

BECOME Dataset and Profiling Methodology

(Real-World Data Infrastructure for Early Identification of Developmental Vulnerabilities)

The BECOME dataset is a longitudinal real-world data infrastructure established within the Clinical Psychology Unit of the Maternal and Child Health Area at Policlinico Universitario A. Gemelli IRCCS in Rome, Italy. Its primary purpose is to facilitate the systematic analysis of cognitive, behavioral, and learning trajectories in children and adolescents undergoing neuropsychological assessment as part of routine clinical care.

This dataset is derived from standardized clinical evaluations conducted between 2019 and 2024 and incorporates structured information obtained from the hospital's electronic health record system (TrackCare).

Methodology

Study Population

The initial dataset comprised 833 patients, with a mean age of approximately 10 years. Following the exclusion of cases with only a single available report, the final analytical sample consisted of 766 patients who had undergone at least two assessments, corresponding to a total of 3,308 clinical reports and an average follow-up duration of 3.2 years.

Results

The age group most prominently represented at the initial assessment was 5 to 10 years, with a higher prevalence of males compared to females.

Assessment Framework

Participants engaged in comprehensive standardized assessments, which included:

- Cognitive evaluations (Griffiths, WPPSI, WISC, WAIS, Leiter scales)
- Behavioral assessments utilizing the Child Behavior Checklist (CBCL)
- Academic achievement evaluations encompassing reading, writing, and mathematics
- Additional neuropsychological measures when clinically indicated (e.g., visual-motor integration, attention tasks)

All evaluations adhered to standardized clinical protocols and were documented through structured clinical reports that were integrated into the electronic health record system.

Data Infrastructure and Extraction

Data extraction and curation were conducted in collaboration with Gemelli Generator Real World Data, a multidisciplinary clinical data science hub. The data pipeline encompassed:

- Extraction, transformation, and loading (ETL) from hospital systems
- Integration of both structured clinical data and unstructured reports
- Implementation of natural language processing (NLP) algorithms for the extraction of diagnostic information
- Anonymization and adherence to GDPR-compliant data processing

The resultant harmonized dataset serves as a longitudinal repository of cognitive, behavioral, and academic data, thereby enabling the investigation of developmental trajectories in real-world clinical environments.

Diagnostic Trajectories

Longitudinal analyses have identified four principal diagnostic trajectories:

- Persistent diagnoses: 34% of patients
- Remission: 26%
- Emerging diagnoses: 14%
- No diagnosis across assessments: 26%

Among children who presented without a diagnosis at baseline, 35% received a diagnosis during subsequent follow-up assessments, indicating that certain developmental conditions may manifest later in childhood.

Cognitive Profile Findings

Overall, cognitive performance predominantly fell within the normative range; however, the dataset illuminated selective vulnerabilities in executive domains, specifically:

- Working memory
- Processing speed

Conversely, verbal and fluid reasoning abilities appeared to be relatively intact. These findings align with the high prevalence of neurodevelopmental conditions characterized by attentional and executive components within the sample.

Value of the BECOME Dataset

The BECOME dataset provides a structured framework that facilitates:

- The early identification of developmental vulnerabilities
- Longitudinal monitoring of cognitive and behavioral trajectories
- Risk pattern profiling in children and adolescents
- The formulation of stratified prevention and intervention strategies

In the context of behavioral prevention research, this infrastructure supports data-driven profiling and aids in identifying subgroups that may derive benefit from targeted preventive interventions.