israel, 31905
nuamana@yahoo.com
[2]Department of Management Information Systems, University of Haifa, Haifa, Israel, 31905
{morpeleg, tsvikak}@mis.haifa.ac.il
[3]Institute for Child Development, Kupat Holim Meuhedet, Central Region, Herzeliya, Israel 46782
mitch_s@meuhedet.co.il

**Abstract**. In the field of developmental disorders, there is no commonly accepted medical vocabulary. Vocabularies, such as ICD-10, are unsatisfying to clinicians, who try to create their own diagnostic lists. This results in inconsistency in the terms used in clinical practice. When attempting to apply automatic computational methods on patients' data, the need for common consistent diagnoses list arises. To this end, we mapped a set of different diagnoses used in clinical practice to UMLS —a well-known Unified Medical Language System that organizes and unifies over 100 vocabularies. Diagnoses that were defined by different terms and were mapped to one concept at UMLS were joined as synonyms. Concepts that were not found at UMLS were mapped to the closest concept found. We found that SNOMED-CT is the most comprehensive vocabulary (85.7% term coverage) in UMLS. We propose a framework that, when applied on inconsistent manually constructed set of diagnoses, leads to minimal and consistent set of diagnostic terms.

## 1 Introduction

High level of co-morbidity has long been recognized in childhood developmental disorders; children who have been diagnosed with one developmental disorder are very likely to meet diagnostic criteria for some other developmental disorder. This symptomatic overlap between different developmental disorders has led to a high rate of misdiagnosis. Clinical vocabularies, such as DSM-IV (American Psychiatric Association 1994) and ICD-10 (World Health Organization 1992), classify developmentally and psychiatrically disturbed children in terms of their emotional symptoms and behaviors while ignoring the underlying mechanisms, resulting in multiple overlapping diagnoses that may refer to the same underlying mechanism. Thus, these vocabularies are deeply unsatisfying to many clinicians, who try to create their own diagnostic lists. This results in inconsistency in the diagnostic terms used in clinical practice.

Developmental disorder diagnoses may be grouped into super diagnoses groups according to their underlying mechanisms that cause common findings that are shared by the disorders. Evidence for the existence of such groups has been shown [1]. To this end, we are applying clustering analysis on patient data, to group developmental disorder diagnoses into clusters of super diagnoses. However, for the clustering results to be valid, the set of diagnoses (features) upon which the clustering methods are used, need to be defined clearly and consistently used by all the clinicians who diagnose pediatric patients. Therefore, we decided to develop a developmental-disorder ontology that is satisfying and acceptable to practitioners in the field and that will ensure that all practitioners who give diagnoses to children are using the same agreed-upon terminology and interpret the terms in a consistent way.

An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them [2]. Some of the reasons for developing ontologies are: (1) to share common understanding of the structure of the information among people or software agents, (2) to enable reuse of domain knowledge, (3) to make domain assumptions explicit.

## 2 Methods

We used the protégé-2000 [2] knowledge modeling tool to create the ontology of developmental disorders. We started from a set of diagnoses that are used in clinical practice at the Institute for Child Development, Kupat Holim Meuhedet, Central Region, Herzeliya. We mapped these diagnoses to medical concepts defined in the UMLS (Unified Medical Language System [3]) Metathesaurus—a well-known vocabulary system that organizes and unifies other vocabularies. Diagnoses that were defined by different terms and were mapped to the same concept at UMLS were joined together as synonyms. The clinician expert on our team (MS) mapped concepts that were not found in the UMLS Metathesaurus to the closest concept in UMLS (e.g., the term "sleep disorder – discontinuous" was mapped to the term "Sleep Initiation and Maintenance Disorders"). We considered each concept that was not found in UMLS, but appeared in the diagnosis set of the clinicians, as a candidate for entering the ontology's hierarchy as subclasses of existing classes. We inserted candidates as subclasses if they were specializations of diagnoses, based on severity (e.g., mild developmental delay), certainty of the diagnosis (e.g., ADHD-preschool is a specialization of ADHD since 50% of preschoolers diagnosed with ADHD are no longer diagnosed with it at school age), or causality (e.g., developmental memory disorders as a subtype of memory disorders). We would not add concepts to the ontology if the missing concepts are related to local environments or genetic influences because this would compromise the acceptance of the ontology by wider community of clinicians, worldwide.

We structured each medical concept in the ontology by defining slots taken from the way by which terms are defined in the UMLS Metathesaurus. These slots include name, synonyms, semantic types, textual definitions, concept_identifier, and the source vocabularies in which it is found. We defined terms that were not found in UMLS using their original name and a unique concept_identifier.

In order to construct the ontological links (hierarchical and non-hierarchical), we found the minimal coverage set (MCS)-the minimal set of vocabularies that cover all the concepts found in UMLS, and checked which vocabulary covers most of the concepts in the ontology (SNOMED-CT, as discussed in section 3). The taxonomy of this vocabulary served as a basis for the ontology's hierarchy, which we call the Main Tree (MT). To create the MT, we manually entered terms to the ontology to ensure using only the MCS set; we entered all the ancestors of the terms found in the selected vocabulary. We then added the terms that were in MCS but not at the vocabulary used to create the MT: we followed the ancestral hierarchies of such a term, as it exists in the vocabularies in MCS-MT, searching for an ancestor that exists at MT. We added the ancestors up to the mutual ancestor to MT. In this way, we were able to link few more concepts into MT. The rest of the concepts remained in separate hierarchies, taken from their source vocabularies.

## 3 Results

Mapping the terms to UMLS minimized the set of 179 inconsistent diagnoses to an equivalent set of 88 consistent diagnoses, by inserting 91 diagnoses as synonyms of the other 88. The 88 diagnoses were entered to the ontology. Seventy seven of the 88 terms were found in UMLS (87.5%), and 11 terms were not found in UMLS but are subtypes of terms found in UMLS (12.5%), as explained in the Methods. The diagnosis list was approved by the clinician expert on our team (MS) to ensure clinical validity of the list.

**Table 1**. Percentage of diagnoses covered by clinical vocabularies found in UMLS

| Vocabulary | Number of diagnoses covered | %diagnoses covered |
|---|---|---|
| SNOMED-CT | 66 | 85.7 |
| SNOMED-Intl-1998 | 59 | 76.6 |
| MedRA | 56 | 72.7 |
| ICD-9-CM, ICD-10, &DSM-IV | 51 | 66.2 |
| MESH | 43 | 55.8 |
| Alcohol and Other Drug Thesaurus | 37 | 48.0 |
| ICD-10 | 34 | 44.2 |
| ICD-9-CM | 31 | 40.3 |
| DSM-IV | 28 | 36.4 |
| Clinical Problem Statements | 25 | 32.5 |
| ICPC2E-1998 | 2 | 2.6 |

We found that 18 vocabularies include terms from the diagnostic list and 7 of them are sufficient to create a MCS. The results showed that the SNOMED-CT vocabulary is the most comprehensive vocabulary, covering all but 11 concepts found in UMLS (85.7%

coverage). We therefore used its taxonomy as the ontology's concept hierarchy and were able to integrate seven of the 11 concepts with SNOMED-CT's hierarchy since ancestors of these concepts were included in SNOMED-CT. The remaining four concepts were integrated as separate hierarchies originating from their source vocabularies. Table 1 shows the percentage of diagnoses that are covered by the MCS and by DSM-IV, ICD-9, and IDC-10, which are the vocabularies most used in clinical practice in this field. The hierarchies of DSM-IV and ICD conformed with the MT of SNOMED-CT for 47 of the 51 overlapping terms.

## 4 Discussion

We reported a framework for taking an inconsistent manually constructed set of diagnoses and forming a consistent set of diagnostic terms that is based on established medical vocabularies. This standardization of the terms used at the diagnostic list brings at least two advantages (1) allowing knowledge sharing in the domain – the ontology may serve as a starting point for establishing an internationally agreed upon list of developmental disorder diagnoses and (2) allowing the application of computational methods on patients' data to discover new knowledge (e.g., we can now apply clustering techniques for finding patients with similar comorbidities and analyze the similarity).

The framework we proposed showed its potential to identify the most comprehensive vocabulary (SNOMED-CT in our study) of the clinical vocabularies that cover a medical sub-domain, suggesting that it may serve as the main vocabulary used in the clinical sub-domain. In clinical practice, clinicians in Israel do not use SNOMED-CT, and instead use other vocabularies as the basis for their diagnoses lists: DSM-IV, ICD-10, or ICD-9-CM. While each of these vocabularies alone is not comprehensive enough, combining the three of them covers 51 (66.2%) of the terms found in UMLS.

The ontology is not finalized and needs to be further validated by more experts. Special care will be paid to the 11 diagnostic terms that were not found in UMLS. In some cases, they represent new knowledge that has not yet been added to existing vocabularies, and clear definitions must be established to allow for consistency in diagnoses made by different clinicians.

## References

1. Gillberg, C.: Deficits in attention, motor control and perception, and other syndromes attributed to minimal brain dysfunction. In Gillberg C, (ed).: Clinical child neuropsychiatry. Cambrigde University Press, Cambridge, UK (1995) 138-172.
2. Noy N., McGuinness D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report No. KSL-01-05 and Stanford Medical Informatics Technical Report # SMI-2001-0880; (2001).
3. Humpreys B., Lindberg D., Schoolman H., Barnett G. The Unified Medical Language: An Informatics Research Collaboration. J Am Med Inform Assoc. 5:1 (1998) 1-11