
Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

# Project No. 611391

# DREAM
# Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Agreement Type: Collaborative Project
Agreement Number: 611391

# D2.2.2 Tools for the assessment of child-robot interaction and diagnostics

Due Date: 01/04/2018
Submission date: **10/05/2019**

Start date of project: **01/04/2014** Duration: **60 months**

Organization name of lead contractor for this deliverable: **Babes Bolyai University**

Responsible Person: **Daniel David**

Revision: **1.0**

| **Project co-funded by the European Commission within the Seventh Framework Programme** | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | **PU** |
| **PP** | Restricted to other programme participants (including the Commission Service) | |
| **RE** | Restricted to a group specified by the consortium (including the Commission Service) | |
| **CO** | Confidential, only for members of the consortium (including the Commission Service) | |

***Contents***

# Executive Summary

Deliverable 2.2.2 summarizes data from studies that test the use of multi-sensory systems for the assessment of child-robot interaction and diagnostic, developed in work packages WP4 and WP5. This deliverable presents the rationale of the task, as well as results from task T2.2 concerning the comparison between the input provided by clinicians and the automatically derived diagnostic data recorded in task T5.4 (i.e., system's annotations).

# Principal Contributors

The main authors of this deliverable are as follows (in alphabetical order):
Daniel David, Babes-Bolyai University
Anca Dobrean, Babes-Bolyai University
Silviu Matu, Babes-Bolyai University
Radu Soflau, Babes-Bolyai University
Aurora Szentagotai, Babes-Bolyai University

# Revision History

**Version 1.0 (28-03-2017)**

First draft, describing the theoretical background for the development of automated diagnosis and intervention tools.

**Version 2.0 (31-03-2019)**

Final version, describing the final results on assessment.

# Task rationale

A series of assessment tools have been developed for the evaluation of autistic symptoms and diagnosis based on DSM criteria. Of these, the Diagnosis Interview Revised (ADI-R; Rutter & Le Couteur, 2003), the Autism Diagnostic Observation Schedule Generic (ADOS-G; Lord et al., 2000), the Childhood Autism Rating Scale (CARS; Schopler, Reichler, DeVellis, & Daly, 1980), the Diagnostic Interview for Social and Communication Disorders (DISCO; Wing, Leekam, Libby, Gould, & Larcombe, 2002), and the Gilliam Autism Rating Scale (GARS; Gilliam, 1995) are among the most relevant and widely used. Although these well-established instruments for the screening and diagnosis of childhood autism made important contributions to the field, they are not without limitations. For example, some of these instruments might have a somewhat reduced sensitivity, especially among younger children (e.g., Corsello et al., 2007; Eaves et al., 2006b; Lord, Rutter, P.C. Dilavore, & Risi, 2002). Another drawback of most of the commonly used ASD assessment tools is that they require time-consuming training that is generally expensive and difficult to secure (Charman & Gotham, 2013).

Moreover, it appears that inter-rater reliability for the available "golden standard" instruments can be particularly low when younger children are assessed (Lord, Rutter, P.C. Dilavore, & Risi, 2002). This can be partially due to the data collection procedure that generally requires clinicians to observe, code and to interpret the behaviors simultaneously. The multitasking might result in errors at any of the three levels. There might also be some small variations between clinicians concerning the manner in which different specific tasks are applied, based on their expertise, which could also lead to different clinical judgments.

Given the role of an accurate diagnosis of ASD for selecting appropriate treatment for individuals and the criticality of early interventions, it is crucial that the data collected be as valid as possible. Therefore, there is a need for methodologies that produce a quantified characterization of the core symptoms in ASD during the diagnosis process. One way forward is to include machine-perception-guided technologies to augment the existing observational diagnoses and judgments made by clinicians

The use of clinicians-based instruments also limits the amount of collected data. It would be almost impossible for the clinician to assess different relevant outcomes in-session, while also delivering the intervention. Thus, an important source of information generally remains unexplored. The data gathered during sessions could provide important insights concerning the evolution of ASD symptoms throughout intervention. Moreover, it has the potential to contribute to the clarifications of involved mechanisms of change.

Therefore the accuracy of direct assessment – observational data has implications for the trustworthiness of the assessments obtained. Given the potentially transformative nature of the interventions programs developed and the data used to guide recommended changes in methods of improving these mechanisms that reinforce accurate data collection is imperative for this objective practice, it is crucial for the data to be as valid as possible.

Data capture and analysis is an important part of decision taking about whether a treatment is working or not. However, manual annotations are time consuming, thus limiting the amount of available data. Using different types of technological tools can help increase the amount of data collected, making it easier to collect, and helping professionals to quickly scan through data in order to take better informed decisions (Kientz, Hayes, Westeyn, Starner, & Abowd, 2007). Moreover, the resulting annotations might be of higher quality and more consistent than manual annotations.

The aim of this task was to evaluate the performance of the automatic diagnostic algorithms by comparing behavioral data obtained from a manually coding system with data extracted from the developed algorithms.

# Method

For the main outcomes we compared the interpretations made by the semi-autonomous system during the interventions session in the clinical trial (see D2.3) with the corrections made by the psychologist supervising its decisions. The psychologist based its decision on the direct recordings of the behavior of the child (i.e., the same images that are used as input by the system). The psychologist's input was considered as a reference and each decision of the algorithms was evaluated based on its matching with this input. These results have been presented and discussed extensively at Review 3.

Because several issues related to the set-up and the child behavior that were not mapped in the initial algorithms had a strong impact on the diagnostic performance, a new comparison between clinician's ratings and those made by the system has been made based on an offline analysis of the child behavior. The procedure and the results of this second analysis, which reflects more accurately the capabilities of the system, has been done for TT task and is presented in detail in D5.3.

# Data analysis

***Direct comparison between system's and clinician's judgements***

This comparison it is based on the first set of data, using the online decisions made by the algorithms and the operator corrections. We present here data for imitation (IM) and joint attention (JA).

The data for IM was based on 690 trials for which we had the system's assessment and the therapist judgement. The percentage of agreement between robot and therapist for each child varied between 23.64% and 78.69%. The overall correlation between robot and therapist ratings across participants was $r = .79$, $p = .039$. When looking at individual trials, the performance of matches between the therapist and the automatic algorithms drops to 48.84%. The percentage of correct assessments of trials made by the system when the children did perform a correct behavior, as judged by the therapist, was 20.41%. The percentage of correct assessments of trials made by the system when the children did not perform a correct behavior, as judged by the

therapist, was 70.51%. This pattern indicates that the system performs significantly better in identifying incorrect behaviors of the child, $\chi^2 = 177.42$, $p < .001$, than correct ones. The average agreement between the clinician and the automatic diagnostic algorithms was Cohen's Kappa = -.29 (SE = .03), $p < .001$. Figure 1 summarizes the number of correct and wrong targets identified by the automatic algorithms for IM.

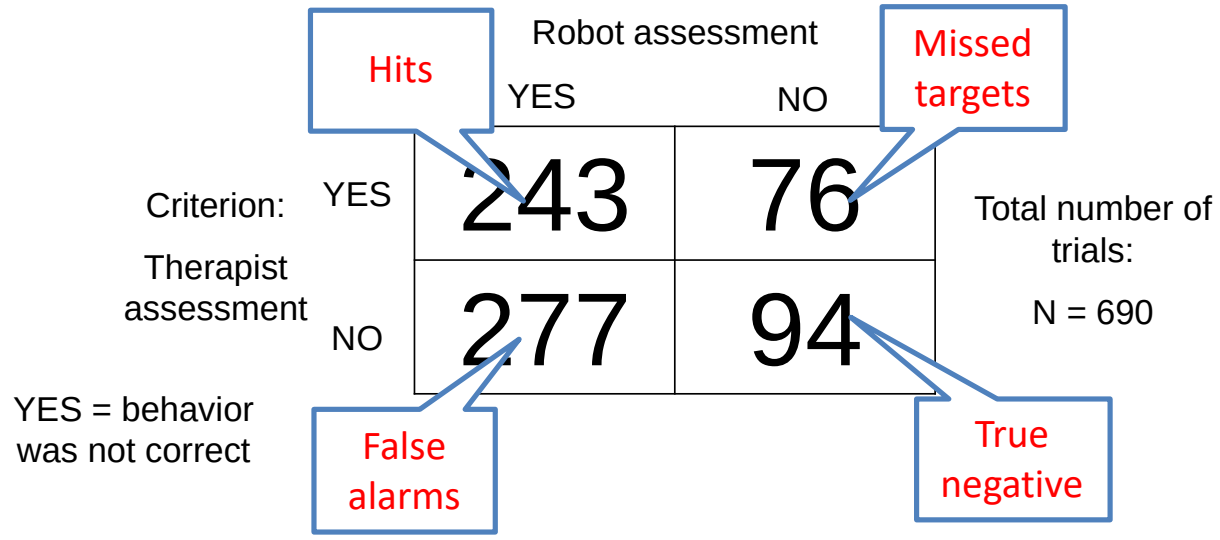


*Figure 1*. Number of correct and wrong targets identified by the robot as compared to the clinician's judgements for **IM**. *Note:* "Yes" means that a behavior was judged as being indicative of ASD symptoms.

The data for JA was based on 429 trials for which we had the robot's assessment and the therapist judgement. The percentage of agreement between robot and therapist for each child varied between 14.63% and 77.50%. The average level of agreement between the automatic algorithms and the clinician was 48.25%. The percentage of correct assessments of trials by the system when the children did perform a correct behavior, as judged by the therapist, was 20.40%. The percentage of correct assessments of trials by the system when the children did not perform a correct behavior, as judged by the therapist, was 89.35%. One again, the performance of the system is skewed towards identifying incorrect behaviors of the child, $\chi^2 = 179.82$, $p < .001$. The average agreement was Kappa = -.39 (SE = .03), $p < .001$. Figure 2 summarizes the number of correct and wrong targets identified by the automatic algorithms for JA.

***Direct comparison between system's and clinician's judgements***

The second of comparisons was made based on an offline analysis of the child behavior with improved algorithms that tried to overcome some of the limitations related to child behavior descriptions and specifications that were not taken into account in the initial phases. For example, the child might largely bend its body over the Sandtry without touching it, which

would be considered by the clinician that he or she is waiting his/her turn. However, this would have been interpreted by the system as a behavior indicating the lack of TT skills. In the offline analysis only the hands and wrists were considered, allowing the child's head and body to move over the sand tray without being interpreted as bad waiting. Also, a time filter was applied to the data in order to remove noise, coming from errors in detecting child's posture.

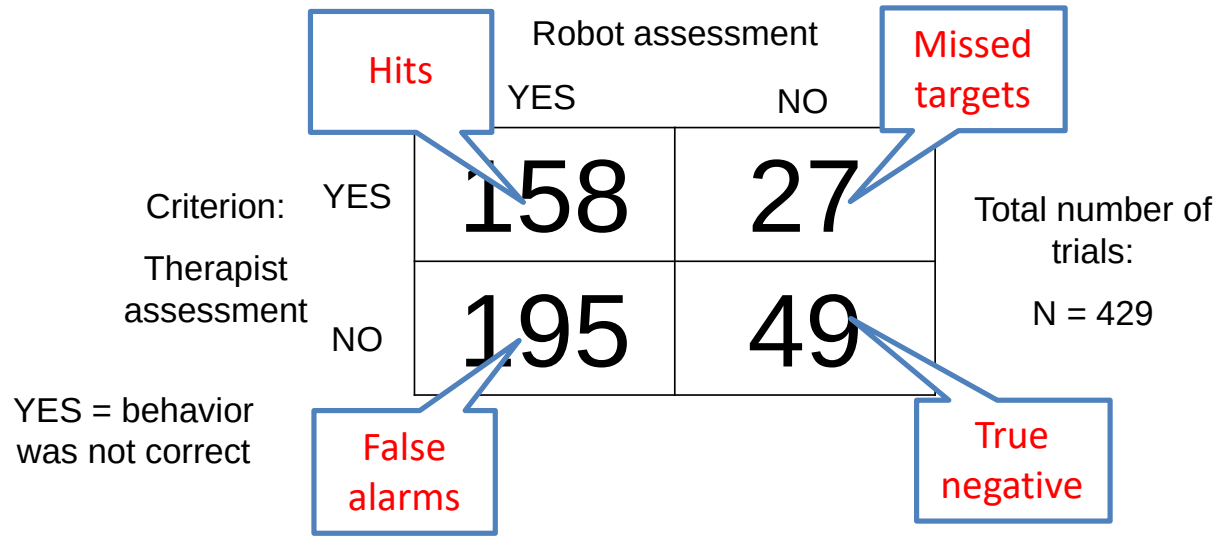


*Figure 2*. Number of correct and wrong targets identified by the robot as compared to the clinician's judgements for **JA**. *Note:* "Yes" means that a behavior was judged as being indicative of ASD symptoms.

The detailed results of the performance based on the offline analysis and the changes in the algorithms (and the rationale for these changes) are presented extensively in D5.3. The performance of the updated assessment methods presented in Figure 3, reached an average performance of 73%.

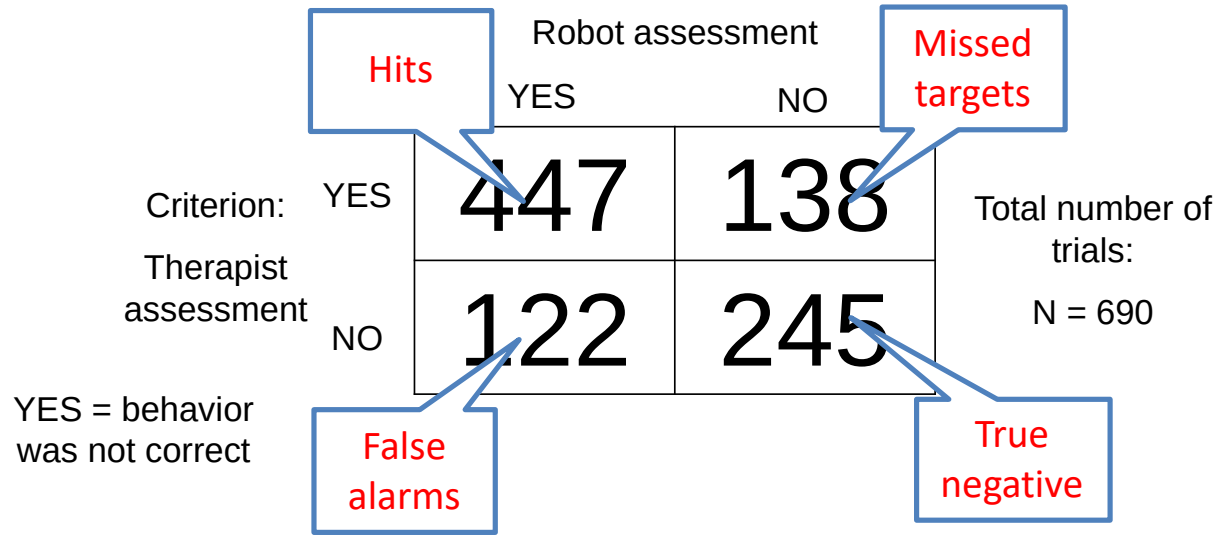


*Figure 3*. Number of correct and wrong targets identified by the robot as compared to the clinician's judgements for **TT**. *Note:* "Yes" means that a behavior was judged as being indicative of ASD symptoms; This data is based on the modified algorithms.

## Conclusion

These results point that automatic diagnosis of ASD symptoms is a challenging task even when using a complex setup to capture and analyze child behavior. Clinician's decisions in judging individual behaviors might depend on more subtle factors that are hard to operationalize into automatic algorithms, such as child progress in developing a particular skill, based on his or her previous performance (sometimes across sessions), or inferences about child's intentions. Moreover, the performance and stability of the algorithms in capturing and representing the behavior and comparing it with the expected pattern might also be improved. The work described in this deliverable and others associated to this one (D2.1, D5.1, and D5.) offer an example of an iterative process for gradually improving the performance of an automated diagnostic system by redefining relevant inputs and adjusting the algorithms to better capture the dynamics of the behaviors that are indicative of ASD symptoms. Despite the initial poor performance, the final results are actually comparable to what would be expected for two individual clinicians judging the same behaviors of child.

# References

Charman, T., & Gotham, K. (2013). Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders–lessons from research and practise. Child and adolescent mental health, 18(1), 52-63.

Corsello, C., Hus, V., Pickles, A., Risi, S., Cook, E. H., Leventhal, B. L., & Lord, C. (2007). Between a ROC and a hard place: decision making and making decisions about using the SCQ. Journal of Child Psychology and Psychiatry, 48(9), 932-940.

Eaves, L. C., Wingert, H., & Ho, H. H. (2006). Screening for autism: Agreement with diagnosis. Autism, 10(3), 229-242.

Gilliam, J. E. (1995).Gilliam autism rating scale: summary response form. Pro-ed

Kientz, J. A., Hayes, G. R., Westeyn, T. L., Starner, T., & Abowd, G. D. (2007). Pervasive computing and autism: Assisting caregivers of children with special needs.IEEE Pervasive Computing,6(1).

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C.,& Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of autism and developmental disorders,30(3), 205-223.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2002). Autism Diagnostic Observation Schedule (Western Psychological Services, Los Angeles)

Rutter, M., Le Couteur, A., & Lord, C. (2003). Autism diagnostic interview-revised.Los Angeles, CA: Western Psychological Services,29, 30.

Schopler, E., Reichler, R. J., DeVellis, R. F., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). Journal of autism and developmental disorders,10(1), 91-103.

Wing, L., Leekam, S. R., Libby, S. J., Gould, J., & Larcombe, M. (2002). The diagnostic interview for social and communication disorders: Background, inter- rater reliability and clinical use.Journal of Child Psychology and Psychiatry,43(3), 307-325.