Libyan Journal of Educational Research and E-Learning (LJERE)
ISSN: 3105-3556
Volume 2, Issue 2, 2026
Page No: 176-195
Website: https://ljere.com.ly/index.php/ljere/index

A Reproducible Rule-Execution and Sensitivity-Analysis Workflow for Synthetic Child-Day Records: A Software-Verification Study

Mohamed Al-Issawi Kajman[1*], Alfaytouri Alsabiri [2], Abdulrahman M. Zargoun [3], Ali Mahmoud Assabri [4]

[1,3] Department of Computing and Information Technology, College of Technical Sciences, Bani Walid, Libya

[2]Department of Computer Eng, High institute of industrial technology, Tripoli, Libya

[4] Department of Engineering Management, College of Technical Sciences, Bani Walid, Libya

منهجية قابلة لإعادة التطبيق لتنفيذ القواعد وتحليل الحساسية لسجلات يومية اصطناعية للأطفال: دراسة في التحقق البرمجي

محمد العيساوي كجمان[1*]، الفيتوري الصابري[2]، عبدالرحمن زرقون [3]، علي محمود الصابري

[1،3] قسم الحاسوب وتقنية المعلومات، كلية العلوم التقنية، بني وليد، ليبيا

[2] قسم هندسة الحاسوب، المعهد العالي للتقنيات الصناعية، طرابلس، ليبيا.

[4] قسم الإدارة الهندسية، كلية العلوم التقنية، بني وليد، ليبيا.

[*]Corresponding author: alimahmoudassabri@gmail.com

Received: 28-04-2026	**Accepted:** 17-07-2026	**Published:** 01-08-2026

Abstract

Families and clinical teams may record daily observations about sleep, environmental stimulation, routine changes, activity, and behavior when supporting children with autism spectrum disorder (ASD). Such records can be difficult to structure and interpret, particularly when observations are incomplete or defined differently across days. A transparent rule-based workflow can serve as a testable engineering approach to organizing these observations, but synthetic demonstrations must not be presented as clinical evidence.

This study aims to develop and evaluate a reproducible rule-based workflow for organizing and analyzing synthetic daily records of children with autism spectrum disorder (ASD). A simulated dataset was generated for three hypothetical children over 90 consecutive days per child, resulting in 270 child-day records, using an explicit generation model based on predefined rules and behavioral and environmental variables.

The system utilized nine rules to calculate a risk score and performed verification and sensitivity analyses, including permutation testing, rule-ablation analysis, and Monte Carlo simulations. The results showed a descriptive association between the rule-based score and the simulated severity level (**r = 0.3687**), with relatively consistent findings across multiple simulation runs.

The study demonstrates the feasibility of using interpretable, rule-based methodologies with software-verification procedures for organizing synthetic behavioral data. However, these findings should not be interpreted as clinical evidence or predictive performance. Future

applications require validation using real de-identified longitudinal data and evaluation by clinical experts.

Keywords: autism spectrum disorder; behavioral observation; rule-based decision support; synthetic data; software verification; digital health; explainable artificial intelligence; reproducible research.

الملخص

قد تقوم الأسر والفرق السريرية بتسجيل ملاحظات يومية حول النوم، والتحفيز البيئي، والتغيرات في الروتين، والنشاط، والسلوك عند دعم الأطفال المصابين باضطراب طيف التوحد (ASD). وقد يكون من الصعب تنظيم هذه السجلات وتفسيرها، خاصة عندما تكون الملاحظات غير مكتملة أو يتم تعريفها بطرق مختلفة عبر الأيام. ويمكن لمنهجية شفافة قائمة على القواعد أن تمثل نهجًا هندسيًا قابلًا للاختبار لتنظيم هذه الملاحظات، إلا أن العروض التوضيحية المعتمدة على البيانات الاصطناعية يجب ألا تُقدَّم باعتبارها أدلة سريرية.

تهدف هذه الدراسة إلى تطوير وتقييم سير عمل قابل لإعادة الإنتاج قائم على القواعد لتنظيم وتحليل سجلات يومية اصطناعية للأطفال المصابين باضطراب طيف التوحد .(ASD) تم إنشاء قاعدة بيانات محاكاة شملت ثلاثة أطفال افتراضيين لمدة 90 يومًا لكل طفل (270 سجلًا)، باستخدام نموذج توليد واضح يعتمد على قواعد محددة ومتغيرات سلوكية وبيئية.

اعتمد النظام على تسع قواعد لحساب درجة الخطر، مع إجراء اختبارات للتحقق من اتساق التنفيذ وتحليل حساسية النتائج، شملت اختبارات التبديل العشوائي، وإزالة القواعد، وتحليل مونت كارلو. أظهرت النتائج وجود ارتباط وصفي بين درجة القواعد وشدة الحالة في البيانات الاصطناعية(**r = 0.3687**) ، مع ثبات نسبي للنتائج عبر عمليات المحاكاة المتعددة.

توضح الدراسة إمكانية استخدام منهجيات قائمة على القواعد قابلة للتفسير والتحقق البرمجي لتنظيم البيانات السلوكية الاصطناعية، مع التأكيد على أن النتائج لا تمثل دليلًا سريريًا أو قدرة تنبؤية، وأن تطبيقها مستقبلاً يتطلب التحقق باستخدام بيانات حقيقية مجهولة الهوية وتقييمات سريرية متخصصة.

الكلمات المفتاحية : اضطراب طيف التوحد؛ دعم القرار القائم على القواعد؛ البيانات الاصطناعية؛ التحقق البرمجي؛ الصحة الرقمية؛ الذكاء الاصطناعي القابل للتفسير.

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent differences in social communication and social interaction, together with restricted or repetitive patterns of behavior, interests, or activities [1]. Its presentation is heterogeneous. Sleep disturbance, sensory reactivity, anxiety, hyperactivity, self-injury, aggression, elopement, and severe distress may occur as associated difficulties, although they are not themselves the defining diagnostic criteria [2,3].

Behavioral distress is also heterogeneous in time and context. Sleep disruption, pain or illness, communication difficulty, sensory load, changes in routine, task demands, and the social environment may interact differently across children and across days. Clinical guidance therefore emphasizes that an acute change in behavior can signal an underlying medical or emotional problem and should not be treated as a purely behavioral event without appropriate assessment [4]. The same guidance recommends a low-stimulus environment, the involvement of familiar carers, and individualized use of familiar sensory or musical stimuli when helpful [4].

Families may record observations in diaries, mobile applications, or informal notes. These records can contain useful information about sleep duration, sleep quality, screen exposure, noise, activity, antecedents, behavior type, duration, severity, and responses to an intervention.

Yet the variables may be incompletely defined, measured with different sources, or recorded at irregular intervals. In principle, a transparent computational workflow could organize these observations and expose the conditions leading to a prompt. That hypothesis remains untested in the present study because no real child or clinical record was used.

Research on autism and digital health has examined ecological assessment, wearable sensing, and multimodal monitoring [5–7] [11]. These approaches illustrate the potential value of repeated observations in natural settings, but they also highlight methodological risks. A model may appear to perform well when repeated records from a small number of people are treated as independent, when the simulated outcome is generated from the same rule thresholds used for evaluation, or when a descriptive association is interpreted as evidence of clinical validity. Reporting guidance for prediction models stresses complete definitions of the data, predictors, outcomes, evaluation procedures, calibration, fairness, and reproducibility [9].

Much digital-health research emphasizes model performance, whereas the reliability of the rule engine that transforms observations into prompts is often treated as an implementation detail. Unit tests can verify individual code paths, but they do not by themselves document how a complete synthetic benchmark, rule registry, score aggregation, sensitivity analysis, and reproducibility audit fit together. The present study addresses this methodological gap by treating rule-engine verification and implementation sensitivity as explicit research outputs.

The present study is intentionally narrower. It documents and audits a rule-based workflow using an explicit synthetic benchmark. The central contribution is not a clinical prediction model. Rather, it is a reproducible software-verification release that distinguishes rule execution from model performance and makes the assumptions behind the generated outcomes inspectable.

2. Research problem and objectives

The practical problem is the difficulty of organizing fragmented daily observations into a transparent structure that can be reviewed by a caregiver or professional. The computational problem is to implement an interpretable rule engine without implying that its numerical output is a calibrated risk probability. A child-specific workflow must preserve the longitudinal structure of observations while making its logic visible and allowing a human reviewer to reject or override a prompt.

The study had five objectives. First, it specified a structured child-day representation for contextual and behavioral observations. Second, it implemented descriptive summaries and candidate association-rule calculations. Third, it represented provisional knowledge through explicit IF–THEN rules. Fourth, it implemented a rule-score display and non-scoring prompt labels with deterministic tests. Fifth, it examined the stability of these software outputs under independent seeds, subgroup summaries, a simple held-out-child baseline, rule-sensitivity scenarios, a within-child negative control, and pre-specified rule ablations.

3. Intended use, definitions, and scope

In this manuscript, a **synthetic record** is a computer-generated record created from prespecified assumptions rather than a record obtained from a child. **Behavioral escalation** is a generator-defined increase in the simulated intensity or safety relevance of a behavioral episode relative to the simulated daily record; it is not a validated clinical endpoint. An **expert rule** is a transparent conditional statement that maps an input pattern to a score increment or a prompt label. **Risk** score is used only as a software term for the output of the rule engine. It is not a calibrated probability and does not represent a probability of harm.

The benchmark represents three hypothetical children aged 4, 8, and 11 years, with 90 consecutive days per child. The age values are design parameters, not evidence of

generalizability to children in those age groups. The workflow is intended for technical inspection and future study design. It does not diagnose ASD, classify clinical severity, prescribe medication, authorize restrictive intervention, or replace clinical assessment.

4. Related literature

4.1. Sleep, behavior, and contextual assessment

Sleep difficulties are common in children with ASD, but their type, severity, measurement source, and relationship with daytime behavior vary across individuals. Reviews have described associations between insufficient or disrupted sleep and daytime difficulties, while also emphasizing that the evidence is heterogeneous and that sleep–behavior relationships may be bidirectional [2,3]. The present benchmark therefore uses sleep-related variables as design inputs, not as estimates of a population effect.

Ecological momentary assessment (EMA) and related diary methods can collect repeated reports in natural settings with less reliance on long-term retrospective recall. A systematic review identified 32 unique autism EMA studies involving 930 autistic participants, but found that the existing evidence was concentrated among adults with average or above-average intelligence and language skills; evidence for children and for a wider range of communication and intellectual profiles was less developed [5]. This distinction is relevant to the present work because a synthetic daily record cannot substitute for a real-world measurement protocol.

4.2. Open-science and wearable resources

The Simons Sleep Project illustrates how future validation could combine home-based wearable and nearable recordings with parent questionnaires and daily sleep diaries. The resource includes data from 102 autistic children and 98 non-autistic siblings aged 10–17 years and more than 3,600 nights of recordings [6,7]. It is relevant for measurement calibration, but it does not automatically provide the same age range, episode definitions, contextual variables, or intervention outcomes used in this benchmark.

Recent reviews of wearable prediction of severe behavior problems have also emphasized methodological limitations in existing studies, including uncertainty around claims of forecasting behavior several seconds before onset [11]. These observations reinforce the need to distinguish sensor-based prediction, descriptive monitoring, rule execution, and clinical benefit.

4.3. Reporting and evaluation guidance

TRIPOD+AI provides reporting guidance for studies that develop or evaluate prediction models using regression or machine-learning methods. It emphasizes complete reporting of data sources, predictors, outcomes, performance, calibration, fairness, and reproducibility [8]. The current study is not a clinical prediction-model development or validation study, but several of these reporting principles remain useful for describing the synthetic benchmark.

The manuscript previously referred to TRIPOD-Cluster as relevant because daily observations are nested within children. That interpretation has been corrected. TRIPOD-Cluster was developed primarily for prediction studies using clustered data sources such as multicentre, registry, or individual-participant-data datasets; the published guidance explicitly states that it is not intended for clustering defined by repeated measurements within the same individuals [9]. Repeated daily observations in a future human study will require longitudinal analyses that account for within-child dependence.

DECIDE-AI addresses early-stage live clinical evaluation of AI-based decision-support systems. It highlights safety, human factors, dataset shift, generalizability, and the gap between in-silico performance and benefit to patient care [10]. The present work remains preclinical and synthetic; DECIDE-AI is cited as a future-stage framework rather than as a claim of compliance or clinical evaluation.

4.4. Synthetic-data utility

Synthetic data can support software development and methodological testing, but its usefulness depends on the purpose for which it is created. Synthetic-data methodology recommends evaluating statistical fidelity, utility for the intended task, and disclosure or privacy risks when a real reference dataset is available [12]. No real reference dataset was used here. Accordingly, the benchmark is a transparent stress-test fixture, not a validated representation of children with ASD.

5. Methods

5.1. Study design

This was an in-silico, synthetic-data proof-of-concept study. No human participant was recruited, no clinical record was accessed, no biological specimen was used, and no intervention was delivered. The study was designed to test whether a documented workflow could ingest child-day records, apply transparent rules, produce descriptive summaries, and reproduce expected outputs under fixed software tests.

No human-population power calculation was performed because the study makes no inferential claim about a target population. The appropriate sample-size rationale for a future human study should be based on the intended prediction horizon, outcome frequency, participant-level split, model complexity, missingness, and clinically relevant performance margin.

5.2. Synthetic benchmark and explicit data-generating process

The benchmark contains 270 records: three hypothetical children and 90 consecutive days per child. Each child has a fixed age and a child-specific effect. Sleep, noise, and activity are updated sequentially with bounded autoregressive-like equations. Routine change and sensory overload are sampled from logistic mechanisms with limited dependence on the preceding day. Let S_t, N_t, and A_t denote sleep hours, noise level, and activity level on day t, respectively. With child-specific means $\mu S, \mu N$, and μA, the updates are:

Equation 1. $S_t = clip[\mu S + 0.68(S_{t-1} - \mu S) + \varepsilon S, t, 4.0, 9.6], where \varepsilon S, t\ Normal(0, 0.85).$
Equation 2. $N_t = clip[\mu N + 0.58(N_{t-1} - \mu N) + \varepsilon N, t, 0, 10], where \varepsilon N, t\ Normal(0, 1.15).$
Equation 3. $A_t = clip[\mu A + 0.45(A_{t-1} - \mu A) + \varepsilon A, t, 0, 10], where \varepsilon A, t\ Normal(0, 1.45).$
Sleep quality is an ordinal value obtained by rounding and bounding $2.2 + 0.48(S_t - 5) + Normal(0, 0.75)$to the interval 1–5. Screen time is sampled from a bounded gamma-based mechanism: $T_t = clip[Gamma(2.2, 1.45) + 0.15(A_t - 5), 0, 10].$
Let $R_{t-1}$and X_{t-1}indicate routine change and sensory overload on the preceding day. The probabilities of routine change and sensory overload are:
Equation 4. $logit[P(R_t = 1)] = -1.05 + 0.48R_{t-1} + 0.10I(child = C02).$
Equation 5. $logit[P(X_t = 1)] = -1.20 + 0.22(N_t - 5) + 0.72X_{t-1} + 0.20I(child = C02).$
The latent severity mechanism is explicit in the release. It does not use RiskRaw or any rule-engine output:
Equation 6. $Z_t = 2.55 + 0.60[0.34(8 - S_t) + 0.17N_t + 0.55R_t + 0.78X_t + 0.075A_t + 0.025T_t + b_i + 0.24R_{t-1}] + \varepsilon Z, t, where \varepsilon Z, t\ Normal(0, 1.00), and Severity_t = clip(Z_t, 0, 8.5).$
The child effects are $bC01 = 0.00, bC02 = 0.45, and bC03 = -0.30.$ An episode indicator is drawn conditionally on the generated severity after the severity value is computed, using:
Equation 7. $p_{episo}d_e, t = clip(0.08 + 0.075 \times Severity_t, 0.02, 0.90).$

Behavior type is sampled from a separate categorical mechanism. For episode records, duration, trigger, intervention label, and resolution code are sampled with prespecified

software-test distributions. The intervention probabilities are not clinical effectiveness estimates. No real reference dataset was used to calibrate any parameter.

5.3. Data dictionary

Table 1 provides the data dictionary used in the release. The complete parameter file and generator are included with the reproducibility package.

Table 1. Data dictionary for the synthetic child-day benchmark**.**

Variable	Type	Range	Unit/scale	Description
ChildID	Categorical	C01–C03	Identifier	Hypothetical child identifier.
Age	Integer	4–11	Years	Fixed age assigned to each profile.
Day	Integer	1–90	Day index	Consecutive simulated day.
SleepHours	Continuous	4.00–9.60	Hours	Simulated sleep duration.
SleepQuality	Ordinal	1–5	Ordinal scale	Simulated sleep-quality rating.
NoiseLevel	Continuous	0–10	Index	Simulated environmental noise exposure.
ActivityLevel	Continuous	0–10	Index	Simulated daily activity level.
ScreenTime	Continuous	0–10	Bounded index	Simulated screen exposure index.
RoutineChange	Binary	False/True	Indicator	Whether a routine change was simulated.
SensoryOverload	Binary	False/True	Indicator	Whether sensory overload was simulated.
BehaviorType	Categorical	NoEpisode plus seven labels	Label	Simulation-coded episode type; NoEpisode is a valid category and not a diagnosis.
Severity	Continuous	0–8.5	Simulation scale	Generator-defined severity outcome.
DurationMin	Integer	0–80	Minutes	Simulation-coded episode duration.
Trigger	Categorical	NoTrigger plus six labels	Label	Simulation-coded antecedent label; NoTrigger is a valid category.
RiskRaw	Integer	0–220 theoretical	Points	Sum of active rule weights; not a probability.
Risk100	Continuous	0–100 theoretical	Display scale	100 × RiskRaw / 220; display convention only.

Variable	Type	Range	Unit/scale	Description
FiredRiskRules	String	R01–R09	Rule IDs	Semicolon-separated active score-generating rules.
Prompts	String	Prompt label or blank	Label	Non-scoring simulation prompt label.
Intervention	Categorical	Ten labels or NoIntervention	Label	Software-test label retained in the technical appendix; NoIntervention is a valid category.
ResolutionCode	Binary/NA	0/1 or NA	Indicator	Software-test resolution code; not treatment efficacy.

The labels NoEpisode, NoTrigger, and NoIntervention are valid categorical values, not missing observations. An empty ResolutionCode denotes structural non-applicability when no intervention label is present. Analyses that import the CSV into software with automatic missing-value conversion should preserve these literal labels; the supplied pipeline does so explicitly. Missing values introduced into rule inputs are handled conservatively: a rule requiring an unavailable input does not fire automatically and should be reviewed before any score is displayed.

5.4. Five-layer architecture

The workflow contains five conceptual layers (Figure 1). The data layer stores structured child-day records. The analysis layer computes descriptive summaries, data-quality checks, and correlations. The mining layer produces candidate association-rule summaries. The knowledge base contains **clinician-reviewable**, not clinician-validated, IF–THEN rules. The decision layer computes the rule-score display and exposes the rules that contributed to it. Human review remains outside the automated score.

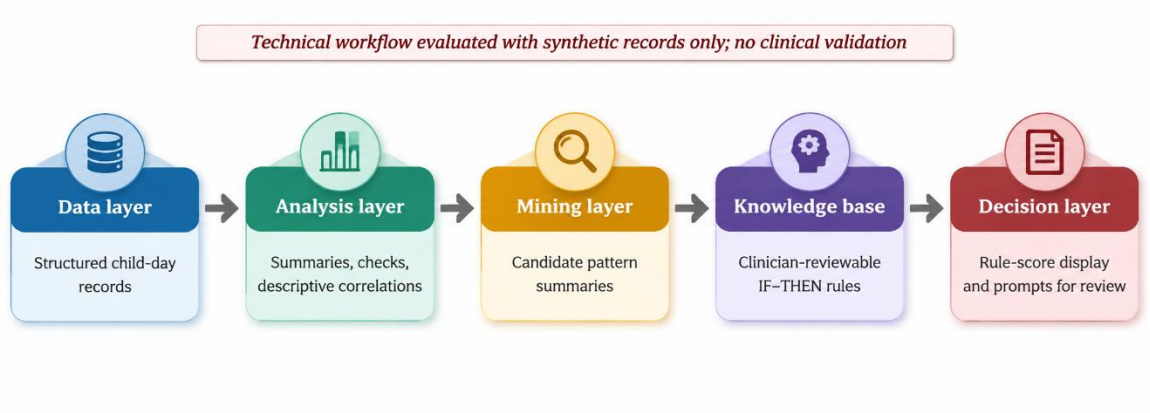

Figure 1. *Five-layer architecture of the rule-based workflow. The workflow is evaluated only with synthetic records and is not clinically validated.*

5.5. Knowledge-base rules and score display

The nine score-generating rules are listed in Table 2. R10–R12 are non-scoring prompt labels for simulation and do not add points.

Table 2. Provisional rule definitions used for software verification.

Rule	Category	Condition	Increment
R01	Sleep	If SleepHours < 5	25
R02	Sleep	If SleepHours < 6 AND SleepQuality < 4	20
R03	Environment	If NoiseLevel > 7	20
R04	Combined	If SleepHours < 6 AND NoiseLevel > 6	30
R05	Routine	If RoutineChange = True	25
R06	Activity	If ActivityLevel > 8	15
R07	Screen exposure	If ScreenTime > 5	15
R08	Sensory	If SensoryOverload = True	30
R09	Critical combination	If SleepHours < 5 AND RoutineChange = True AND NoiseLevel > 5	40
R10	Prompt label	If BehaviorType = Meltdown	Non-scoring simulated crisis-prompt label
R11	Prompt label	If BehaviorType = Self-injury	Non-scoring simulated safety-prompt label
R12	Prompt label	If BehaviorType = Elopement	Non-scoring simulated safety-prompt label

The maximum theoretical sum of R01–R09 is 220 points. The display transformation is:
Equation 8. $Risk100 = 100 \times RiskRaw/220$.

The weights were retained as provisional software-test parameters from the supplied prototype. They were not derived from a completed clinician-adjudication process, an empirical risk model, or a validated clinical instrument. The rules overlap conceptually: the same low-sleep or high-noise state can activate more than one rule. Sensitivity scenarios therefore vary the weights and selected thresholds, and future clinical work should evaluate alternative aggregation methods.

5.6. Evaluation measures

The primary outputs are descriptive implementation summaries. Pearson correlation describes the linear association between a rule score and the generator-defined severity outcome in a particular synthetic realization. It does not establish discrimination, calibration, causality, or clinical utility. Because the records are nested within three hypothetical children, no p-value or population confidence interval is reported. AUROC, area under the precision–recall curve, Brier score, calibration plots, and decision-curve analysis were not reported as clinical performance measures because the benchmark contains neither a validated binary clinical endpoint nor a prespecified prediction horizon or calibrated probability target. Imposing an arbitrary threshold on the generator-defined Severity variable would create a

metric that reflects the simulation design rather than clinical discrimination or utility. These metrics should be pre-specified for a future human validation study.

An exploratory linear baseline was fitted using the same seven input variables as predictors of Severity. The model was evaluated by leaving one child out at a time. This participant-level split prevents the same hypothetical child from appearing in both development and evaluation partitions. A mean-only model provided a simple reference. The baseline comparison is a software stress test and not a clinical validation.

Association-rule summaries use antecedent count, support count, confidence, and support percentage. The five queries were pre-specified demonstration queries for the mining layer rather than hypotheses discovered after inspecting the benchmark outcomes. The summaries are descriptive and are not interpreted as clinical probabilities. Rules with small support are retained only to show the behavior of the mining layer and are discussed as unstable.

5.7. Deterministic verification and sensitivity analysis

The deterministic audit checks the worked example, the strict screen-time boundary, the theoretical all-rules maximum, prompt behavior, and the equality between the stored score and the independently recomputed score. It also checks that each printed rule can be activated by at least one purpose-built test case, including R09, which is rare in the fixed-seed benchmark.

Rule sensitivity is assessed on the fixed-seed records using global ±20% weight scaling, R09 changed from 40 to 30 points, removal of R05, a lower R03 noise threshold of 6, and a lower R07 screen threshold of 4. Global scaling changes score magnitude but not Pearson correlation; the non-uniform scenarios are therefore more informative about rank and overlap sensitivity.

A within-child permutation negative control was conducted for 1,000 permutations. Severity values were shuffled within each hypothetical child, preserving the child-specific severity distribution while breaking same-day alignment with RiskRaw. The analysis recorded both the pooled correlation and a child-centered correlation after subtracting each child's mean RiskRaw and mean Severity. The observed correlations were compared with their corresponding null distributions; these are diagnostic controls, not population-level significance tests.

Rule ablation was evaluated by removing each score-generating rule separately and by removing four pre-specified rule families. For each scenario, the analysis recorded the percentage of records with a changed score, mean and maximum absolute score change, Spearman rank correlation with the baseline score, and overlap of the highest-scoring decile. These metrics assess score and ranking sensitivity rather than clinical performance.

5.8. Monte Carlo analysis

The generator was run for 100 independent seeds from 20260001 through 20260100. Each replicate recorded episode count, mean raw and display scores, the descriptive RiskRaw–Severity correlation, and selected association-rule summaries. These percentiles describe variation across generated datasets. They are not confidence intervals for a human population and should not be interpreted as external validation.

6. Results

6.1. Fixed-seed benchmark

The fixed-seed release contains 270 records, with 90 days for each of the three hypothetical children. It contains 105 episode-coded records (38.89%) and 165 records without a coded episode (61.11%). Behavior counts are shown in Table 3 and Figure 2.

Table 3. Behavior categories in the fixed-seed synthetic benchmark.

BehaviorType	Count	Percentage of all child-day records
NoEpisode	165	61.11
Tantrum	23	8.52
Aggression	19	7.04
Withdrawal	17	6.30
Meltdown	15	5.56
Stereotypy	14	5.19
Elopement	10	3.70
Self-injury	7	2.59

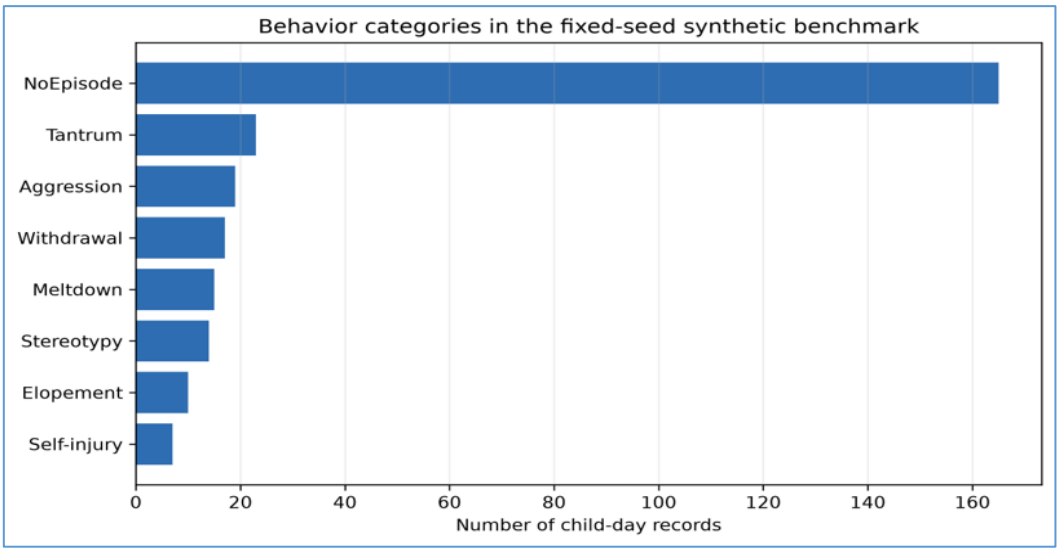

Figure 2. *Behavior categories in the fixed-seed synthetic benchmark. Percentages are generated-record summaries, not prevalence estimates.*

6.2. Descriptive correlations

Table 4 and Figure 3 report pooled descriptive correlations with the generator-defined Severity outcome. Sleep hours had a correlation of −0.2625, while sensory overload and noise level had correlations of 0.2677 and 0.2014, respectively. Activity level and screen time showed weak correlations in this realization. These values reflect the explicit synthetic data-generating process; they do not estimate relationships in children with ASD.

Table 4. Descriptive correlations with generator-defined severity in the fixed-seed benchmark.

Variable	Pearson r
Sleep hours	−0.2625
Sleep quality	−0.1565
Noise level	0.2014
Activity level	0.0470
Screen time	−0.0043
Routine change	0.1432
Sensory overload	0.2677
RiskRaw	0.3687

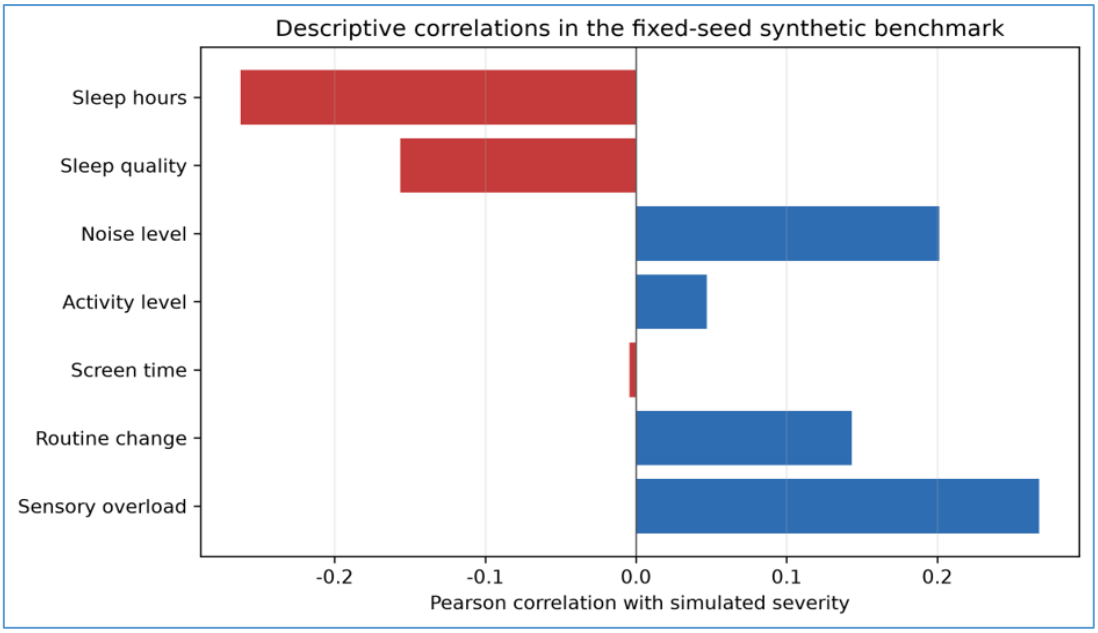

Figure 3. *Descriptive correlations in the fixed-seed synthetic benchmark. The coefficients are pooled summaries over 270 generated child-day records.*

The rule score is constructed from thresholded versions of several contextual variables that also enter the severity equation. Consequently, a positive association is expected under the declared generator. This is a construct-consistency check, not an unbiased performance estimate. The low pooled correlation for activity level and the near-zero correlation for screen time also demonstrate that inclusion in a rule base does not guarantee a strong association with the chosen synthetic outcome.

6.3. Participant-level summaries

Child-level summaries are shown in Table 5. Mean RiskRaw was highest for C02, whose profile also had the highest mean generator-defined severity. Within-child correlations varied, illustrating why pooled record-level summaries should not be treated as participant-level evidence.

Table 5. Participant-level summaries and within-child descriptive correlations.

Child	Age	Days	Episode records (%)	Mean RiskRaw (SD)	Mean Severity (SD)	Within-child RiskRaw–Severity r
C01	4	90	30 (33.33)	24.50 (22.60)	3.7796 (1.0769)	0.0984
C02	8	90	38 (42.22)	50.67 (33.13)	4.5452 (1.1345)	0.3217
C03	11	90	37 (41.11)	26.22 (28.44)	3.7937 (1.1818)	0.3884

6.4. Association-rule summaries

Table 6 reports five pre-specified demonstration queries. AR4 and AR5 have small support counts and are retained only to demonstrate the mining-layer output. Their confidence values must not be interpreted as clinical probabilities.

Table 6. Descriptive association-rule summaries from the fixed-seed synthetic benchmark.

Rule	Antecedent	Consequent	Antecedent count	Confidence (%)	Support count	Support (%)
AR1	NoiseLevel > 7	Severity ≥ 5	60	36.67	22	8.15
AR2	RoutineChange = True	Severity ≥ 5	91	29.67	27	10.00
AR3	SensoryOverload = True	Severity ≥ 5	95	35.79	34	12.59
AR4	ScreenTime > 4 AND SleepHours < 6	Severity ≥ 5	19	31.58	6	2.22
AR5	ActivityLevel > 7 AND NoiseLevel > 5	Aggression or Tantrum	37	21.62	8	2.96

6.5. Exploratory comparison with a simple baseline

A linear model using SleepHours, SleepQuality, NoiseLevel, ActivityLevel, ScreenTime, RoutineChange, and SensoryOverload was fitted on two children and evaluated on the held-out child. Table 7 gives the results alongside a mean-only reference. The model reduced RMSE relative to the mean-only reference in each held-out-child analysis, but the differences are specific to the declared generator and the three-child benchmark. The table is therefore included to demonstrate a comparison protocol, not to establish superiority of either method.

The rule score cannot be compared directly with the linear model's RMSE because RiskRaw is a point score on a 0–220 scale whereas Severity is a generator-defined 0–8.5 outcome. The comparable descriptive statistic is the correlation: the rule-score correlations for C01, C02, and C03 were 0.0984, 0.3217, and 0.3884, while the corresponding held-out linear-model correlations were 0.1095, 0.3236, and 0.3809. A decision-tree comparator was not added because a three-profile synthetic benchmark cannot support a meaningful model-complexity comparison; it is reserved for a prespecified participant-level evaluation with a clinically defined endpoint.

Table 7. Exploratory leave-one-child-out baseline comparison.

Held-out child	Test days	Linear-model RMSE	Linear-model MAE	Linear-model r	Mean-only RMSE	Mean-only MAE
C01	90	1.1212	0.9342	0.1095	1.1397	0.9549
C02	90	1.1922	0.9731	0.3236	1.3595	1.1117
C03	90	1.1606	0.9408	0.3809	1.2317	0.9760

6.6. Worked example

The worked example uses SleepHours = 4.5, SleepQuality = 3, NoiseLevel = 8, ActivityLevel = 7, ScreenTime = 5, RoutineChange = True, SensoryOverload = True, and BehaviorType = Meltdown. R01, R02, R03, R04, R05, R08, and R09 are active. R06 is inactive because ActivityLevel is not greater than 8. R07 is inactive because ScreenTime is exactly 5 and the condition is strictly > 5.

Table 8. Inputs for the worked example.

Input variable	Value
Sleep hours	4.5
Sleep quality	3
Noise level	8
Activity level	7
Screen time	5
Routine change	True
Sensory overload	True
Behavior type	Meltdown

The raw score is:

Equation 9. $RiskRaw = 25 + 20 + 20 + 30 + 25 + 30 + 40 = 190.$

The display score is:

Equation 10. $Risk100 = 100 \times (190/220) = 86.36.$

This is a provisional high rule-score display, not a clinical risk category, probability of harm, or emergency-management protocol. The Meltdown label activates a non-scoring simulated crisis-prompt label. It does not prescribe a clinical response.

6.7. Deterministic software verification

The audit passed the fixed-seed score recomputation, worked-example, strict boundary, all-rules maximum, and prompt-separation tests. R09 was covered by a purpose-built test case even though it appeared only once in the fixed-seed generated benchmark. When all nine score-generating rules are activated, the raw score is 220 and the display score is 100.

Table 9. Deterministic rule-engine verification cases.

Verification case	Expected raw score	Expected display score	Expected behavior
Worked example	190	86.36	R01, R02, R03, R04, R05, R08, R09; non-scoring simulated crisis-prompt label
Neutral profile with ScreenTime = 5	0	0.00	R07 remains inactive because the condition is strictly > 5
Same profile with ScreenTime = 5.0001	15	6.82	R07 becomes active
All nine score-generating rules active	220	100.00	R01–R09 active; no behavior-specific prompt
R09 coverage fixture	40	18.18	R09 activates in a dedicated test fixture

6.8. Rule sensitivity

Sensitivity results are shown in Table 10 and Figure 4. Global weight scaling changes the magnitude of RiskRaw but leaves Pearson correlation unchanged because it applies a common positive scalar. The non-uniform scenarios are more informative. Removing R05 reduced the

pooled descriptive correlation from 0.3687 to 0.3363, whereas changing the R09 weight from 40 to 30 produced only a small change. Lowering the R03 threshold from 7 to 6 increased the mean score and changed the correlation to 0.3756. These are properties of the fixed-seed generator, not evidence for an optimal threshold.

Table 10. Sensitivity of the rule score to provisional weights and selected thresholds.

Scenario	Mean raw score	SD raw score	Maximum observed score	Correlation with Severity
Baseline weights / baseline thresholds	33.80	30.71	140	0.3687
All weights −20%	27.04	24.57	112	0.3687
All weights +20%	40.56	36.85	168	0.3687
R09 weight = 30	33.76	30.58	130	0.3704
R05 removed	25.37	28.62	125	0.3363
Noise R03 threshold = 6	36.91	30.29	140	0.3756
Screen R07 threshold = 4	35.63	31.03	140	0.3676

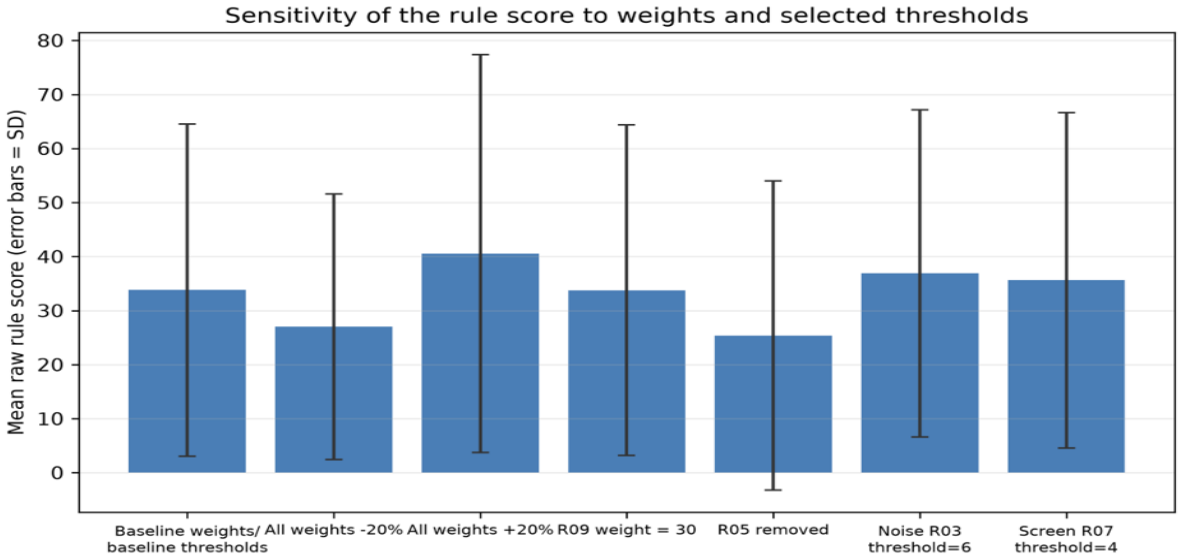

Figure 4. *Sensitivity of the rule score to provisional weights and selected thresholds. Error bars show the within-benchmark standard deviation of RiskRaw.*

6.9. Negative-control and rule-ablation analyses

The within-child permutation negative control produced a null mean pooled RiskRaw–Severity correlation of 0.1184 (SD 0.0536; 5th–95th percentile, 0.0279–0.2103), whereas the observed pooled correlation was 0.3687. After child-centering, the observed correlation was 0.2857, compared with a null mean of 0.0005 (SD 0.0611; 5th–95th percentile, −0.1026–0.1052). Both observed values exceeded all 1,000 corresponding permuted correlations in this diagnostic run. The result is consistent with same-day alignment contributing information beyond preserved child-level distributions, but it does not convert the synthetic association into predictive accuracy or clinical evidence.

Table 11. Within-child permutation negative control.

Analysis	Observed r	Null mean r	Null SD	Null 5th–95th percentile	Observed percentile
Pooled correlation; 1,000 within-child permutations	0.3687	0.1184	0.0536	0.0279–0.2103	100.0
Child-centered correlation; 1,000 within-child permutations	0.2857	0.0005	0.0611	−0.1026–0.1052	100.0

The ablation results show that the rule display is most sensitive to the environmental/sensory family and to R08 in this benchmark. Removing R03 and R08 changed 124 of 270 records (45.93%), reduced the mean score by 15.00 points, and produced a Spearman rank correlation of 0.7417 with the baseline score. Removing the sleep-related family changed 49 records (18.15%). At the individual-rule level, R08 changed 35.19% of records and had a rank correlation of 0.8205, whereas rare R09 changed only one record (0.37%). These findings identify implementation sensitivity; they do not identify clinically optimal rules.

Table 12. Pre-specified rule-family ablation and ranking sensitivity.

Ablated rule family	**Records changed (%)**	**Mean absolute score change**	**Maximum absolute score change**	**Spearman rank correlation**
Sleep-related rules (R01, R02, R04, R09)	49 (18.15)	6.93	85	0.9073
Environmental/sensory rules (R03, R08)	124 (45.93)	15.00	50	0.7417
Routine-change rule (R05)	91 (33.70)	8.43	25	0.8913
Activity/screen rules (R06, R07)	59 (21.85)	3.44	30	0.9765

The complete individual-rule ablation results are supplied in rule_ablation_v2.csv, and the 1,000 null correlations are supplied in null_control_correlations_v2.csv. Because these scenarios are evaluated on the same synthetic records and use the same provisional logic, they should be read as software-sensitivity diagnostics rather than feature-importance estimates.

6.10. Monte Carlo sensitivity across independent generated datasets

Across 100 independent seeds, the mean pooled RiskRaw–Severity correlation was 0.3905 (SD 0.0617; 5th–95th percentile, 0.2857–0.4867). The mean number of episode-coded records was 100.9 (SD 8.42; 5th–95th percentile, 88.95–116.00). The mean raw rule score was 33.7519 (SD 3.0510; 5th–95th percentile, 28.9389–38.0398), and the mean display score was 15.3418 (SD 1.3868; 5th–95th percentile, 13.1540–17.2908). AR4 confidence had a mean of 30.70% and a 5th–95th percentile range of 14.29%–50.00%, with a mean support count of 4.93. These

values show sensitivity of the generator and rule summaries; they are not confidence intervals for a human population.

Table 13. Monte Carlo sensitivity summaries across 100 independent generated datasets.

Metric	Mean	SD	5th percentile	Median	95th percentile
Risk–severity correlation	0.3905	0.0617	0.2857	0.3901	0.4867
Episode records per 270 records	100.90	8.42	88.95	101.00	116.00
Mean raw risk score	33.7519	3.0510	28.9389	33.7963	38.0398
Mean normalized display score	15.3418	1.3868	13.1540	15.3620	17.2908
AR1 confidence (NoiseLevel > 7 → Severity ≥ 5)	31.93	5.67	23.41	31.85	40.92
AR1 support count	19.49	4.64	12.00	19.00	27.05
AR4 confidence (ScreenTime > 4 AND SleepHours < 6 → Severity ≥ 5)	30.70	10.69	14.29	28.57	50.00
AR4 support count	4.93	2.43	2.00	5.00	9.00

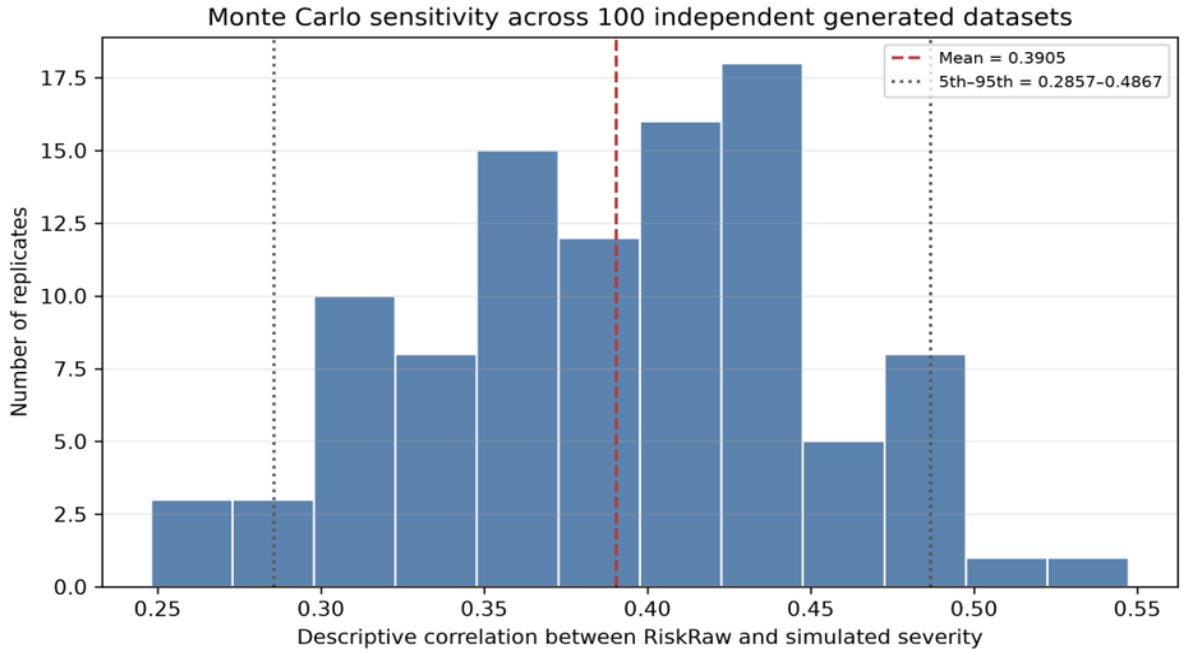

Figure 5. *Distribution of the pooled descriptive RiskRaw–Severity correlation across 100 independent generated datasets. The percentile range reflects generator variation, not uncertainty in a human population.*

7. Discussion

7.1. Principal findings

The study demonstrates that an interpretable rule-based workflow can be implemented and audited on an explicit synthetic longitudinal benchmark. The release specifies the data-generation equations, rule definitions, score normalization, deterministic test cases, baseline comparison, sensitivity scenarios, and Monte Carlo seeds. This is an engineering result. It does not establish that the workflow predicts behavioral escalation in children with ASD or that its prompts improve care.

The worked example is useful because it makes the distinction between rule execution and interpretation visible. Seven score-generating rules activate and produce 190 raw points, corresponding to a display value of 86.36 on the provisional 0–100 scale. The score is not a calibrated probability. The display scale makes the arithmetic easier to inspect, but it does not create clinical meaning.

The fixed-seed pooled correlation of 0.3687 and the Monte Carlo mean of 0.3905 reflect the declared data-generating process. Because Severity includes several variables that also appear in the rule base, the direction of the association is partly structural. The child-centered correlation of 0.2857 is lower than the pooled estimate, which shows that between-child profile differences contribute to the pooled association. Both the pooled and child-centered permutation controls were designed to diagnose same-day alignment rather than to provide population-level p-values. The association should therefore be described as a construct-consistency check. It cannot be used to claim discrimination, calibration, or clinical utility.

The child-level summaries reinforce the importance of the participant-versus-record distinction. C02 had a higher mean rule score and a higher mean generator-defined severity than C01 and C03, while within-child correlations differed. A future human study should report participant-level distributions and should split development and evaluation data by participant when the intended claim concerns generalization to new children.

7.2. Comparison with existing literature

The benchmark is deliberately distinct from empirical autism digital-health resources. Ecological momentary assessment can capture repeated real-world experiences, but the available autism EMA literature remains concentrated in selected populations and does not make a synthetic benchmark equivalent to a child-centered measurement protocol [5]. The Simons Sleep Project demonstrates the value of multi-device home recordings and parent-reported measures for future calibration, yet it uses different ages, measurements, and outcomes [6,7]. Recent wearable reviews likewise support caution about advance forecasting claims rather than validation of the present rule score [11]. These distinctions define the contribution of this study as software verification, not clinical generalization.

7.3. Methodological implications: baseline and rule sensitivity

The leave-one-child-out linear model provides a simple comparator and a protocol for future evaluation. Its lower RMSE than the mean-only reference in this benchmark does not show that a linear model is clinically preferable; both models were evaluated on outcomes generated by the same synthetic mechanism. A future study should compare the rule system with prespecified statistical and machine-learning baselines using participant-level splits and a clinically meaningful outcome horizon.

Rule sensitivity shows that global weight scaling changes score magnitude without changing rank-based correlation. The ablation analysis adds a more informative result: removing environmental/sensory rules changed 45.93% of records whereas and removing the rare R09 rule changed only one record. This result does not imply that any rule should be removed. It shows instead that provisional weights and thresholds should not be treated as self-validating. Because several rules overlap, future work should consider priority logic, capped contributions, or clinically reviewed aggregation rather than assuming that additive scoring is optimal.

7.4. Association summaries and intervention labels

The association-rule summaries are appropriate as demonstrations of data mining output only. AR4 and AR5 have small support counts, and their confidence values vary across generated datasets. They should not appear as clinical probabilities. The intervention labels were removed from the main results because their simulation-coded resolution values could be mistaken for treatment-effect estimates. They are retained in Technical Appendix Table A2 solely to document the generated file and to support output-integrity checks.

7.5. Implications for future data collection

A validation study should recruit a sufficiently large and diverse cohort from more than one setting. Each observation should include a timestamp, child identifier, measurement source, antecedent window, episode onset and end time, operational severity definition, intervention start time, outcome definition, and missingness indicator. Sleep should be distinguished by source, such as caregiver estimate, actigraphy, wearable output, or clinical measurement. Noise and activity should have defined units or validated ordinal scales. Where feasible, clinicians and trained caregivers should receive the same operational definitions, and inter-rater reliability should be quantified.

The data should be divided by participant rather than by day when the research question concerns generalization to new children. If personalization within a child is intended, it should be stated explicitly and evaluated with a prospective temporal split. The prediction horizon must be defined before analysis. Variables recorded after episode onset cannot be used as predictors of that same episode without introducing leakage.

7.6. Safety, human factors, and governance

A future system should display the evidence behind each prompt, identify missing or out-of-range data, and provide an explicit override or suppression mechanism. Evaluation should record false alarms, missed episodes, user overrides, reasons for disagreement, time to response, usability, and adverse events. The system should not direct users toward restrictive or physically intrusive interventions. Acute distress should prompt appropriate human and medical assessment, particularly when pain, illness, injury, or immediate danger is possible [4].

The rule base should be versioned and accompanied by a change log. Any update to thresholds or prompts should be evaluated as a new version. If future work uses data from hospitals, treatment centers, or controlled-access repositories, the investigators should obtain the required ethics determination, comply with consent and data-use conditions, minimize re-identification risk, and address parental permission and developmentally appropriate child assent.

7.7. Strengths and limitations

The strengths are the explicit separation between data generation and rule execution, the reproducible fixed seed, the participant-level subgroup summaries, the held-out-child baseline protocol, the non-uniform rule-sensitivity and ablation analyses, the within-child negative control, the deterministic audit, and the removal of simulated intervention outcomes from the main results. The data dictionary and parameter file make the benchmark easier to inspect than an undocumented simulation.

The limitations are substantial. The benchmark contains only three hypothetical children, uses no real reference dataset, and has no validated clinical outcome. The generator was designed by the investigators and may encode assumptions that are unrealistic or incomplete. Missingness, irregular reporting, measurement error, caregiver burden, device failure, and dataset shift are not modeled. The outcome mechanism shares variables with the rule base, so the reported correlations are structurally induced. The weights were not adjudicated by clinicians. No external validation, prospective observation, usability study, safety assessment, or treatment comparison was performed. These limitations restrict the work to software verification and methodological demonstration.

8. Conclusions

This study presents a reproducible rule-based workflow for organizing synthetic daily behavioral observations related to ASD. The release specifies the data-generating equations, documents the rule engine, tests boundary behavior, compares the workflow with a simple

held-out-child baseline, examines rule sensitivity, and reports Monte Carlo variation across independent generated datasets.

The results should be interpreted as implementation and construct-consistency summaries. They do not provide evidence of clinical predictive accuracy, probability of harm, therapeutic effectiveness, caregiver benefit, or safety. The next research stage requires independent de-identified longitudinal data, participant-level evaluation, clinically reviewed outcomes, explicit missingness mechanisms, comparison with prespecified baselines, calibration and utility analyses, and staged human-factors and safety evaluation. Whether the present technical feasibility extends to real-world clinical utility is unknown.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Association; 2013. https://www.psychiatry.org/psychiatrists/practice/dsm

2. Cohen S, Conduit R, Lockley SW, Rajaratnam SMW, Cornish KM. The relationship between sleep and behavior in autism spectrum disorder (ASD): a review. Journal of Neurodevelopmental Disorders. 2014;6:44. https://doi.org/10.1186/1866-1955-6-44

3. Cortese S, Wang F, Angriman M, Masi G, Bruni O. Sleep disorders in children and adolescents with autism spectrum disorder: diagnosis, epidemiology, and management. CNS Drugs. 2020;34:415–423. https://doi.org/10.1007/s40263-020-00710-y

4. Royal Children's Hospital Melbourne. Autism and developmental disability: management of distress/agitation. Clinical Practice Guideline. Updated June 2022. https://www.rch.org.au/clinicalguide/guideline_index/Autism_and_developmental_disability__Management_of_distress/agitation/

5. Chen Y, Xi Z, Greene T, Mandy W. A systematic review of ecological momentary assessment in autism research. Autism. 2025;29(6):1374–1389. Published online 18 December 2024. https://doi.org/10.1177/13623613241305722

6. Hacohen M, Levy A, Kaiser H, et al. An open science resource for accelerating scalable digital health research in autism and other neurodevelopmental conditions. Nature Neuroscience. 2026;29:467–478. https://doi.org/10.1038/s41593-025-02146-3

7. Simons Foundation Autism Research Initiative. Simons Sleep Project: open-science data resource. https://www.sfari.org/resource/simons-sleep-project/

8. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378

9. Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ. 2023;380:e071018. https://doi.org/10.1136/bmj-2022-071018

10. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine. 2022;28:924–933. https://doi.org/10.1038/s41591-022-01772-9

11. Romani PW, D'Mello SK, Moulder RM, Berkowitz LN. Using wearable technology to predict the occurrence of severe behavior problems among neurodiverse individuals: a

systematic review. Perspectives on Behavior Science. 2026;49:361–383. https://doi.org/10.1007/s40614-026-00497-1
12. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Medical Research Methodology. 2020;20:108. https://doi.org/10.1186/s12874-020-00977-1
Amna Ahmed Juma Awheedah. (2026). Psychological and Social Stress among Families of Children with Autism Spectrum Disorder and the Role of Psychological Support. Libyan Journal of Educational Research and E-Learning (LJERE), 2(1), 27-43. https://doi.org/10.65417/ljere.v2i1.62
Najat Nouri Naseer, & Farhat Al-Mabrouk Ali Zayed. (2025). Parenting Styles and Their Relationship to Aggressive Behavior Among Children: A Field Study. Journal of Libyan Academy Bani Walid, 1(4), 230–246. https://doi.org/10.61952/jlabw.v1i4.274

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.