Onto-clust –A methodology for combining clustering analysis and ontological methods for identifying groups of comorbidities for developmental disorders

Mor Peleg, PhD[a*] , Nuaman Asbeh, MA[b], Tsvi Kuflik, PhD[a], and Mitchell Schertz, MD[c]

[a]*Department of Management Information Systems, University of Haifa, Israel*

[b]*Department of Statistics, University of Haifa, Israel*

[c]*Institute for Child Development, Kupat Holim Meuhedet, Central Region, Herzeliya, Israel*

All communication should be with:

Mor Peleg

10367 Vista Knoll Blvd.

Cupertino, CA, 95014

peleg.mor@gmail.com

Tel: (408) 733-1531

Fax: (413) 375-3755

Abstract

Children with developmental disorders usually exhibit multiple developmental problems (comorbidities). Hence, such diagnosis needs to revolve on developmental disorder groups. Our objective is to systematically identify developmental disorder groups and represent them in an ontology. We developed a methodology that combines two methods (1) a literature-based ontology that we created, which represents developmental disorders and potential developmental disorder groups, and (2) clustering for detecting comorbid developmental disorders in patient data. The ontology is used to interpret and improve clustering results and the clustering results are used to validate the ontology and suggest directions for its development. We evaluated our methodology by applying it to data of 1175 patients from a child development clinic. We demonstrated that the ontology improves clustering results, bringing them closer to an expert generated gold-standard. We have shown that our methodology successfully combines an ontology with a clustering method to support systematic identification and representation of developmental disorder groups.

Keywords: Ontology, clustering, developmental disorders, comorbidity.

I INTRODUCTION

Childhood developmental disorders affect a wide range of abilities, including cognitive processes, social interaction, medical function, motor skills, and behavior [1]. Although the mechanisms causing disorders are hardly understood, many practitioners are realizing that children who suffer from one developmental disorders often exhibit other, comorbid developmental disorders [1-5], and that the treatment of a child should consider collections of disorders. However, very few comorbidity groups have been reported so far, despite the fact that comorbidities are recognized to be more prevalent than single developmental disorder [1, 3].

Since comorbidities are the norm, the diagnosis of developmental disorders needs to revolve on developmental disorder *groups*. Current classification systems, such as DSM-IV (American Psychiatric Association 1994) and ICD-10 (World Health Organization 1992) do not reflect developmental disorders groups. Our goal is to develop an ontology of developmental disorders that provides clear and detailed definitions and citations of developmental disorders and developmental disorder groups, including among others, risk factors, comorbidities, and manifestations. Such ontology may improve the quality of diagnoses given to children, by assisting practitioners in considering a wider range of options following developmental disorder groups rather than individual disorders. The principles of our approach are presented schematically in Figure 1 and discussed briefly in the rest of this section.

Because the current literature does not contain many publications on developmental disorder groups, the sources of knowledge for constructing the ontology may also come from clinical data of patients diagnosed with developmental disorders. After standardizing the data, unsupervised machine-learning methods may be applied to the data to detect clusters of patients suffering from particular comorbid disorders. These could be suggested as candidates for developmental disorder groups in the ontology, directing the search for literature that would corroborate the clusters' existence or its likelihood for existing, or even trigger new research.

The patient data may contain noise, which may cause machine-learning methods to generate clusters that do not agree with the current medical knowledge. In these cases, we can utilize domain knowledge to help improve the automatic clustering results, leaving clinically-valid clusters that may direct ontology development. Close to its conception, the ontology will contain literature evidence for only part of the relevant knowledge in the field. The knowledge contained in the ontology may have to be supplemented by clinical expert knowledge to improve the automatic clustering results. As more knowledge becomes available in the literature, and as clustering results help direct literature searches, the ontology may become more complete and may serve as a knowledge-base for clinicians.

In this paper, we report our methodology for creating a knowledge-base of developmental disorders and developmental disorder groups, as suggested above, and its evaluation. Throughout the paper, we use the term *cluster* to denote a collection of vectors formed by machine-learning methods, Super Diagnosis Group to denote a group of comorbid developmental disorders derived from the literature and represented in the ontology, and developmental disorder group to denote a data-based group of patients suffering from the same set of comorbid developmental disorders as defined also by an ontology Super Diagnosis Group.

II RELATED WORK

We discuss related research that combined clustering analysis with an ontology and research work that used clustering methods to discover clusters of developmental disorders.

A. Combining clustering methods with ontologies

Several studies in biomedical informatics and other domains have presented methods that combine ontologies with clustering analysis [6-11]. Some approaches attempt to interpret or improve clustering results based on knowledge extracted from the ontology [6-8]. For example, clustering gene-expression [6, 9], where Gene Ontology (GO) [12] was used to narrow the search-space and the hierarchical links in GO were used for interpreting the clusters [6] or to infer the degree of similarity between genes [9], using the graph-based structure of GO. Another example is clustering documents based on a concept hierarchy [7, 8]. Pre-processing steps utilized the ontology to resolve synonyms and introduce more general concepts to easily identify related topics [7].

Other researchers used clustering analysis to support ontology development [10, 11, 13]. Examples include creating an ontology based on clustering of songs that were similarly rated by users [10], constructing an ontology of a computer-science department based on web pages of faculty members [11], and creating a taxonomy of images from hierarchical clustering of images, based on similarity of color and shape [13]. In all of the works discussed above, hierarchical (but not other kinds of) relationships in the ontology were used in clustering analysis or derived from it.

B. Using clustering methods to discover clusters of developmental disorders

Clustering [14, 15] has been used to identify subtypes of pervasive developmental disorders (PDD). Recently, PDD data was clustered using an ensemble of clustering methods [16] to increase the validity of clusters formed, However, none of these studies used an ontology to interpret and improve clustering results.

III RESEARCH DESIGN AND METHODS

Our research objectives were (1) to develop a literature-based and patient data-based ontology of developmental disorders and developmental disorder groups, and (2) systematically identify and represent associations between developmental disorders by combining the domain ontology with clustering of patient data in a complementary and bi-directional way.

To reach these objectives, we developed a methodology that includes the following steps (Figure 1): (A) data cleanup, (B) ontology development, (C) clustering (concurrently with step B), and (D) combining clustering and ontology. The methodology's steps are currently processed manually except of the automatic clustering analysis. The steps are discussed in the rest of this section. As the focus of this paper is mainly on the integration of the ontology with clustering, we cover steps A and B briefly and provide more details in appendices.

A. Creating Consistent Diagnoses List and Cleaning the Data set

Patients' data may contain inconsistent terminology that results in duplication of terms representing the same diagnosis. We created a consistent list by searching for concepts in the Unified Medical Language System (UMLS) [17] Metathesaurus and joining as synonyms diagnoses that corresponded to the same UMLS concept. Several specializations of UMLS concepts that were used in the data were also added [18]. Originally, the data contained 208 medical terms including synonyms and non-standard diagnostic terms. Mapping the terms to the UMLS reduced the set to an equivalent set of 95 consistent diagnoses. 82 of these terms were found in the UMLS (86.3%), and 13 terms were subtypes of terms found in the UMLS (13.7%). Table A.1 in Appendix A shows the terms of the consistent diagnostic list and their Semantic Types. Not all of these terms were diagnoses, as can be seen in Table A.2 of Appendix A, which shows the number of terms belonging to each semantic type. While most of the terms were *Mental or Behavioral Dysfunction* (31) or *Disease or Syndrome* (20), some represented *Findings*, *Signs*, *Mental Processes*, etc.

After creating the consistent diagnosis list, we cleaned the data by replacing the terms found in the clinical data with the terms taken from the consistent diagnosis list.

B. Ontology Development

We developed the ontology in three steps: creating a concept hierarchy, creating detailed concept definitions, and defining super-diagnosis groups. We created the ontology with Protégé-2000 [19]. Figure 2 shows the hierarchy of medical concept classes that represent developmental disorder diagnoses found in the consistent diagnosis list and related medical terms, such as medical procedures and anatomical terms that are used to define the diagnoses. We based the hierarchy on the hierarchy of the vocabulary found in UMLS that covered most of the concepts in the consistent diagnosis list (SNOMED-CT, which covered 86.7% of the terms found in UMLS), as explained in [18]. We defined two metaclasses [19] in the ontology: Concept metaclass and Super-Diagnosis metaclass. Each medical concept class (e.g., Developmental_Coordination_Disorder, DCD – shown in Figure 2) is an instance of the Concept metaclass. The basic slots of the Concept metaclass correspond to the structure of terms in the UMLS Metathesaurus: name, synonyms, semantic types, textual definitions, concept identifier, and source vocabularies. 13 additional slots represent horizontal links (relationships) to other concepts in the ontology and are described in detail in Appendix B. In this paper we focus only on comorbidity and hierarchical (is-a) concepts relationships that are important for the methodology that combines the ontology with clustering.

The knowledge that we entered as detailed concept definitions was based on a literature search performed by clinical experts[1]. In considering which knowledge should be entered initially into the ontology, we turned to our data to see which diagnoses were most prevalent. We assume that inserting detailed knowledge into the most prevalent diagnosis first, will enable ontology-assisted labeling of more clusters than would be possible by any other approach. Two of the 95 consistent diagnoses, DCD and ADHD, are of prevalence above 12% (Table A.3 in Appendix A shows the 12 most prevalent diagnoses). Therefore, our clinical experts conducted a literature review and provided definitions (including comorbidity relationships) for the two most prevalent concepts. Figure 2 shows part of the definition of DCD.

We used the Super-Diagnosis metaclass to represent cliques of developmental disorders that can be inferred from concept relationships represented in the ontology. Super Diagnosis Group classes are defined as instances of this metaclass and correspond to groups of co-occurring disorders. The *core* slot of a Super-Diagnosis Group holds comorbid concepts (i.e., concepts associated with comorbidity relationships). To find cores of Super Diagnosis Groups, we viewed the ontology as a graph in which the links are the comorbidity relations defined in the ontology, and the nodes are the concepts in the ontology. Based on the pair-wise comorbidity links in the graph, we found 27 cliques of concepts that are connected to each other via comorbidity links, shown in Appendix C. For example, the set {DCD, Attention Deficit Disorder with Hyperactivity (ADHD), Learning Disorder (LD)} is a clique since any two of these three concepts are comorbidities of each other; hence it becomes a core of a Super Diagnosis Group. Note that single medical concepts (e.g., {DCD}) that are known from the literature to occur in children without comorbid developmental problems are also considered as Super Diagnosis Groups.

C. Clustering analysis

A wide variety of clustering analysis tools, models, and algorithms exist, from statistics, artificial intelligence, and databases, to machine learning [20]. In this research, we used Self Organizing Map

[1] The experts were Mitchell Schertz and Luba Zuk

(SOM) [21] for clustering analysis. A SOM [21] is an unsupervised competitive artificial neural network used to map high-dimensional data onto a lower, usually 2D representation space. The resultant maps are organized in such a way that similar data are mapped onto the same node or to neighboring nodes in the map. What distinguishes SOMs from other clustering methods is that the clusters on the map are organized. The arrangement of the clusters in the map reflects the topological relationships of these clusters in the input space, and no prior assumption regarding the number of clusters is needed. The fact that SOMs visualize clusters makes it intuitive to detect cluster boundaries within clustering results. SOM is considered as one of a number of high-dimensional data visualization techniques [22] used in visualization of biomedical information, which is often heterogeneous and multi-dimensional. We utilized the unified distance matrix (U-matrix) [23] to identify the clusters generated by SOM. U-matrix visualizes distances between neighboring map units using grey levels (Figure 1, part D).

We selected SOM as our clustering and visualizing technique for the following reasons: (1) SOM is commonly and successfully used in medical informatics [24-27]; (2) Other visualization techniques, such as RadVis [28], RankViz, and FreeViz are more suited to classification problems rather than clustering [22]. In classification problems, the classes are already known and the target is to identify the attributes that most effectively discriminate among members of different classes. In contrast, SOMs do not require prior knowledge on the classes that may exist in the data [21], as in this research; (3) In SOMs there is no need to specify explicitly the number of clusters a-priory as in other clustering algorithms (e.g., K-means) [21]. This characteristic was crucial in our research because we did not have prior knowledge regarding the number of comorbidity groups. Due to the exploratory nature of the work we did not consider any pre-processing process for determining the possible number of clusters (like gap statistics), but preferred to let the expert evaluate the results.

In order to cluster patient data to identify clusters of comorbid diagnoses, we first transferred the consistent patient data from their original textual representation to binary representation. To do so, we assigned each diagnosis from the consistent diagnosis list a unique number between 1 and *n*. For each patient vector, we created an *n*-slot binary vector in which the i^{th} slot holds 1 if the diagnoses vector has the diagnosis numbered by *i*, and 0 otherwise. In cases of diagnoses that had more than binary values (e.g., prematurity<28 week, prematurity 29-32 weeks, prematurity 33-37), we used multiple binary slots, one for recording the existence or inexistence of each value.

We then clustered the vectors using SOM Toolbox [29] using the default sequential training algorithm with Gaussian neighborhood function and the default topology - a 2-dimensional sheet map with hexagonal lattice. While SOM does not require specifying the exact number of clusters a-priori, SOM users are required to choose a map size. The map size used in the SOM algorithm has an impact on the clustering results; starting from a good map that produces clusters that are not too large to aggregate different patient populations and are not too small so that they partition a population into several clusters. While the ontology-assisted corrections may eventually identify the same developmental disorder groups, starting with a good map shortens the number of iterations in the ontology-assisted corrections. To determine the best map size, several maps are produced until the “right” size is selected, usually by domain experts, as suggested for example by [30-32]. However, these works did not suggest a method for determining the best size. Moreover, due to the exploratory nature of the work we did not attempt to find an optimal number of clusters directly, but preferred to examine several maps before selecting the one best suited to our needs. SOM's visualization helps in detecting the boundaries between clusters. Since there was no a priori knowledge about the number of anticipated clusters, we started from a relatively large map (2000) and tried to reduce the size while preserving clusters separation. The map sizes we used were 2000, 1750, 1500, 1250, 1000, 750, 500, and 250. Too dense (few clusters) and too sparse (many small clusters) maps were removed, leaving map sizes of 750, 1000 and 1250. The possibly valid remaining maps were presented to our clinical expert who chose a good map size, i.e., a map containing members that are indeed clinically close (which turned out to be of size 1000). Once the good map had been chosen, we manually partitioned the set of diagnoses in each cluster into two parts. The *common part* (which was chosen as the centroid representing the cluster) contains the set of diagnoses that appear in all of the vectors that belong to a cluster. The *varied part* contains the diagnoses not common to all the vectors.

D. Combining clustering analysis and ontological methods

The ontology and clustering complement each other in identifying developmental disorder groups.

1. Using the ontology to interpret and improve clustering results

SOM produces clusters using a quantitative similarity measure without exploiting prior knowledge regarding the data. As a result, SOM may assign patient vectors into a less appropriate cluster. However, this situation can be fixed using knowledge found in the ontology. A clinical expert can suggest corrections to clusters by splitting clusters that represent more than one comorbidity group and joining clusters that represent the same comorbidity group. Our objective was to develop a methodology that uses the ontological knowledge to suggest the appropriate corrections for clustering results.

The ontology is used to interpret and improve clustering analysis by (1) providing labels to clusters, (2) improving the clustering results, and (3) classifying the clusters according to evidence from the ontology. The first two tasks are explained informally below and in Figure 3, while Appendix D provides a formal specification.

Providing initial labels. Although the ontology does not contain complete definitions of all of the diagnoses, it can still be used for labeling some of the clusters. For a complete discussion of the ontology's labeling potential, please refer to Appendix E. At a given version of the ontology *O*, we determine *C* – the set of all the Super Diagnosis Groups' cores currently defined in the ontology. Providing labels from *C* is done by comparing each core in *C* to the common part of each cluster in a set of clustering results *S*. If the Super Diagnosis Group's core and the common part of the cluster contain the same concepts, then we label the cluster with the Super Diagnosis Group's core (see Figure 3-a). In this comparison, a concept from the cluster's common part that is a specialization (e.g., ADHD-preschool) of a concept from the Super Diagnosis Group's core is considered equal to.its parent concept (ADHD).

Improving clustering results and providing final labels. SOM clusters similar vectors together, based on vector similarity. As a result, on the one hand vectors representing similar patterns with respect to a *common part* (centroid) may be assigned to different clusters, because of their *varied part*. While this is reasonable from the clustering algorithm point of view, this may be wrong clinically. On the other hand, some times, again, due to the nature of the algorithm, vectors that clinically represent different phenomena are assigned to the same cluster, again, due to the concepts included by the SOM in the *common part* of the cluster. In these cases, the knowledge in the ontology may be used to correct the clustering errors, by applying cluster splitting operations (see Figure 3b-i) followed by joining cluster parts that were labeled by the same label (see Figure 3b-ii).

We split clusters that were not split by SOM when the ontology shows that they correspond to more than one Super Diagnosis Group; when the *varied part* of a cluster contains additional concepts (e.g., PDD) that together with the concepts of the *common part* (e.g., DCD-ADHD) represent a core of another Super Diagnosis Group in the ontology (e.g., DCD-ADHD-PDD), we split these vectors from the entire cluster into a new cluster that we label by the more comprehensive core from the ontology.

Classifying the clusters according to evidence from the ontology

We distinguish between three sets of clusters that can be interpreted by the ontology:

(1) Actual clusters that are known from the literature and have literature citations in the ontology

(2) Actual clusters that are *pair-evidenced* - clusters for which there is literature evidence between each pair of concepts belonging to the core; but there is no literature evidence defined in the ontology for the comorbidity between all the core's concepts

(3) Impossible clusters - clusters containing diagnoses that exclude each other (i.e., the ontology contains exclusion criteria for them). These situations arise when invalid patient data was entered by the clinician (e.g., "normal psychomotor development" with "developmental coordination disorder").

2. Using the SOM clustering results to support ontology development

SOM clusters can be used to (1) validate Super Diagnosis Groups defined in the ontology. In addition, clusters that were found by SOM and are not represented in the current ontology can be used to support ontology development by (2) focusing the literature search for possible comorbidity relationships between diagnoses belonging to the same clinically valid cluster, and (3) finding possible exclusion criteria based on invalid clusters.

IV EVALUATION METHODS

We report the methods that we used to evaluate our methodology.

A. Clinical Data

We used clinical data from the Institute for Child Development, Kupat Holim Meuhedet, Central Region, Herzliya, Israel. The data comprise records for 1175 patients who were diagnosed at the institute during the past six years (from Aug 1, 2000 to May 10, 2006). Ethical approval for use of this data was obtained from the Helskinki Committee at the Chaim Sheba Tel Hashomer Medical Center. It should be noted that a major part this data (2000-2005) was also used for defining the methodology.

Following data cleanup, we grouped all diagnoses given to the same patient at a clinical visit into a single vector that uses terms from the consistent diagnoses list. We noticed that a patient can have more than one visit and that some of these visits were medication follow-ups in which the doctors record just diagnoses for which the patients received medication. Therefore, a patient could have a mixture of vectors – some representing diagnoses and some representing medications follow ups. In addition, the number of visits for each patient was not uniform. Hence, we selected one vector per patient.

We examined three strategies to create one vector for each patient: (1) creating one vector that unifies all the diagnoses that were given to the patient based on all of his visits; (2) selecting the first-visit's vector. This strategy is based on the clinical practice in which the clinician strives to provide a comprehensive assessment of the patient at the first visit; (3) the most comprehensive vector - the diagnoses vector composed of the visit on which the patient receives the maximum number of diagnoses.

The first strategy of creating a unified vector of comorbid disorders was ruled out as being clinically invalid. The reason being, that a vector that unifies all the diagnoses given to a patient over several visits may contain diagnoses that were never, at a single point in time comorbid, but were a temporal progression of one diagnosis to another (e.g., ADHD-preschool and ADHD). Future modifications to this strategy that would automatically unify the temporal diagnoses into a single diagnosis may yield additional results, as the two other strategies may miss important diagnoses. In the current study, we used the “first visit” and “most comprehensive” strategies, creating two data sets from the data.

Using the Mann-Whitney test, we found that the distribution of the vector lengths (the number of diagnoses per patient vector) in the two data sets were significantly different (P-value < 0.001) with vectors being longer in the most comprehensive data set. The average vector length for the first visit data set is 2.111 and for the most comprehensive data set is 2.375. In addition, we found that 70.8% of the vectors in the first visit data set and 78.1% of the vectors in the most comprehensive data set contained at least two diagnoses, supporting the premise of this research that developmental disorders are often comorbid.

B. Evaluation of clustering results and ontology-assisted improvements

Our clinical expert evaluated the clinical validity of clusters and produced a partition of vectors, which we used to evaluate the clustering results and their improvement via the ontology.

1. Expert evaluation of clinical validity of clusters

The clinical expert (MS) evaluated the clinical validity of the clusters, differentiating:

1) Clinically valid comorbid clusters – clusters containing patient vectors that reflect a group of disorders that, based on the clinical expert's knowledge and experience, may be comorbid

2) Doubtful clusters – clusters containing diagnoses that normally do not appear together (e.g., "Feeding Disorder of Infancy of Early Childhood" and "Normal psychomotor development")

3) Irrelevant clusters –clinically valid clusters that do not reflect a group of comorbid disorders and therefore were irrelevant to our study. They included:

- Single diagnosis – clusters containing vectors representing different patients that have a single diagnosis each, where the patients' diagnoses were clinically unrelated (e.g., patients with "neglect of child" was grouped with patients with "mathematical disorder")
- Unclear diagnoses – vectors containing the diagnosis "diagnosis deferred", which indicates that the child was not diagnosed
- Retired diagnosis - vectors containing diagnoses that the clinician had used several years ago but no longer uses (e.g., “Memory Disorders”)
- Duplicative diagnoses - vectors containing diagnoses that duplicate each other: “Normal psychomotor development” and “Average intellect”
- Non-specific diagnosis - vectors that contained the non-specific diagnosis “Behavior Disorders” but did not contain other, more specific, diagnoses

4) Invalid clusters – clusters containing diagnoses that exclude each other (e.g., "Gross Motor Developmental Delay" and "Normal psychomotor development")

2. Comparing clustering results to an Expert partition of labeled patient vectors

Clustering algorithm partitions the data set, represented as vectors, into clusters of similar vectors. To evaluate the quality of a SOM partition, it can be compared to a partition of vectors that represents a "gold standard" partition, based on expert opinion (e.g. the "right" way to partition the data set). Following Milligan and Cooper's [33] recommendation, we adopted the adjusted Rand index [34] as the index for comparing the two partitions (e.g., SOM and the "Gold Standard" expert partitions). The adjusted Rand index measures the agreement between two different partitions of the same set of objects (patient vectors), by looking at pairs of objects in the original data set and counting and comparing how many pairs of objects were assigned to the same cluster in both partitions (agreement), and how many pairs of objects were not assigned to the same clusters in both partitions (a disagreement). It includes a correction for chance agreement. The maximum value of the adjusted Rand index is 1 and its expected value in the case of random partitions is 0. A higher index means better correspondence between the partitions. Steinley [35] suggested the following heuristics for determining the quality of cluster recovery relative to adjusted Rand index: (a) values greater than 0.90 are viewed as excellent recovery, (b) values greater than 0.80 are considered good recovery, (c) values greater than 0.65 reflect moderate recovery, and (d) values less than 0.65 reflect poor recovery.

In evaluating the ontology-assisted clustering, we used just the vectors contained in valid clusters that received final labels after applying the split and join methodology. We used the ontology to provide labels to as many of the valid clusters as possible and evaluated only labeled clusters.

To assess the quality of the clusters with final labels, we compared the clustering partition against the clinical expert partition using adjusted Rand index. Our clinical expert partitioned the same set of patient vectors found in the labeled clusters into classes. While theoretically, it may be good to have the clinical expert produce an independent partition, from a practical perspective it is not feasible for the clinical expert to partition 500 vectors in each data set. He thus used the SOM partitioning as a starting point.

After applying the ontology-assisted split and join operations, a final set of labeled clusters was produced and was compared with the expert partition, to evaluate the ontology contribution to clustering.

V RESULTS

In this paper we focus on the methodology that combines clustering with an ontology. Therefore, we report the results of applying the clustering and the ontology-assisted corrections part of the methodology and the evaluation steps, following parts C-E of the flowchart of Figure 1.

A. SOM Clustering Results

The clustering results that were obtained by the map sizes 250 and 500 contained clusters that were too large and those obtained by the map sizes 1500, 1750 and 2000 contained clusters that were too small. Therefore we removed those maps. Our clinical expert (MS) examined the clustering results of the possibly valid map sizes (750, 1000 and 1250) and selected the 1000-cell as the best map.

SOM produced 43 clusters for the first visit set and 55 clusters for the most comprehensive set. 37 and 51 of them, respectively, were determined by our clinical expert to be clinically valid (see Table 1).

B. Combining clustering analysis and ontological methods

We report the results of using SOM clusters to support ontology development followed by the results for using the ontology to label and improve the SOM clusters.

1. Using the SOM clustering results to support ontology development

As the results for the two data sets are similar, we refer in the text to the results of the first visit set and report the results of the two data sets in tables.

(1) Validating Super Diagnosis Groups.17 Super Diagnosis Groups (out of the 27 Super Diagnosis Groups defined in the ontology) were supported by patient data, in both data sets (see Table 1). Hence, the status of these 17 groups is “Actual Group”. 12 of these 17 Super Diagnosis Groups contained multiple diagnoses, supporting the existence of patients with comorbid diagnoses, while the data also supported five Super Diagnosis Groups containing single diagnoses.

(2) Focusing the literature search. Using the 37 valid clusters identified in the clustering results, we identified 26 possible pair-wise comorbidity relationships between concepts, of which 9 relationships were already defined in the ontology, and 17 were possible comorbidity relationships on which the literature search should focus. Table 2 shows the possible comorbidity relationships (26 using the first visit set and additional 6 relationships from the most comprehensive visit data set).

(3) Finding possible exclusion criteria that should be defined in the ontology. Based on the two invalid clusters identified at the clustering results (see Table 1, footnote b), we identified two exclusion criteria that were not defined in the ontology and should be added to it.

2. Ontology-assisted clustering improvement and labeling

After applying the ontology-assisted split and join operations, we used the 27 Super Diagnosis Groups in the ontology to provide final labels to the resulting final clusters. The ontology-assisted operations helped in discovering more clusters with unique labels. As shown in Table 3 for the first visit data set (and Table 4 for the most comprehensive data set), the ontology provided 11 labels to 16 clusters in the SOM results. 6 additional labels were utilized only as a result of applying ontology-assisted split operations, resulting in 17 labels provided by the ontology to label the final cluster set in each data set. In this way, we provided a better interpretation for those clusters and may identify more groups of comorbidities. As reflected in Table 3, the ontology-assisted improvements utilized 4 split operations and 7 join operations, raising the total number of clusters by one - from 43 to 44.

For both data sets, the same set of 17 labeled clusters was produced. We classified the clusters according to evidence from the ontology (see section III.D.1). Out of these clusters, 14 were classified as “known from the literature” (see tables 3 and 4 for specific references), and 3 as pair-evidenced.

As shown in Table 3 (Table 4 for the most comprehensive data set), 7 clusters in the final clusters results were problematic since they were very small clusters, containing at most 3 vectors. In this research, we accepted these clusters as separate clusters. We note that the classification of 5 of these clusters is "known from the literature" (rows 10-14 of Table 3 and rows 11-14 of Table 4). This justifies the decision to accept them as separate clusters despite their size. The two other small clusters were "pair-evidenced".

C. Evaluation of Ontology-Assisted Clustering using adjusted Rand index

We used the set of heuristics for determining cluster recovery quality, related to adjusted Rand index (see section IV.B.2), to evaluate the quality of the SOM clusters that could be labeled by the ontology and the quality of the final set of ontology-improved clusters, both in relation to the expert partition.

We measured the agreement of clustering results with the clinical expert partition using the adjusted Rand index. Our clinical expert partition contained 21 classes of vectors. For the first visit data set, we found that the adjusted Rand index improved from 0.8234 (good) to 0.964 (excellent) when applying the ontology corrections. For the most comprehensive data set, the adjusted Rand index improved from 0.641 (poor) to 0.953 (excellent).

VI DISCUSSION

We were able to demonstrate the feasibility and value in combining clustering analysis with an ontology of developmental disorders that we developed in order to support systematic identification and representation of developmental disorder groups. As we demonstrated, clustering yields in addition to valid clusters, invalid clusters, irrelevant clusters, and clusters that contain a mixture of several patterns of comorbidity groups. Applying ontological domain knowledge improves the clustering results in terms of (1) their agreement with a clinical expert's partition, (2) the homogeneity and number of clusters that could be labeled and interpreted, and (3) the number of actual labeled comorbid clusters that could be proposed as new diagnostic terms in the field of developmental disorders. Similarly, we have shown that using the ontology alone yields theoretical groups that are not corroborated by patient data and may not represent reality. Combining the ontology with patient data obtained via cluster analysis allowed identification of developmental disorder groups that exist in clinical practice. Hence, our demonstration of combining cluster analysis with an ontology was beneficial over each technique used separately. Most significantly, this combination allowed us to begin systematic identification of developmental disorder groups that occur in reality. Such identification has not been done to date and this approach may improve the ability to accurately identify cases of developmental disorder groups, which indirectly can improve research about developmental disorders.

A prerequisite to the work performed was the development of an ontology of developmental disorders. We created a basic but functional ontology that allowed us to convert the data into consistent terms, provide labels for clusters created by machine-learning techniques, and improve the clusters based on domain knowledge. However, the ontology development is a work in-progress. The ontology's potential to interpret and improve the clustering results will grow as more knowledge is added into it. Currently, the ontology consists of the current 27 Super Diagnosis Groups defined in it (Appendix C), based on complete definitions for two concepts only (DCD and ADHD). Despite the limited number of concepts that are fully defined, the ontology, when combined with the clustering results, was able to interpret and provide labels for 17 clusters detected in real patient data (table 3 and 4). These 17 developmental disorder groups will be suggested as new diagnostic terms in the field of developmental disorders. Whereas 16 of these developmental disorder groups are already known from the literature, one developmental disorder group – DCD-ADHD-SLI – is a novel developmental disorder group that has not been reported. Based on the ontology-improved clustering results, we found a cluster of 32 patients that share these three co-occurring diagnoses. Its existence is theoretically possible based on literature citations for co-occurrence of pairs of these three concepts. Of course, we need to reproduce these findings using different validation data sets and use expert partitions formulated by other experts than our clinical co-author.

Apart from the novel findings that we discovered from combining the ontology and clustering results, the stand-alone SOM results also proved beneficial. From the SOM clusters that related to developmental disorders for which we did not have complete ontological definitions, we were able to identify comorbidity links that may direct the literature search for knowledge that could be incorporated into the ontology (see the 23 possible comorbidity relationships in Table 2).

While currently, the literature search and knowledge incorporation is manual, future research should consider the best method to automatically maintain an updated ontology as the literature evolves, such as natural language processing methods.

Our long-standing goal is to develop an ontology of developmental disorders that provides clear and detailed definitions and citations of developmental disorders and developmental disorder groups, based on evidence from the literature and from patient data; the clustering results serve the purpose of developing this ontology. When the ontology of developmental disorders will be completed and validated with more experts and more patient data sets, it could potentially:

(1) Increase the consistency of terminology in the field

(2) Ease communication between experts in the domain and ease research collaboration

(3) Yield consistent sets of data that can be analyzed to discover diagnoses clusters

Combining the ontology with evidence from naturally-occurring developmental disorder clusters may

(4) Suggest candidates for new terms in the field which reflect developmental disorder groups, akin to other known entities, such as the previously mentioned deficits in attention, motor control, and perception (DAMP) or Non-Verbal Learning Disabilities (NVLD). Based on the results of the two data sets examined in this work, twelve such new terms were suggested (see Table 2)

(5) Serve as an online aid for reminding clinicians of the definitions of particular diagnoses as well as for directing clinicians in examining relevant comorbidities

(6) Provide researchers with more homogenous groups of subjects. This should significantly contribute to our ability to investigate etiology as well as to evaluate treatment and intervention programs, which are sorely lacking in the field.

The methodology that we have developed could contribute to informatics research. Our methodology is different from previously published work discussed in Section II.A [6-9], where researchers also used ontological knowledge to guide and improve clustering results or used clustering results to guide ontology development [10, 11, 13], but never used both approaches together. In the related work, the degree of similarity of co-occurring concepts corresponded to *hierarchical* links in the ontology, denoting is-a or part-of relationships between concepts. In our work, co-occurring developmental disorders are not necessarily hierarchically similar to each other; we used mainly knowledge about comorbidity relationships between concepts as a measure of concept co-occurrence. Apart from its novelty, our methodology may contribute by its potential to be applied to other domains, where a high level of co-existing information exists..

Comparing the two strategies used for creating the diagnoses vectors for the patients using the clinical data, the first visit data set yielded better results in terms of agreement with the clinical expert partition, even before applying the ontology-assisted methodology of split and join operation. This is probably due to the fact that the most comprehensive data set had more diagnoses per vector, which resulted in greater degree of variability between the vectors, creating a larger number of clusters.

However, applying the methodology to the most-comprehensive set yielded better results, in terms of fewer problematic clusters, and better justification for accepting these clusters despite their small size. The evaluation of the final clustering results obtained from both data sets using adjusted Rand index demonstrated our methodology's potential to bring the clustering results closer to the clinical expert partition (to the same extent) despite the difference of the initial SOM partition.

In future work, we recommend using informative data sets, such as the most comprehensive visit strategy and the unified visits strategy while unifying temporal diagnoses (see Section IV.A) because, theoretically, they could suggest more comorbidity relationships that should be searched for in the literature, as compared to the first visit data.

In this study, we chose SOM as a clustering method due to its flexibility and the fact that it visualizes the clustering results in a way that helps get an impression of how the clusters are spaced. These options were extremely important for us, due to the exploratory nature of the work. This is why a mathematical approach of gap statistics was not considered, even though this may have focused us better to the "vicinity" of the right map size. We do not believe that different algorithms would give different results, but due to the exploratory nature of the work, the option of comparing several clustering approaches was not relevant at this time, yet this remains to be demonstrated in future work.

A. Study limitations

This study had several limitations. First, about 80% of the data was initially used for the development of our methodology and later on for its evaluation. Although we used two data sets to validate our methodology, obtaining similar results for both, these data sets were derived from the same source and 82% of their vectors are identical, since the most comprehensive visit for a patient may be his first visit. In addition, the clinical data used to evaluate our methodology was gathered at a single clinic for one population of patients (children from the central region in Israel). This data was created by a single clinician who did not make use of standardized vocabularies to diagnose children and inconsistencies in the terminology and partial diagnoses given to patients were found in the clinical data. More evaluation remains to be done on other data sets.

Because the clinical data was collected at the same institute as that of the clinical expert, who also developed the ontology, and partitioned the patients into clusters to which the ontology-assisted clustering is compared, more potential problems exist. First, although the comorbidity relationships in the ontology were generated from the literature, they are nevertheless likely to be influenced by the experiences of the expert. Since the expert, being from the same institute, is familiar with the comorbidity patterns in the data set, will likely generate an ontology that contains comorbidity relationships that match those in the data set. Thus, the ontology may be unusually well suited for distinguishing the patterns observed in the data set, and the performance measurements may not apply to data from another institute that may have different comorbidity patterns.

Second, the expert partitioned the patients into clusters that are compared against the ontology-assisted clusters. Because the same expert created the ontology, in effect, the same knowledge base (the ontology and expert), and potentially the same algorithm is being used to refine the clusters. Thus, it would not be surprising that the final clusters are similar. Future studies should be conducted to evaluate the results with other experts.

We also acknowledge that taking the SOM partitioning as a starting point for creating the expert partition, while being practically the only feasible way (as manually classifying 500 vectors is not reasonable), could potentially bias the results.

Another limitation is that the ontology is still incomplete. Therefore our evaluation focused only on clusters that we may label and improve by the ontology.

In addition, we did not target the issue of problematic small clusters that were produced after applying our methodology. We justified their acceptance based on medical literature. But future research needs to define acceptance criteria.

Lastly, all of our methodology's steps were processed manually except for (1) cleaning the data set, (2) transferring both data sets to binary representation, and (3) applying clustering analysis by SOM using the SOM Toolbox. The fact that many steps were created manually could result in errors that may be introduced by manual processing. We suggest automating our methodology in future work, especially the step of defining Super Diagnosis Groups in the ontology, which we modeled as finding cliques in a graph. Although this problem is NP-complete, in our case it should not be problematic since the number of nodes in the graph (representing concepts in the ontology that are associated via comorbidity links) is considerably small – up to 95 concepts for our data set. We also suggest automating the labeling methodology, for which we provided a formal definition using set theory notations. We further suggest automating the interpretation of the U-matrix visualization of the SOM maps labeled with the BMUs and generating an output file with the clustered patient's vectors.

VII CONCLUSION

We have reported and evaluated a methodology that combines clustering analysis with an ontology of developmental disorders in order to support systematic identification and representation of literature-based and data-evidenced developmental disorder groups. From the perspective of medical informatics, this research is novel in that it combines clustering analysis with a knowledge base (ontology) in a bi-directional way, utilizing different types of concept links in the ontology (hierarchical links and comorbidity links), and may potentially be applied to other domains apart from developmental disorders.

ACKNOWLEDGEMENTS

Supported in part by Caesarea Rothschild Foundation Institute for Interdisciplinary Applications of Computer Science, Haifa University. The authors thank Dr. Lidia Gabis and Rubi Tamir for making this work possible and Dr. Luba Zuk for her part in the literature review on DCD. We thank the anonymous reviewers for their insightful comments that helped us to improve the paper, including its focus and structure.

REFERENCES

[1] Gilberg C. Deficits in attention, motor control, and perception: a brief review. Arch Dis Child 2003;88:904-10.

[2] Webster RI, Majnemer A, Platt RW, Shevell MI. Motor function at school age in children with a preschool diagnosis of developmental language impairment. J Pediatr 2005;146(1):80-5.

[3] Piek JP, Dyck MJ. Sensory-motor deficits in children with developmental coordination disorder, attention deficit hyperactivity disorder and autistic disorder. Hum Mov Sci 2004;23(3-4):475-88.

[4] Gross-Tsur V, Shalev RS, Manor O, Amir N. Developmental right-hemisphere syndrome: clinical spectrum of the nonverbal learning disability. J Learn Disabil, 1995;28(2):80-6.

[5] Wilson PH. Practitioner review: approaches to assessment and treatment of children with DCD: an evaluative review. J Child Psychol Psychiatry 2005;46(8):806-23.

[6] Liu J, Wang W, Yang J. A framework for ontology-driven subspace clustering. In: Proc of the 10th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining (SIGKDD); 2004. p. 623-628.

[7] Hotho A, Maedche A, Staab S. Text Clustering Based on Good Aggregations. Kunstliche Intelligenz (KI) 2002;16(4):48-54.

[8] Yoo I, Hu X. Clustering Ontology-enriched Graph Representation for Biomedical Documents based on Scale-Free Network Theory. In: 3rd Intl IEEE Conf on Intelligent Systems; 2006. p. 851-858.

[9] Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, et al. A knowledge-based clustering algorithm driven by Gene Ontology. J Biopharm Stat. 2004;14(3):687-700.

[10] Clerkin P, Cunningham P, Hayes C. Ontology discovery for the Semantic Web using hierarchical clustering. In: Semantic Web Mining Workshop; 2001.

[11] Elliman D, Rafael J, Pulido G. Automatic Derivation of On-line Document Ontology. In: Intl Workshop on Mechanisms for Enterprise Integration: From Objects to Ontology (MERIT 2001) 15th European Conf on Object Oriented Programming; 2001.

[12] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25(1):25-9.

[13] Khan L, Wang L. Automatic Ontology Derivation Using Clustering for Image Classification. In: Proc. of Eighth Intl Workshop on Multimedia Information Systems; Tempe, Arizona; 2002. p. 56-65.

[14] Rescorla L. Cluster analytic identification of autistic preschoolers. J Autism Dev Disord 1988;18(4):475–92.

[15] Beglinger LJ, Smith HS. A review of subtyping in autism and proposed dimensional classification model. J Autism Dev Disord. 2001;31(4):411-22.

[16] Shen JJ, Lee PH, Holden JJA, Shatkay H. Using Cluster Ensemble and Validation to Identify Subtypes of Pervasive Developmental Disorders. In: Proc AMIA Symp; Chicago; 2007. p. 666-70.

[17] Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. J Am Med Rec Assoc 1990;61(5):40-42.

[18] Asbeh N, Peleg M, Schertz M. Creating Consistent Diagnoses List for Developmental Disorders Using UMLS. In: Next Generation Information Technologies and Systems, Lecture Notes in Computer Science, Vol. 4032, Springer Berlin / Heidelberg; Kibbutz Shefayim, Israel; 2006. p. 333-6.

[19] Gennari J, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, et al. The Evolution of Protege: An Environment for Knowledge-Based Systems Development. Intl J of Human-Computer Interaction 2002;58(1):89-123.

[20] Jain AK, Murty MN, Flynn PJ. Data Clustering: A Review. ACM Computing Surveys 1999;31(3):264-323.

[21] Kohonen T. Self-Organizing Maps. New York: Springer-Verlag; 1997.

[22] Grinstein G, Trutschl M, Cvek U. High-Dimensional Visualizations. In: Proceedings of the Visual Data Mining workshop, 2001; San Francisco, CA; 2001.

[23] Ultsch A, Guimaraes G, Korus D, Li H. Knowledge Extraction from Artificial Neural Networks and Applications. In: TAT &World Transpter Congress; Aachen, Germany: Springer; 1993. p. 194-203.

[24] Nelson DW, Bellander B, MacCallum RM, Axelsson J, Alm M, Wallin M, et al. Cerebral microdialysis of patients with severe traumatic brain injury exhibits highly individualistic patterns as visualized by cluster analysis with self-organizing maps. Critical Care Medicine 2004;32(12):2428-2436.

[25] Chen D, Chang RF, YL YLH. Breast cancer diagnosis using self-organizing map for sonography. Ultrasound Med Biol 2000;26(3):405-11.

[26] Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Lett 1999;451(2):142-6.

[27] Yan S, Abidi SS, Artes PH. Analyzing Sub-Classifications of Glaucoma via SOM Based Clustering of Optic Nerve Images. Stud Health Technol Inform 2005;116:483-8.

[28] McCarthy JF, Marx ka, Hoffman PE, Gee AG, O'Neil P, Ujwal ML, et al. Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. Ann N Y Acad Sci 2004;1020:239-62.

[29] SOM toolbox. In.http://www.cis.hut.fi/projects/somtoolbox, last accessed May 12, 2008

[30] Niskanen M, Silven O, Kauppinen H. Experiments with SOM based inspection of wood. In: Proc Intl Conf on quality control by artificial vision; Le Creusot, France; 2001. p. 311-316.

[31] Wu S, Chow TWS. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. Pattern Recognition 2004;3(2):175-188.

[32] Park YS, Cereghino R, Compin A, Lek S. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling 2003;160:265-280.

[33] Milligan GW, Cooper MCa. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research 1986;21:441-458.

[34] Hubert L, Arabie P. Comparing partitions. J of Classification 1985;2:193-218.

[35] Steinley D. Properties of the Hubert-Arabie adjusted Rand index. Psychol Methods 2004;9(3):386-96.

Table 1. The number of valid and invalid clusters obtained for the two data sets

Table 2. Possible pair-wise comorbidity relationships identified in the two data sets, based on clinically valid clusters.

ADHD - Attention Deficit Disorder with Hyperactivity; DCD- Developmental Coordination Disorder; SLI - Specific Learning Disorder; LD - Learning Disorders; PDD - Pervasive Developmental Disorder; TIC - Tic disorders; ODD - Oppositional Defiant Disorder; AI - Average intellect; GMDD - Gross motor development delay, TOR- Torticollis; Hyp - Hypotonia; SLI-C – SLI-Combined; PRE - Prematurity; BIF - Borderline Intellectual Functioning ; NPD - Normal psychomotor development;

Table 3. Final table presenting the labels and labeled clusters (first visit data set). The 'Actual Labels' column shows the 17 final labels provided by the ontology after the split and join operations. The initial 11 labels that were provided for original SOM clusters (before the splits and joins) are shown in grey. The labels provided for new clusters, identified only after applying the labeling methodology that used ontology-assisted split and join operations, are shown in white. The 'labeled clusters' column shows the temporary identifiers given to the original SOM clusters after applying the split operations followed by the cluster numbers after applying the labeling methodology. More than one identifier at an arrow's tail reflects a join operation. Latin letters were added to the end of the identifiers of clusters that were split (cluster parts). The numbers in parentheses denote the sizes of cluster (or cluster parts). Citations are provided for clusters with comorbid diagnoses that are known from the literature.

Table 4. Final table presenting the labels and labeled clusters for the most comprehensive visit data set

Figure 1. A flowchart representing the methodology for identifying and defining groups of comorbid disorders. Steps in our methodology are depicted as ovals. A child with development disorders (i) is diagnosed by a doctor who uses domain concepts from a diagnostic list (ii) to insert the child's list of comorbidities into the electronic medical record (EMR) (iii). Based on knowledge found in the Unified Medical Language System (UMLS), domain concepts are arranged in a consistent diagnostic list (A). The consistent list serves as the basis for selecting knowledge from the UMLS and from the medical literature to develop an ontology (B) of developmental disorders. Next, clustering is applied to patient data that have been cleaned-up using concepts from the consistent diagnoses list (C). The ontology and clustering results are combined (D) in the following way. Clusters found by clustering techniques direct future literature searches for entering detailed knowledge into the ontology and validating the super-diagnosis groups defined in the ontology (D-1). The ontology is used to interpret (provide labels) and correct clustering results (D-2). Finally, the ontology-assisted clustering results are evaluated (E). SDG – Super Diagnosis Group

Figure 2. Part of the Ontology showing the concept hierarchy on the left and a detailed definition of one of the medical concepts (Developmental Coordination Disorder (DCD)) on the right. The name, concept ID, (source) vocabularies, semantic types and synonyms were taken from UMLS. The figure shows three of the 13 relationships (horizontal links) that the DCD concept has with other concepts, namely the comorbidities, risk factors, and functional manifestations relationships. The insert on the bottom shows the details of one of the comorbidity relationships (DCD co_occurs_with ADHD), where the relationship type (taken from UMLS) is co-occurs-with.

Figure 3. Ontology-assisted labeling and split & join operations for improving the clustering results. (a) When a cluster contains patient vectors that equal to the core of a Super Diagnosis group at the ontology, SD2, this cluster can be labeled by that SD. (b) Split and join operations. The clinical expert in our team (MS) suggested to (b-i) split cluster #18 that contains 30 DCD-ADHD patient vectors and 4 DCD-ADHD-PDD patient vectors into two separate cluster parts. He also suggested (b-ii) to join into one cluster the DCD-ADHD-PDD cluster part with another cluster, #22, which contained 20 DCD-ADHD-PDD vectors. The ontology can suggest the appropriate corrections instead of the clinical experts in the following way. When a cluster contains patient vectors that equal to the cores of two different super-diagnosis groups defined in the ontology, SD1 and SD2, this cluster can be split into two cluster parts (b-i). Then, clusters and cluster parts which contain patient vectors that equal to the core of the same super-diagnosis-group (SD2) can be joined (b-ii).

Cluster type		First visit	Most comprehensive
Total number		43	55
Clinically-valid		35	45
Doubtful		1[a]	1[a]
Irrelevant	total	5	7
	Single-diagnosis	1	1
	Retired-diagnosis	2	3
	Duplicative-diagnoses	1[b]	2[b]
	Non-specific diagnosis	1	1
Clinically invalid		2[c]	2[c]

[a] Feeding Disorder of Infancy of Early Childhood & Normal psychomotor development

[b] Normal psychomotor development & Average intellect

[c] Gross Motor Developmental Delay & Normal psychomotor development; Specific Language Impairment & Normal psychomotor development

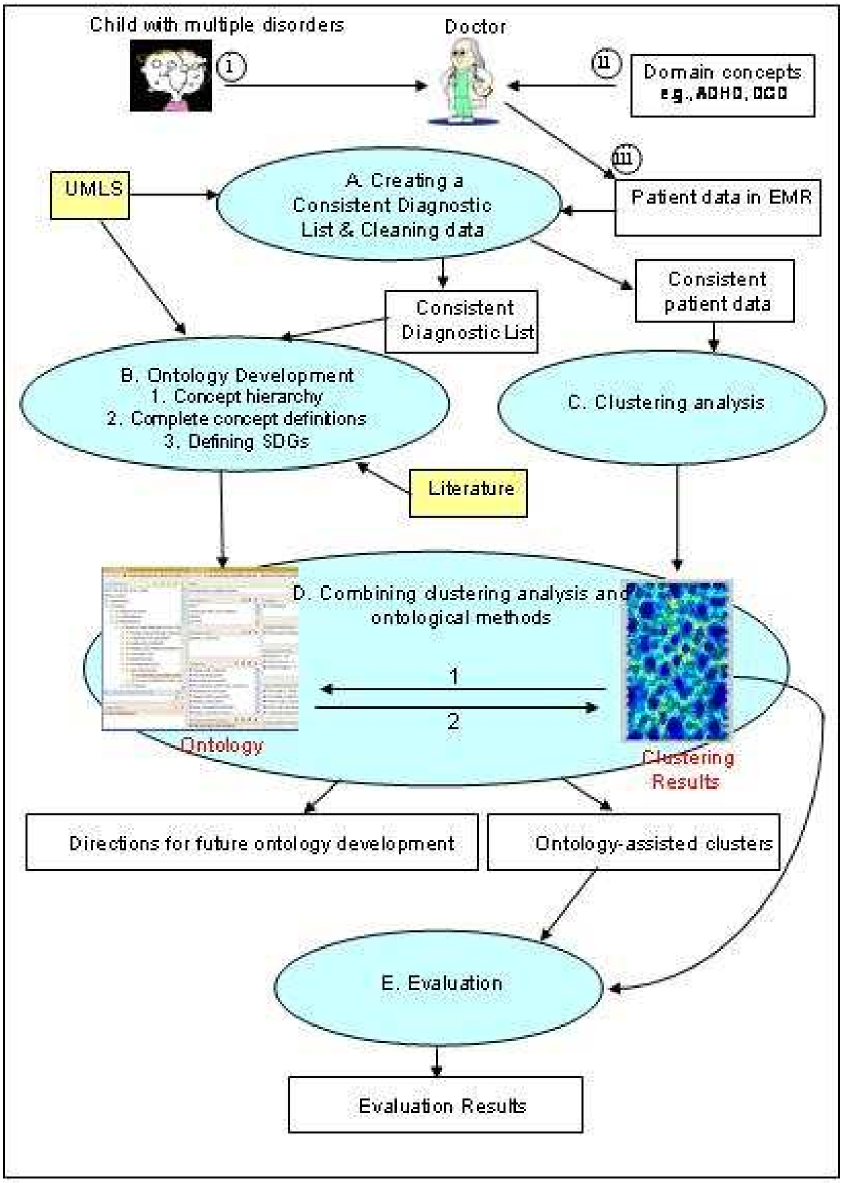
	Possible Comorbidity Relationship	Evidencing Clusters (most comprehensive data set)	Evidencing Clusters (first visit data set)	Defined in the Ontology
1	ADHD, SLI	36, 49, 52,53,54	28, 31, 41, 43	Yes [18]
2	DCD, LD	40, 55	40, 34	Yes [21]
3	DCD, SLI	42, 49, 50, 51, 53, 54	41, 42, 43	Yes [2, 17]
4	DCD, ADHD	45, 46, 476, 49, 53, 54, 55	35, 36, 40, 41, 43	Yes [1, 12]
5	ADHD, LD	38, 39, 55	26, 29, 40	Yes [19]
6	PDD, ADHD	28, 46, 55	26, 40	Yes [19]
7	ADHD, TIC	34, 38	27	Yes [19]
8	PDD, DCD	40, 46, 55	34, 40	Yes [20]
9	ADHD, ODD	34, 38	25, 26	Yes [19]
10	AI, GMDD	1, 2, 3, 4	1, 2, 6	No
11	TOR, AI	2	2	No
12	GMDD, HYP	1, 5, 11	6, 7, 9, 16	No
13	HYP, AI	1	6	No
14	HYP, SLI	30	33	No
15	PDD, DD	13, 21	12	No
16	DCD, HYP	30	32, 33	No
17	SLI-C, ADHD	47	36	No
18	SLI-C, DCD	47, 48	36, 37	No
19	TOR, GMDD	2, 5, 6	2, 3, 9	No
20	PRE, TOR	10	5	No
21	TOR, HYP	5	9	No
22	GMDD, BIF	11	7	No
23	HYP, BIF	11	7	No
24	NPD, PRE	10	5	No
25	DCD, BD	32	38	No
26	SLI, AI	42		No
27	TOR ,NPD	9, 10	5	No
28	AI, PRE	3		No
29	GMDD, PRE	3		No
30	SLI-C, BIF	24		No
31	SLI-C, DD	24		No
32	DCD, AI	41, 42		No

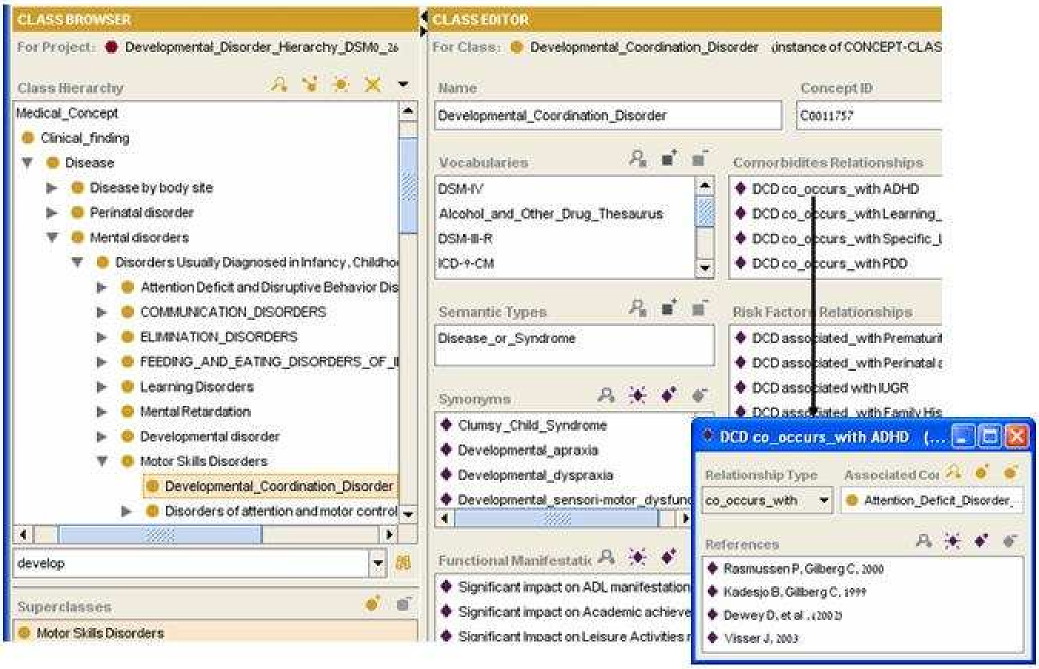
	Actual Label	Label evidence (in the ontology)	Labeled Clusters	Cluster size
1	DCD	known from literature	#34A(95) → #34(95)	95
2	DCD-ADHD	known from literature [1, 12]	#35(34),#40A(46) →#35(81)	81
3	DCD-SLI	known from literature [2, 17]	#42(62) → #42(62)	62
4	ADHD	known from literature	#25A(15),#26A(31) →#26(46)	46
5	SLI	known from literature	#30(45) →#30(45)	45
6	ADHD-SLI	known from literature [18]	#28(10),#31(18) →#31(28)	28
7	PDD	known from literature	#12(11), #20(13) →#20(24)	24
8	ADHD-LD	known from literature [19]	#29(15),#26C(2) →#29(17)	17
9	LD	known from literature	#23(8) →#23(8)	8
10	ADHD-TIC	known from literature [19]	#27(3) →#27(3)	3
11	ADHD-ODD	known from literature [19]	#25B(2),#26B(1) →#25(3)	3
12	DCD-LD	known from literature [21]	#34B(1) →#28(1)	1
13	DCD-PDD	known from literature [20]	#34C(1) →#41(1)	1
14	ADHD-PDD	known from literature [19]	#26D(1) →#12(1)	1
15	DCD-ADHD-SLI	Pair-evidenced	#41(20), #43(12) →#43(32)	32
16	DCD-ADHD-LD	known from literature [1]	#40B(2) →#44(2)	2
17	DCD-ADHD-PDD	known from literature [12[	#40C(1) →#40(1)	1

DCD - Developmental Coordination Disorder; ADHD - Attention Deficit Disorder with Hyperactivity; SLI - Specific language impairment; LD - Learning Disorders; PDD - Pervasive Developmental Disorder; ODD - Oppositional Defiant Disorder

	Actual Label	Label evidence in the ontology	Labeled Clusters	Cluster Size
1	DCD-ADHD	known from literature	#45B(1),#46A(23),#55A(59)→#53(83)	83
2	DCD	known from literature [1, 12]	#40A(53),#45A(7)→#40(60)	60
3	DCD-SLI	known from literature [2, 17]	#40B(1), #50(45),#51(10)→#49(56)	56
4	ADHD	known from literature	#34A(15),#38A(30)→#38(45)	45
5	SLI	known from literature	#37(37)→#37(37)	37
6	ADHD-SLI	known from literature [18]	#36(20),#52(8)→#36(28)	28
7	ADHD-LD	known from literature	#38C(4),#39(22)→#39(26)	26
8	PDD	known from literature [19]	#13(11),#23(7)→#23(18)	18
9	LD	known from literature	#28(8)→#28(8)	8
10	ADHD-TIC	known from literature [19]	#34C(1),38E(4)→#50(5)	5
11	ADHD-ODD	known from literature [19]	#34B(1),#38B(2)→#34(3)	3
12	ADHD-PDD	known from literature [19]	#38D(3)→#13(3)	3
13	DCD-LD	known from literature [21]	#40D(1)→#51(1)	1
14	DCD-PDD	known from literature [20]	#40C(1)→#44(1)	1
15	DCD-ADHD-SLI	Pair-evidenced	#49(24),#53(11),#54(10)→#48(45)	45
16	DCD-ADHD-LD	known from literature [1]	#55B(6)→#52(6)	6
17	DCD-ADHD-PDD	known from literature [12]	#46B(1),#55C(4)→#45(5)	5

ADHD - Attention Deficit Disorder with Hyperactivity; DCD- Developmental Coordination Disorder; SLI - Specific Learning Disorder; LD - Learning Disorders; PDD - Pervasive Developmental Disorder; TIC - Tic disorders; ODD - Oppositional Defiant Disorder ;


Child with multiple disorders
i
Doctor
ii
Domain concepts
e.g., ADHD, OCD
iii
Patient data in EMR
UMLS
A. Creating a Consistent Diagnostic List & Cleaning data
Consistent Diagnostic List
Consistent patient data
B. Ontology Development
1. Concept hierarchy
2. Complete concept definitions
3. Defining SDGs
C. Clustering analysis
Literature
D. Combining clustering analysis and ontological methods
1
2
Ontology
Clustering Results
Directions for future ontology development
Ontology-assisted clusters
E. Evaluation
Evaluation Results



CLASS BROWSER

For Project: Developmental_Disorder_Hierarchy_DSM0_26

Class Hierarchy

- Medical_Concept
- Clinical_finding
  - Disease
    - Disease by body site
    - Perinatal disorder
    - Mental disorders
      - Disorders Usually Diagnosed in Infancy, Childho
        - Attention Deficit and Disruptive Behavior Dis
        - COMMUNICATION_DISORDERS
        - ELIMINATION_DISORDERS
        - FEEDING_AND_EATING_DISORDERS_OF_I
        - Learning Disorders
        - Mental Retardation
        - Developmental disorder
        - Motor Skills Disorders
          - Developmental_Coordination_Disorder
          - Disorders of attention and motor control

develop

Superclasses

- Motor Skills Disorders

CLASS EDITOR

For Class: Developmental_Coordination_Disorder (instance of CONCEPT-CLAS

Name: Developmental_Coordination_Disorder

Concept ID: C0011757

Vocabularies

- DSM-IV
- Alcohol_and_Other_Drug_Thesaurus
- DSM-III-R
- ICD-9-CM

Comorbidites Relationships

- DCD co_occurs_with ADHD
- DCD co_occurs_with Learning_
- DCD co_occurs_with Specific_I
- DCD co_occurs_with PDD

Semantic Types

- Disease_or_Syndrome

Risk Factors Relationships

- DCD associated_with Prematurit
- DCD associated_with Perinatal c
- DCD associated with IUGR
- DCD associated_with Family His

Synonyms

- Clumsy_Child_Syndrome
- Developmental_apraxia
- Developmental_dyspraxia
- Developmental_sensori-motor_dysfunc

Functional Manifestatic

- Significant impact on ADL manifestation
- Significant impact on Academic achieve
- Significant Impact on Leisure Activities r

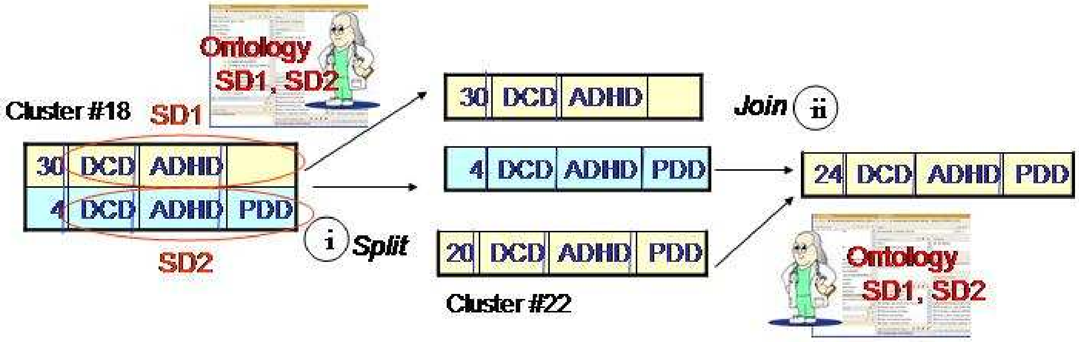
DCD co_occurs_with ADHD (...

Relationship Type: co_occurs_with

Associated Co: Attention_Deficit_Disorder_

References

- Rasmussen P, Gilberg C, 2000
- Kadesjo B, Gilberg C, 1999
- Dewey D, et al., (2002)
- Visser J, 2003



Cluster #18

30 | DCD | ADHD

4 | DCD | ADHD | PDD

30 | DCD | ADHD

4 | DCD | ADHD | PDD

20 | DCD | ADHD | PDD

Cluster #22

24 | DCD | ADHD | PDD