An algorithm for deduplication and insurance of quality in a database for ASD diagnosis: proposal and qualitative evaluation

Sarah Klock Mauricio[1], Helena Brentani[2], Joana Portolese[2], Luciana Madanelo[2], Ariane Machado-Lima[1], Lima Fátima L. S. Nunes[1]

[1]Laboratory of Computer Applications for Health Care,
Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Arlindo Bettio street, 1000, São Paulo – SP, Brazil, 03828-000

[2]Institute of Psychiatry of the Hospital if Clinics, Hospital das Clínicas,
Faculdade de Medicina da Universidade de São Paulo,
Dr. Ovídio Pires de Campos street, 785, São Paulo - SP, Brazil, 05403-903

{sarahkm, ariane.machado, fatima.nunes}@usp.br,

{helena.brentani, joanaportolese, lucianamadanelo}@gmail.com

Abstract. *Autism Spectrum Disorder (ASD) diagnosis requires the action of well-trained health professionals, which limits access to diagnosis. Computer-aided diagnosis using biomarkers can be an alternative to make diagnosis more accessible. However, public databases to support the development of CAD systems are still a challenge. This study presents an algorithm to identify flaws in data quality in a database for ASD diagnosis, such as duplicate records and missing data, in the Research Electronic Data Capture platform. The tool automates error detection and generates structured reports to assist health professionals in correcting data. A qualitative evaluation confirmed its usefulness and indicates that the time to identify errors can decrease approximately 15 times, contributing to minimizing the effort necessary to maintain a consistent database.*

Resumo. *O diagnóstico do Transtorno do Espectro Autista (TEA) requer a atuação de profissionais de saúde altamente treinados, o que limita o acesso ao diagnóstico. O diagnóstico assistido por computadorutilizando biomarcadores pode ser uma alternativa para tornar o diagnóstico mais acessível. No entanto, a disponibilidade de bases de dados públicas para apoiar o desenvolvimento de sistemas CAD ainda é um desafio. Este estudo apresenta um algoritmo projetado para identificar falhas na qualidade dos dados em um banco de dados para diagnóstico de TEA, como registros duplicados e dados ausentes, na plataforma Research Electronic Data Capture. A ferramenta automatiza a detecção de erros e gera relatórios estruturados para auxiliar profissionais de saúde na correção dos dados. Uma avaliação qualitativa confirmou sua utilidade e indicou que o tempo necessário para identificar erros pode ser reduzido em cerca de 15 vezes, contribuindo para minimizar o esforço necessário para manter uma base de dados consistente.*

1. Problem definition

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by difficulties in communication and social interaction, as well as the presence of repetitive behaviors, affecting approximately 1% of the global population [Zeidan et al. 2022]. Early treatment of ASD significantly increases the likelihood of success [Mandell et al. 2005]. However, early diagnosis is also necessary to favor therapies. The two main diagnostic instruments currently considered the "gold standard", (Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS)), require extensive and costly training, resulting in limited accessibility to clinical diagnosis [Wall et al. 2012]. Therefore, more accessible methods to support ASD diagnosis are needed, resulting in the concept of Computer-Aided Diagnosis (CAD) systems. Among the existing CAD systems, research conducted by different teams focuses on classifications based on visual attention identification using eye-tracking (ET), and by analyzing anthropometric characteristics.

To address these issues, our team currently stores patient data on the Research Electronic Data Capture (REDCap) platform. However, this dataset presents challenges related to inconsistencies in stored data, such as duplicates and missing data. These inconsistencies are highly problematic. Missing data can cause biased results, and duplication may lead to inconsistencies and inefficient use of space. [Council et al. 2010].

To mitigate the data scarcity issue and facilitate the use of already collected data, a project is underway to study, improve, and unify existing databases used in previous research conducted our team. Among the stages developed, this article aims to present the an algorithm that detects inconsistencies and evaluates the data quality of the ASD diagnostic data within the REDCap platform. To verify the efficacy of this stage, a qualitative evaluation was conducted with health professionals.

2. Related work

Data quality management has been extensively studied. In the fields of public health information systems, Chen et al. [2014] reviewed the assessment methods and proposed a framework based on three dimensions: data, data use, and data collection process. Their findings highlight completeness, accuracy, and timeliness as the attributes most evaluated, with surveys and audits as key quantitative methods, and interviews and documentation reviews as qualitative approaches. However, they identified gaps, including inconsistent attribute definitions, limited attention to data use and collection processes, and the absence of systematic assessment procedures.

Regarding databases deduplication, which was used in this work along with the assessment of data quality efforts, Bhatt et al. [2013] use demographic data, alongside with biometric characteristics related to fingerprints, to perform deduplication through algorithms that calculate string distance for demographic data and Support Vector Machine (SVM) for biometric data. The results indicate that this methodology is only useful when demographic data are genuine and that there was difficulty in handling the fact that demographic data are updated and change over time as new profiles are created. With a different approach, Kaushik et al. [2012] developed an algorithm that reduces names based on phonetic rules to facilitate the deduplication of names that are identical but written differently; their work an average precision of 94%.

The present study offers a contribution by focusing on structured quality evaluation for ASD diagnosis data. Unlike traditional data quality models, our approach aligns with specific requirements of clinical datasets.

3. Material and methods

The algorithm was built using Python [Van Rossum and Drake Jr 1995] programming language, specifically with the PyCap [Burns et al. 2014] library to import REDCap data via an Application Programming Interface (API), Pandas [The pandas development team 2020] for data manipulation, and difflib [Python Software Foundation 2025] to assist in the deduplication process by calculating string distances. The final output of the algorithm is an spreadsheet file, obtained by running the Python code file in the terminal. The use of the data processed by the algorithm was approved by the Ethics Committees of the associated institutions [Pinheiro 2018].

A study with multiple interviews with healthcare professionals was conducted to determine which forms and fields would be used. From this process, we determine the following forms that should be analyzed to eliminate duplication and ensure data quality: **Data** and **Mothers's Life Form** - used to collect general demographic and medical history information about patients and their mothers, with **Data** being used in this study to identify the patients and their duplicates, rather than being analyzed - **Psychiatric Interview**, **Autistic Behavior Checklist (ABC)**, **Modified Checklist for Autism in Toddlers (M-CHAT)**, **Social Communication Questionnaire (SCQ)**, **Childhood Autism Rating Scale (CARS)**, **Autism Diagnostic Observation Schedule (ADOS)**, **Eye-Tracking Data**, **Anthropometric Data**, **Child Behavior Checklist (CBCL)**, **Vineland Adaptive Behavior Scales (Vineland)**, **Social Responsiveness Scale (SRS)**, **Wechsler Abbreviated Scale of Intelligence (WASI)**, **Wechsler Preschool and Primary Scale of Intelligence (WPPSI)**, **Wechsler Intelligence Scale for Children (WISC)** and **SON-R (Snijders-Oomen Nonverbal Intelligence Test - Revised)**, all used for the ASD diagnosis process.

The dataset contains various types of forms, some purely textual, while others generate scores within REDCap or through external tools, which must then be manually entered. Additionally, some instruments have multiple versions that vary based on patient characteristics. All forms include an automatically assigned status ('Complete', 'Incomplete', or 'Unverified'), but this can be manually altered, leading to inconsistencies. Errors in 'Complete' records can be detected by comparing the form's status with its corresponding score form. In this study, problematic records include those marked as 'Incomplete', 'Unverified', or those where form and score statuses diverge, as they are unsuitable for research. Unfilled forms, however, are not considered missing data, as not all forms are required in every case.

The first step of the algorithm is to import the data using the PyCap library. The relevant columns are processed, unifying form entries based on the patient's ID, adjusting data types, and calculating the patient's age based on their date of birth. Finally, the coded values in certain columns are replaced with their textual representations, defined by the REDCap system. Once the basic processing is complete, the code identifies duplicate entries for the same patient. This function compares full names using the string distance function implemented in the Difflib library, along with the exact comparison of the dates

of birth. If the birth dates of two records match, the function verifies if their name similarity exceeds 0.9, and, if so, they are referenced as duplicates in an additional column. The name and date of birth columns are removed from the final spreadsheet to prevent data privacy issues. To improve the performance of this function, considering that it compares all registers in an initially exponential complexity, the concept of index-based comparison was applied, making the complexity linear.

The main errors related to the 'Complete'/'Incomplete'/'Unverified' status are identified and colored in the final sheets accordingly with their severity, especially in cases where an instrument has one form marked as 'Complete' for the questions, but the other form for scoring is marked as 'Incomplete', 'Unverified', or left unfilled.

Regarding the existence of multiple versions of a single instrument, in the main sheet, their values are integrated as if they were only one, so that if one version is marked as 'Complete' or filled, that value is selected to appear. To preserve detailed information, the original versions of the columns, along with specific data related to the instruments according to their particularities, is available in additional sheets in the final spreadsheet, with one sheet per instrument. Returning to the main page, the algorithm classifies patient diagnoses based on numeric scoring rules from specific scores, such as SCQ, M-CHAT, CARS and ABC, inserting the results into new columns.

The assessment of the final product involved a qualitative analysis conducted with the target users - two health professional experts in ASD diagnosis who work with the REDCap platform. An interview was conducted in two separate online sessions with two psychologists. Both professionals have post-graduate degrees, with 12 and 27 years of experience in ASD diagnosis. The participants were presented with a project summary, followed by a demonstration of how to use the spreadsheet generated by the algorithm. Then, five questions were presented, whose main discussions are presented in Section 4.2.

4. Results and discussion

4.1. Data Quality Statistics

Analyzing the algorithm's outputs allowed the extraction of some interesting statistics about the data quality of the database. Disregarding unfilled records and considering records whose status is not 'Complete' as problematic records, a percentage of records presenting any inconsistency in any column was calculated for each instrument. That is because it is expected that some forms would not be filled in all cases, and those marked as 'Complete' do not indicate any inconsistency. The results demonstrate that the current mean percentage of problems in the instruments is approximately 52.4%. To enable a more detailed examination, the plot in Figure 1 presents the percentage of problematic records for each form ordered in descendant order in the black line, again disconsidering unfilled records. In the same plot, the stacked bars show the raw number of records divided by their status, allowing the visualization of the overall quality of each instrument and the main cause of problematic records.

Examining the instruments with the highest percentages of problematic records, many records are marked as complete while their scores remain incomplete or unfilled, as seen in the stacked bars of Figure 1. Additionally, certain mandatory forms, filled

out online by patients' guardians before professional follow-up, are more prone to errors due to the lack of guidance on the usage of the platform. Further analysis indicated that 9.14% of the database is composed by duplicates, and 1.26%, by patients without a name or any information that would allow identification, which may or may not be duplicates of already registered patients.

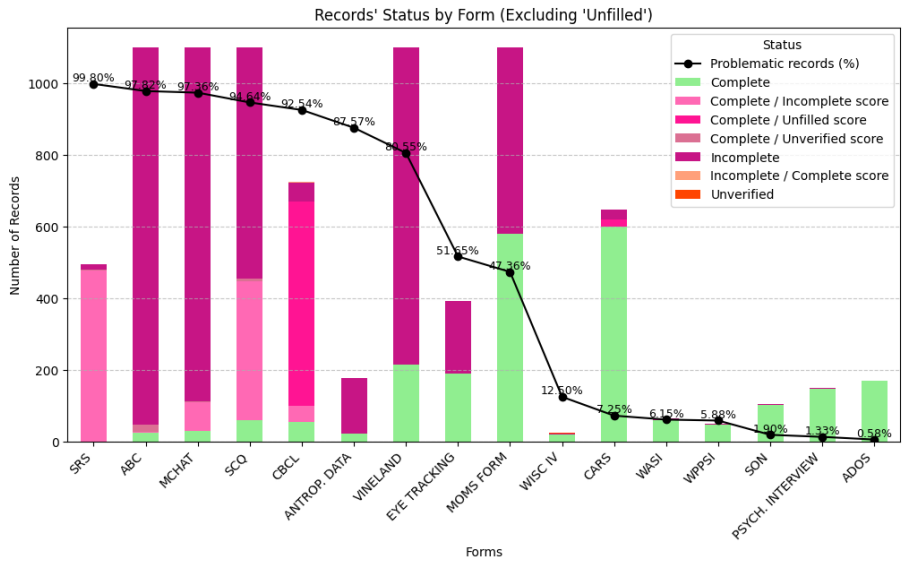

Figure 1. Amount and status of the records and a line describing the percentage of problematic records for each form.

4.2. Qualitative evaluation and algorithm performance

As mentioned in Section 3, an interview was conducted with two professionals, U1 and U2, to assess the tool's usefulness in hospital routines. Both confirmed that no existing application provides similar functionalities, with U1 previously creating manual spreadsheets that required days to complete. U2 noted that while REDCap allows data downloads, retrieving files is time-consuming and prone to failures. The time estimations for key tasks indicate that the algorithm-generated spreadsheet significantly reduces workload, cutting routine activity times by at least 15 times. Both participants highlighted the tool's benefits in improving data maintenance, streamlining access to patient records, and supporting training in REDCap usage. They rated the tool's utility and quality with the highest score (10/10), emphasizing its value in managing multiple records and tracking patient cases. Additionally, they suggested integrating real-time alerts to notify professionals about errors and expanding the output with additional columns and improved visualization methods for related scales.

Considering the current amount of data in the dataset, which includes approximately 1130 registers, the overall algorithm takes around 30 seconds to finish its execution, but the processing itself only takes 5 seconds. Most of its performance time is dedicated to formatting the final spreadsheet with colors according to the cells' values.

5. Conclusion

This study developed an algorithm to improve the quality of ASD diagnostic data stored in the REDCap platform to aid ASD diagnosis. By identifying duplicate records, miss-

ing data, and improving the readability of the information, the tool supports professionals in maintaining high-quality datasets, which is essential for reliable diagnostic support and research in the field, contributing to the broader goal of enhancing ASD diagnostic methods. These benefits were confirmed in a qualitative evaluation with experts, which demonstrated a significant reduction in effort and time to identify inconsistencies. This study has certain limitations, such as the algorithm's output being static, requiring periodic execution to update reports, which may delay error detection and correction. Future improvements could integrate real-time monitoring with automated alerts.

Acknowledgments

This work was supported in part by the São Paulo Research Foundation (FAPESP) through a undergraduate research scholarship (process 2024/21470-0), the National Council for Scientific and Technological Development (CNPq) (processes 2024-46 and 2024-2953), and the Office of the USP's Vice-Provost for Research and Innovation.

References

Bhatt, H. S., Singh, R., and Vatsa, M. (2013). Can combining demographics and biometrics improve de-duplication performance? In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 188–193.

Burns, S. S., Browne, A., Davis, G. N., Rimrodt, S. L., and Cutting, L. E. (2014). Pycap (version 1.0) [computer software].

Chen, H., Hailey, D., Wang, N., and Yu, P. (2014). A review of data quality assessment methods for public health information systems. *International Journal of Environmental Research and Public Health*, 11(5):5170–5207.

Council, N. R., of Behavioral, D., Sciences, S., Education, on National Statistics, C., and on Handling Missing Data in Clinical Trials, P. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press.

Kaushik, V. D., Bendale, A., Nigam, A., and Gupta, P. (2012). An efficient algorithm for de-duplication of demographic data. In Huang, D.-S., Jiang, C., Bevilacqua, V., and Figueroa, J. C., editors, *Intelligent Computing Technology*, pages 602–609, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mandell, D. S., Novak, M. M., and Zubritsky, C. D. (2005). Factors associated with age of diagnosis among children with autism spectrum disorders. 116.

Pinheiro, T. D. (2018). Classificação de imagens faciais para o auxílio ao diagnóstico do transtorno do espectro autista.

Python Software Foundation (2025). difflib — helpers for computing deltas. Accessed: 28-02-2025.

The pandas development team (2020). pandas-dev/pandas: Pandas.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., and DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One*.

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., and Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism Res*, 15:778–790.