*applied sciences*

MDPI

*Review*

# From Unimodal Tools to Multimodal Systems: A Structured Narrative Review of Selected Early-Childhood Screening, Assessment, and Monitoring Applications

**Jihoon Moon**

Department of Data Science, Duksung Women's University, Seoul 01369, Republic of Korea; jmoon25@duksung.ac.kr

**Abstract**

Early-childhood digital screening, clinician-supervised assessment, and monitoring use smartphone images, home video, interaction audio, caregiver questionnaires, and non-contact sensors. Yet "multimodal" is sometimes applied when distinct patient-level inputs are not jointly modeled. This structured narrative review examines selected decentralized domains—neonatal jaundice, autism-related screening or assessment, and cry or sleep monitoring—in children from birth through six years. A structured targeted update through 29 July 2026 used PubMed, IEEE Xplore, version-of-record pages, citation chaining, and official guidance. Multimodality required the joint use of two or more conceptually distinct patient-level modalities for model development, inference, or decision generation. Of 22 retained empirical reports, 9 used multimodal fusion, representing four system lineages. Validation was task-specific; prospective or external evaluation was more developed in neonatal jaundice and autism-related applications, while remote-home, postmarket, or implementation evidence was concentrated in selected autism-related lineages. Limitations included matched unimodal comparisons, calibration, determinate result coverage, missing-input handling, subgroup transportability, independent replication, referral outcomes, and deployment evaluation. A non-aggregated decision-oriented matrix organizes these requirements across technical, clinical, operational, equity, caregiver, privacy, and safety dimensions. Future studies should test whether multimodal fusion improves clinically relevant decisions over matched unimodal baselines, and whether gains persist across independent settings and care pathways.

**Keywords:** early childhood; multimodal artificial intelligence; digital screening; neonatal jaundice; autism; home monitoring; clinical validation; child-centered governance

Check for updates

Academic Editor: Bernardo Sotto-Maior Peralva

Received: 8 August 2026
Revised: 10 September 2026
Accepted: 12 September 2026
Published: 15 September 2026

**Copyright:** © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Developmental surveillance and screening identify children who may benefit from further assessment or timely intervention, whereas diagnostic assessment integrates multiple sources of evidence under qualified professional judgment. This distinction is particularly important during the first years of life, when developmental trajectories change rapidly, behaviors are context-dependent, and children may not be able to describe symptoms reliably. Pediatric recommendations place screening within a longitudinal process that includes observation, caregiver concerns, standardized instruments, referral, assessment, and follow-up [1]. Digital tools may support parts of this pathway, including clinician-supervised assessment, but their outputs are clinically interpretable only when the intended user,

action threshold, abstention or failure pathway, and relationship to professional judgment are explicit.

The number of artificial intelligence (AI)-enabled medical devices has grown rapidly, yet pediatric-specific evidence remains limited. A 2026 analysis of 952 U.S. regulatory submissions found that 42 (4.4%) included pediatric age ranges, and only five technologies were exclusively pediatric [2]. A complementary analysis examined 876 FDA-authorized AI/ML-enabled devices and identified 149 with pediatric labeling [3]. Among those 149 devices, only 28 (18.8%) explicitly reported pediatric participants in validation datasets, 22 (14.8%) were validated using adult data only, and 99 (66.4%) did not clearly report pediatric representation [3]. Together, these findings show that pediatric labeling and age-specific validation are not equivalent. Performance derived from adult or age-ambiguous cohorts therefore cannot be assumed to transfer to young children. Consistent with this concern, a recent systematic review of 133 pediatric AI studies found that 76% used single-center retrospective data, only 21% reported external validation, and fewer than 5% had been integrated into routine clinical workflows [4].

The term "multimodal" also requires precision. In machine learning, modalities are distinct information channels that capture different aspects of a patient or context—for example, interaction audio and a caregiver questionnaire—rather than different transformations or views of the same signal [5–7]. In this review, a study is classified as multimodal only when two or more conceptually distinct patient-level modalities jointly contribute to model development, inference, or decision generation. Multiple images of the same body region, several views from one camera, repeated measurements from the same sensor family, or the use of a color-calibration card are not considered multimodal fusion. Likewise, converting an audio waveform into a spectrogram changes the representation of the acoustic signal but does not create an additional patient-level modality.

This distinction matters because a collection of studies may include images, video, audio, sensors, and questionnaires even when the individual models remain unimodal. By contrast, Bae et al. [8] jointly used parent–child interaction audio and standardized screening instruments in 1242 children aged 18–48 months, providing a direct example of patient-level multimodal risk stratification. Studies of this kind permit questions that a catalog of separate modality-specific tools cannot answer: whether each modality provides incremental value, how the system behaves when one input is unavailable, whether the fused output is calibrated, and whether any performance gain justifies the additional acquisition burden.

The scope of this review is deliberately narrower than that of broad surveys of pediatric AI. It focuses on children from birth through six years of age and on screening, risk stratification, clinician-supervised diagnostic assessment, or longitudinal monitoring that could occur in home, community, telehealth, or primary-care-facing pathways. Hospital-based studies are included when the acquisition method or intended workflow is plausibly transferable to these settings. Autonomous diagnosis without qualified professional interpretation, inpatient decision support, and general pediatric AI applications fall outside the primary scope of this review.

Rather than attempting to cover every use of AI in early childhood, this review examines three selected application domains: neonatal jaundice, autism-related screening or clinician-supervised assessment, and cry or sleep monitoring. These domains represent three distinct decentralized uses of digital data: physiological measurement through optical sensing, developmental-behavioral assessment through interactive or caregiver-mediated inputs, and repeated state monitoring through audio or non-contact sensing. The findings should be interpreted within these selected application domains and not as a census of early-childhood AI research as a whole.

Within these selected application domains, five research questions organize the synthesis.

- RQ1. Which early-childhood screening, clinician-supervised assessment, and monitoring tasks, populations, and care settings have been studied?
- RQ2. Which data modalities have been used, how do genuinely multimodal systems differ from collections of unimodal tools, and how have fusion, temporal alignment, and missing modalities been handled?
- RQ3. How have discrimination, calibration, subgroup performance, and external or prospective validity been assessed?
- RQ4. How have caregiver-led acquisition, abstention, clinical workflow integration, privacy, implementation outcomes, and post-deployment considerations been evaluated?
- RQ5. Which evidence gaps currently prevent translation from model performance to safe screening, assessment, and monitoring practice?

Throughout the synthesis, direct pediatric empirical evidence is distinguished from secondary reviews, cross-cutting methodological literature, regulatory or reporting guidance, and evidence-informed design implications. Contextual and methodological sources can identify requirements or design considerations, but they do not establish prospective validation in early-childhood care. Multimodality is treated as an empirically testable source of incremental decision value rather than as a label for heterogeneous data collection. Claims of clinical or deployment readiness are bounded by direct pediatric evidence on validation, calibration, coverage, robustness, and workflow performance.

## 2. Review Design and Methods

### *2.1. Design and Scope*

This article is a structured narrative review with a documented targeted search update, not a systematic review, scoping review, or meta-analysis. A complete reproducible record of the original search and screening process was unavailable; the updated search was designed to reassess and correct the seed corpus rather than to support a PRISMA claim. The review emphasizes explicit scope boundaries, source-level traceability, transparent reporting of both report and system-lineage counts, and critical interpretation, consistent with core quality domains for narrative reviews [9]. No PRISMA flow count is reported, and the charted counts should not be interpreted as prevalence estimates for the broader literature.

The review uses three anchor application domains identified in the source corpus and confirmed during the targeted update. Neonatal jaundice represents a physiologically anchored screening task with a laboratory reference standard; autism-related applications represent screening and clinician-supervised developmental assessment using behavioral and caregiver-derived inputs; and cry or sleep applications represent repeated state monitoring. These domains were selected to support a focused cross-domain comparison rather than exhaustive coverage of pediatric or early-childhood AI applications.

The target population was children from birth through six years of age. The intended-use boundary included screening, referral support, risk stratification, clinician-supervised diagnostic assessment, repeated monitoring, and acquisition-quality control in home, community, telehealth, or primary-care-facing settings. Specialist or hospital-based studies were retained when the input method—such as a smartphone image, home video, mobile interaction recording, or contactless sensor—was directly relevant to decentralized acquisition. Treatment selection, autonomous diagnosis without qualified professional interpretation, and laboratory-only modalities without a plausible decentralized pathway were outside the primary scope.

### 2.2. Eligibility Criteria and Operational Definitions

The eligibility criteria, evidence roles, and operational definitions used to guide study classification and evidence charting are summarized in Table 1.

**Table 1.** Eligibility criteria, evidence roles, and operational definitions.

| Element | Operational Definition | Reason for Inclusion in Charting |
| --- | --- | --- |
| Population | • Birth through 6 years<br>• Mixed-age studies only when early-childhood results were identifiable | • Limits developmental heterogeneity<br>• Aligns with pre-school-age screening, assessment, and monitoring pathways |
| Clinical task | • Screening<br>• Risk stratification<br>• Referral support<br>• Clinician-supervised diagnostic assessment<br>• Repeated monitoring<br>• Acquisition quality assessment | • Includes professionally interpreted diagnostic support<br>• Excludes autonomous diagnosis and treatment selection |
| Setting | • Home<br>• Community<br>• Telehealth<br>• Primary care<br>• Hospital-based study with plausibly decentralized acquisition | • Maintains translational relevance<br>• Retains validation under clinical supervision when acquisition is decentralizable |
| AI/digital method | Computational model producing a:<br>• Risk estimate<br>• Classification<br>• Measurement<br>• Monitoring state | • Requires patient-level digital input<br>• Excludes administrative tools and non-evaluated technology descriptions |
| True multimodal study | Two or more conceptually distinct patient-level modalities jointly used for:<br>• Model development<br>• Inference<br>• Decision generation | • Distinguishes patient-level fusion<br>• Avoids classifying collections of separate unimodal tools as multimodal |
| Fusion | • Joint integration of two or more conceptually distinct patient-level modalities<br>• Common representation, prediction, or decision<br>• Fusion stage: input/early, representation/intermediate, prediction/late, or decision [5–7]<br>• Excludes mere co-collection, parallel reporting, reference/calibration-only use, or representation change of one signal | • Separates model-level integration from multimodal data collection alone<br>• Supports consistent identification of the stage at which modalities interact |

**Table 1.** *Cont.*

| Element | Operational Definition | Reason for Inclusion in Charting |
| --- | --- | --- |
| Not multimodal | • Multiple views of one image type<br>• Repeated measurements from one sensor family<br>• Calibration card<br>• Derived representation such as an audio spectrogram | • Prevents modality inflation<br>• Keeps classification at the patient input-modality level |
| Direct empirical evidence | Pediatric evaluation at the:<br>• Participant level<br>• Recording level<br>• Workflow level | Supports claims about:<br>• Observed performance<br>• Modality use<br>• Validation<br>• Implementation maturity |
| Contextual evidence | • Secondary reviews<br>• Technical/methodological studies<br>• Reporting standards<br>• Ethics/governance frameworks<br>• Regulatory documents | • Interprets requirements and evidence gaps<br>• Does not establish pediatric clinical performance |
| Exclusions | • Adult-only or prenatal-only studies<br>• Autonomous diagnosis without professional interpretation<br>• Laboratory-only EEG/MRI without a plausible decentralized pathway<br>• Treatment optimization<br>• Non-health monitoring<br>• Commentary without relevant guidance<br>• Duplicate reports without added evidence | • Maintains clinical coherence<br>• Maintains developmental coherence<br>• Preserves the decentralized intended-use boundary |

### *2.3. Source Identification and Updating*

The source corpus was reassessed and expanded through a structured targeted search update, last run on 29 July 2026. Domain-specific PubMed searches and limited IEEE Xplore publisher record checks were supplemented by publisher version-of-record verification and backward and forward citation chaining. Exact database-specific queries and search approaches are reported in Supplementary Table S1. Official pages from the FDA, WHO, the European Union, the U.S. Federal Trade Commission, and Health Level Seven International were used for regulatory and governance context. Additional recent literature through 9 September 2026 was also reviewed to strengthen the discussion of pediatric AI, multimodal fusion, autism-related digital assessment, caregiver and human-factors issues, privacy, explainability, and model monitoring. These additional papers were used only as contextual support for the discussion and later study design recommendations. They were not added to the 22-report direct empirical analytic set, and the 29 July 2026 cutoff for that set remained unchanged. Source selection and initial charting were conducted in a single-reviewer workflow because the article was designed as a structured narrative review that revisited a pre-existing seed corpus. Because a complete reproducible record of the original screening process was unavailable, the updated search was not presented as duplicate systematic screening. To reduce subjectivity in classification, the age, intended-use, acquisition-setting, and patient-level multimodality criteria listed in Table 1 were applied consistently.

A report was classified as multimodal only when the version of record or available Supplementary Material explicitly showed that two or more conceptually distinct patient-level modalities jointly contributed to model development, inference, or decision generation. Multiple views of the same signal family, repeated same-family measurements, calibration artifacts, and derived representations were coded as non-multimodal; borderline cases were also coded as non-multimodal unless joint contribution was explicit. All 22 retained empirical reports were subsequently checked again against the same criteria and source documents. This additional check was intended to reduce inconsistent classification, but it does not replace duplicate independent screening or eliminate the possibility of selection and interpretation bias. The complete database-specific search strings and source-retention log are provided in Supplementary Table S1. The log is intended to improve the transparency and reproducibility of the targeted update; it is not presented as a PRISMA screening flow or as evidence of exhaustive retrieval. Source-specific counts may overlap and are not combined into a single screening denominator. Figure 1 summarizes the review boundary and evidence roles. It is an evidence-identification map rather than a PRISMA screening flow; the targeted update was used to reassess a pre-existing seed corpus within three selected application domains, while contextual methodological and governance sources were retained separately from the direct pediatric empirical set.

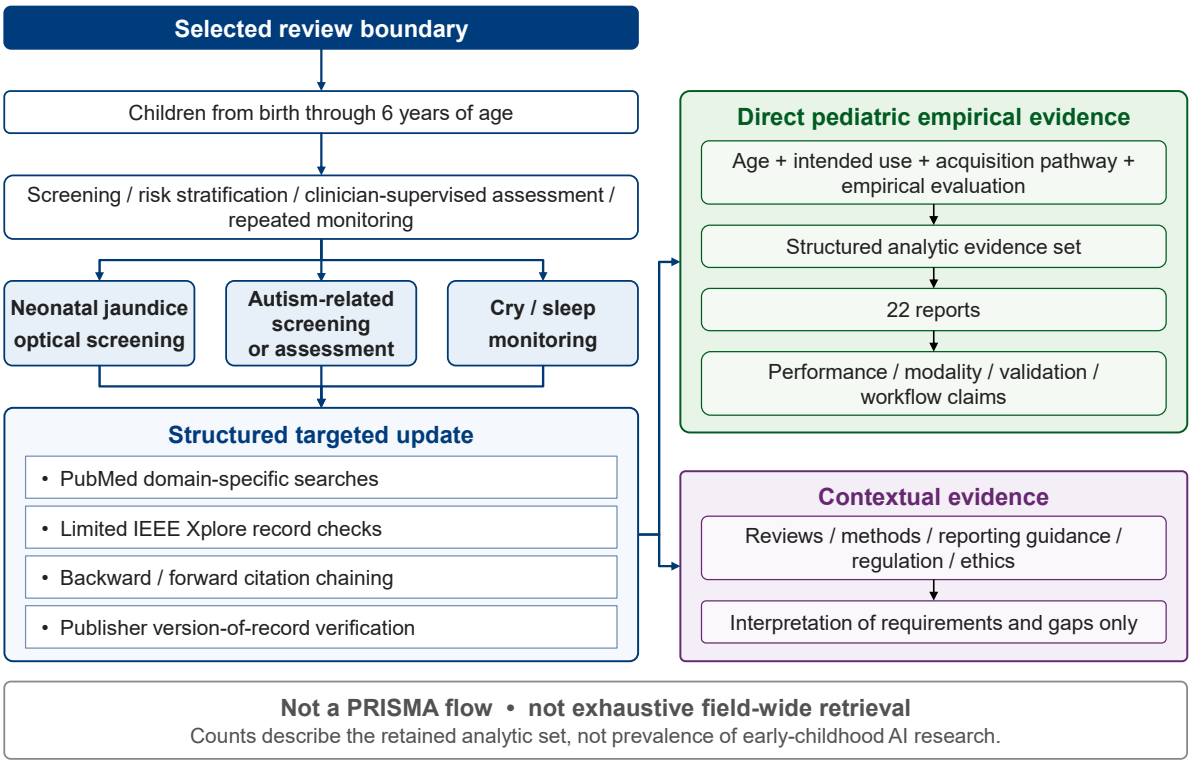



**Figure 1.** Review scope, targeted evidence-identification process, and evidence–role boundary. The review focuses on three purposively selected early-childhood application domains and uses a structured targeted update to reassess and expand a pre-existing source corpus. Direct pediatric empirical reports contribute to the analytic evidence set, whereas secondary methodological, reporting, regulatory, and governance sources are used only for contextual interpretation. The diagram is not a PRISMA screening flow and does not imply exhaustive field-wide retrieval.

### *2.4. Data Charting and Synthesis*

For direct empirical studies, charted variables included clinical domain, population and setting, patient-level input, multimodal status, fusion level when applicable, reference standard, validation design, principal performance signal, and the most decision-relevant

limitation. Evidence was summarized at three levels: study characteristics, distribution by clinical task, and cross-cutting validation or deployment gaps. Heterogeneity in target conditions, reference standards, thresholds, and intended uses made quantitative pooling and direct performance ranking inappropriate. Performance estimates are reported within the context of each task rather than combined into a common score.

Contextual sources were assigned one of four roles: secondary pediatric synthesis; cross-cutting methodological evidence; reporting, regulatory, or ethical guidance; and deployment context. Only direct pediatric empirical studies were used to support statements about observed discrimination, calibration, validation, workflow, or implementation outcomes. Architecture, privacy, uncertainty, and governance sources informed design requirements and interpretation unless a pediatric empirical study had directly evaluated the relevant approach.

## 3. Characteristics of the Analytically Charted Evidence

### *3.1. Distribution by Task and Modality*

Twenty-two direct empirical reports were retained for structured appraisal after application of the specified age, intended-use, acquisition setting, and modality criteria during the targeted update. Eight concerned neonatal jaundice, eleven concerned autism-related screening or assessment, and three concerned cry or sleep monitoring. Twenty-one reports evaluated performance or validation, whereas one prospectively evaluated a primary care implementation. Eleven reports used a single patient-level signal family, two used a single patient-level signal together with a non-patient calibration artifact, and nine jointly modeled conceptually distinct patient-level modalities.

The nine multimodal reports represented four system lineages. Within the Cognoa/Canvas Dx lineage, an early Cognoa report [10] was followed by a later multi-module report [11] and three subsequent reports evaluating Canvas Dx validation, postmarket use, or implementation [12–14]. Two reports evaluated the SenseToKnow lineage [15,16]. Bae et al. evaluated a separate audio-plus-screening-instrument framework [8], and Liang et al. combined neonatal skin imaging with structured clinical data within the BiliSG lineage [17]. Report counts and system lineage counts are presented separately to avoid treating successive publications or deployments of the same technological lineage as independent systems.

Figure 2 illustrates the two distinctions that govern the synthesis: first, multimodality is assigned at the patient-level model-input level rather than from the diversity of data types present across a literature set; second, publication counts are separated from technological lineages so that successive evaluations of the same platform are not interpreted as independent systems. To summarize the principal systems and approaches without implying report-level independence, Table 2 presents their patient-level modalities, evidence stage, principal evidence, and main limitations.

Table 2 summarizes systems or approaches rather than independent reports. Full report-level characteristics, sample information, performance metrics, coverage, and decision-relevant limitations are provided in Supplementary Table S2.

The distribution of reports also differed substantially across clinical areas in both multimodal representation and evidentiary stage. Table 3 summarizes the report counts, multimodal system-lineage counts, principal evidence signals, and persistent gaps within each selected application area.

Counts summarize reports identified through the structured targeted update and should not be interpreted as field-wide prevalence estimates. Parenthetical values indicate distinct system lineages rather than independent studies. “True multimodal” follows the patient-level fusion definition specified in Table 1.

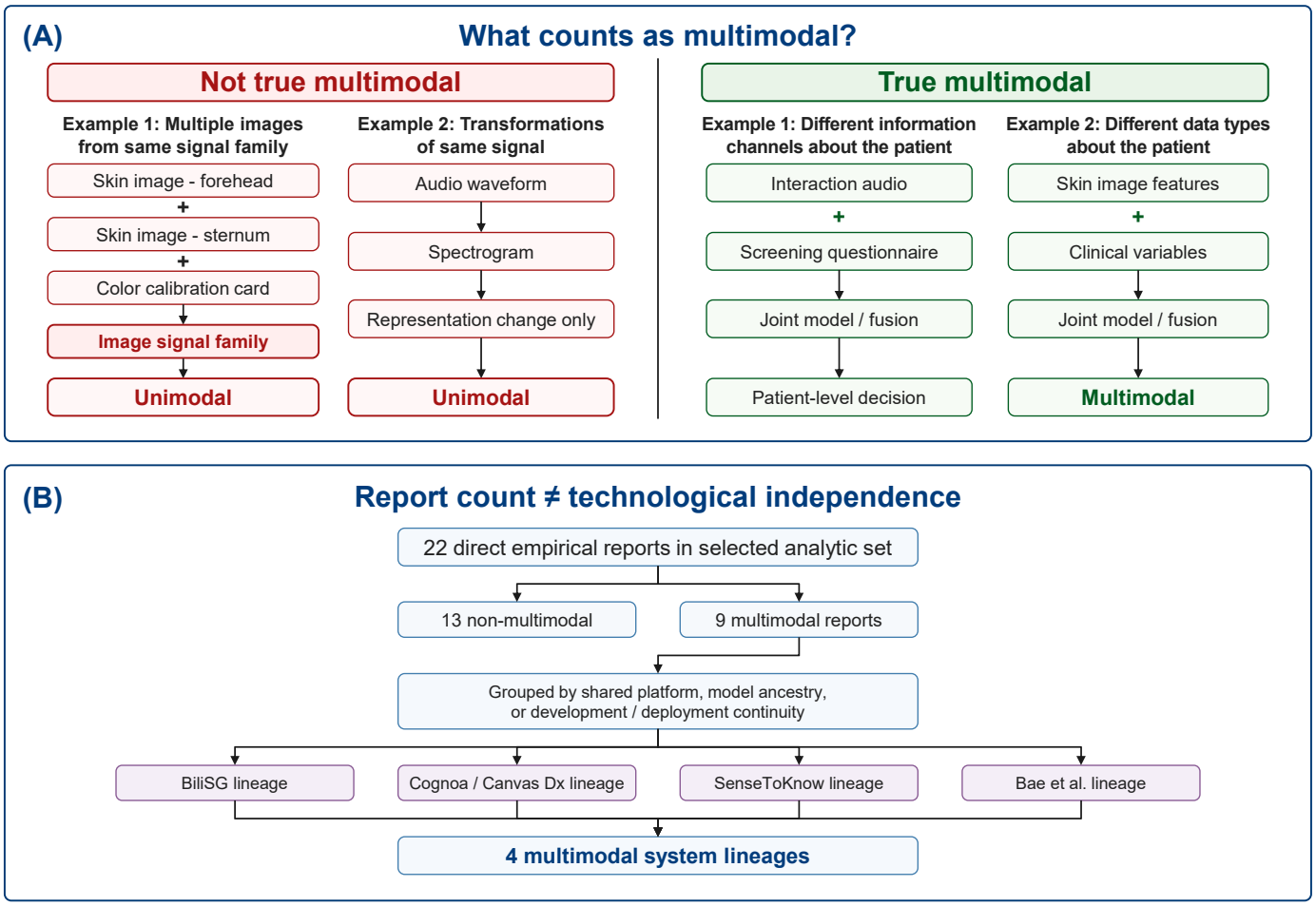



**Figure 2.** Operational definition of patient-level multimodality and separation of reports from system lineages. (**A**) Multiple views, transformations, or calibration artifacts from one patient-level signal family are not counted as distinct modalities, whereas jointly modeled conceptually distinct patient-level information channels are classified as multimodal. Structured clinical or questionnaire variables are treated as a distinct modality when they provide information conceptually different from image, audio, video, or sensor measurements. (**B**) The 22 direct empirical reports in the selected analytic set include 9 multimodal reports, but these map to four system lineages because several publications evaluate successive versions, cohorts, or deployments of the same underlying platform. The diagram represents analytical classification rather than a screening flow.

**Table 2.** System- and approach-level summary of the selected pediatric empirical evidence.

| Domain | System or Approach | Patient-Level Modalities | Evidence Stage | Principal Evidence | Main Limitation |
|---|---|---|---|---|---|
| Jaundice | BiliCam/Picterus/ BiliSG image tools | • Image | • Prospective<br>• Temporal<br>• Multicountry | • Comparison withtotal serum bilirubin | • Device transportability<br>• Home transportability |
| Jaundice | BiliSG multimodal model | • Image<br>• Clinical variables | • Secondaryfeature-fusion analysis | • Matched unimodal comparison | • Same acquisition cohort<br>• No external fusion validation |
| Autism | Cognoa/Canvas Dx lineage | • Questionnaires<br>• Video | • Development<br>• Pivotal evaluation<br>• Postmarket analysis<br>• Implementation | • Clinician-supervised evidence | • Lineage dependence<br>• Variable determinate-result coverage |

**Table 2.** *Cont.*

| Domain | System or Approach | Patient-Level Modalities | Evidence Stage | Principal Evidence | Main Limitation |
|---|---|---|---|---|---|
| Autism | SenseToKnow | • Camera-derived behavior<br>• Touch/inertial data | • Prospective clinic evaluation<br>• Remote-home validation | • Caregiver-led remote acquisition | • Independent pathway outcomes limited |
| Autism | Bae et al. [8] | • Interaction audio<br>• Screening instruments | • Multisite risk stratification | • Calibration<br>• Ablation analysis | • Missing audio<br>• External transportability |
| Autism | Eye tracking | • Gaze | • Multisite evaluation<br>• Replication | • Objective behavioral measurement | • Unimodal<br>• Referral effect untested |
| Cry | Acoustic classification | • Audio | • Proof of concept | • Signal classification | • Weak clinical labels |
| Sleep | Video/sensor monitoring | • Single signal family | • Laboratory feasibility | • State estimation | • Small sample<br>• No action-linked outcome |

**Table 3.** Descriptive distribution and evidentiary interpretation of the charted studies.

| Domain | Reports | Multimodal (Lineages) | Evidence Signal | Key Gap |
|---|---|---|---|---|
| Neonatal jaundice | 8/22 | 1 (1) | • Prospective TSB comparisons<br>• Multicountry or temporal evaluation<br>• One matched multimodal feature analysis | • Independent external fusion validation<br>• Device/operator transportability<br>• Skin-tone and high-bilirubin performance<br>• Caregiver-led transportability |
| Autism-related assessment | 11/22 | 8 (3) | • Multisite prospective testing<br>• Remote-home validation<br>• Replication<br>• Postmarket analysis<br>• Primary care implementation | • Independent lineage replication<br>• Screening-to-referral outcomes<br>• Missing-modality robustness<br>• Cultural/language validity |
| Cry/sleep monitoring | 3/22 | 0 | • Automated audio-derived state classification<br>• Video-derived state classification<br>• Feasibility evidence | • Patient-level sampling<br>• Clinically meaningful labels<br>• Independent validation<br>• Action-linked utility |

**Table 3.** *Cont.*

| Domain | Reports | Multimodal (Lineages) | Evidence Signal | Key Gap |
| --- | --- | --- | --- | --- |
| Overall | 22 | 9 (4) | • Discrimination comparatively well developed<br>• Calibration less consistently evaluated<br>• Implementation/workflow evidence limited in selected settings | • Independent comparative validation<br>• Calibration and coverage<br>• Prospective workflow evaluation<br>• Deployment evaluation |

### *3.2. Validation Patterns*

Validation patterns differed more clearly by clinical task than by algorithm family. Jaundice studies generally compared an app-derived estimate with total serum bilirubin (TSB), providing a relatively well-defined diagnostic accuracy reference pathway. Autism studies used clinical diagnosis, expert assessment, or established risk groups as reference standards, but their designs ranged from retrospective feature extraction to double-blind, multisite evaluation. Cry and sleep studies more often evaluated recognition accuracy against dataset labels or manually annotated states. These outcomes should not be interpreted along a single maturity scale because estimating bilirubin, supporting specialist diagnosis, screening an asymptomatic population, and labeling an infant state involve different reference standards, error costs, and downstream actions.

Discrimination was the most consistently reported performance dimension. Calibration was explicitly evaluated in the recent multimodal ASD study [8] but was otherwise uncommon. Some studies used independent, temporal, multicountry, or replication cohorts, yet external validity remained constrained in several reports by supervised acquisition, a limited device set, previously diagnosed populations, or narrow distributions of high-risk cases. Operational measures—including failed capture, time to result, repeat-acquisition burden, network failure, false-referral workload, and caregiver comprehension—were seldom evaluated as co-primary outcomes alongside model performance.

## 4. Clinical Evidence by Application Area

### *4.1. Smartphone-Based Neonatal Jaundice Screening*

Among the selected application domains, smartphone bilirubin assessment provides a relatively well-defined example of decentralized pediatric screening with an established confirmatory reference. BiliCam evaluated 530 newborns and reported high sensitivity with moderate specificity under threshold-specific rules [18]. Subsequent systems used sternal or multisite skin imaging or scleral imaging. Biliscan's early validation showed only moderate agreement with TSB [19], whereas the later BiliSG model used a larger multiethnic cohort and temporal evaluation, reporting 100% sensitivity and 70% specificity at a selected screening threshold [20]. This performance profile may be appropriate for a rule-out or referral tool, but high sensitivity alone does not establish suitability for treatment decisions. False positives, underestimation, and the consequences of threshold selection still require explicit management.

Geographic diversity was greater in later evaluations. A sclera-based application evaluated in Ghana achieved sensitivity of 94%, specificity of 73%, and area under the receiver operating characteristic curve (AUROC) of 0.92 for severe jaundice [21]. The Picterus system was evaluated in Mexico [22] and across Mexico, Nepal, and the Philip-

pines [23]. These studies supported feasibility in populations often underrepresented in imaging research, while also revealing site-dependent bias, wide limits of agreement, and underestimation at clinically important higher bilirubin values. A 2026 prospective single-center study in Türkiye reported a strong association with transcutaneous bilirubin but worsening agreement at higher values [24]. Because TcB rather than TSB was used as the comparator, the findings should be interpreted as a comparison between noninvasive measurements rather than as definitive laboratory validation.

Most jaundice reports remain unimodal under this review's definition. Color cards, standardized backgrounds, and multiple body sites may improve image acquisition or normalization, but they do not constitute a second patient-level modality. However, Liang et al. [17] provided an important exception. In 633 neonates, skin image features from three body regions were combined with structured clinical and demographic variables. The selected LightGBM model achieved RMSE 2.0477 mg/dL and Pearson correlation 0.8435, with more than a 10% RMSE improvement over single-modality models; clinical data contributed more than one-quarter of feature importance, led by hour of life. This matched comparison provided direct evidence of incremental multimodal value within the analyzed cohort rather than merely an architectural claim. However, the study was a secondary analysis of the BiliSG acquisition cohort [20], used controlled nurse-led iPhone 12 capture in Singapore, and lacked external or caregiver-led home validation.

A meta-analysis of smartphone applications found a broad range of sensitivity and specificity and judged the certainty of evidence to be moderate, reinforcing the need for well-designed threshold-specific studies [25]. The practical question is not simply whether smartphone-derived estimates correlate with bilirubin under controlled acquisition, but whether caregiver or community worker acquisition can reduce missed severe hyperbilirubinemia without producing excessive confirmatory testing, delay, or false reassurance across devices, lighting conditions, and skin appearances.

### *4.2. Autism Screening and Clinician-Supervised Assessment*

Autism-related studies span screening, diagnostic support, and objective behavioral measurement. Tariq et al. [26] showed that blinded raters could extract sparse behavioral features from brief home videos; an eight-feature model achieved AUROC 0.93 in a held-out set and 0.86 in an independent set of 66 videos. The findings supported the use of home video for rapid behavioral characterization, but participants already had diagnostic labels, video quality affected inclusion, and the behavior tags were human-mediated. The study provided evidence for a video-derived assessment approach rather than for population screening effectiveness.

Abbas et al. [10] combined a short parent questionnaire with behavioral tags from home video in a multicenter study of 162 children. This study was classified as true multimodal because two distinct patient-level information sources contributed to the combined decision. Its allowance for inconclusive results was also clinically relevant: abstention may reduce errors among conclusive cases, although coverage and follow-up burden must be reported alongside accuracy. The limited sample and missing-video pattern meant that the incremental benefit of fusion required confirmation in larger and more representative cohorts.

A later Cognoa study added a clinician questionnaire and evaluated parent, video, and clinician modules in 375 at-risk children aged 18–72 months [11]. With up to 30% of cases classified as inconclusive, the combined approach achieved AUROC up to 0.92, sensitivity of 90%, and specificity of 83%. This was a distinct report but not a fully independent replication; the 2016 cohort of 162 participants overlapped the earlier report [10], refinements were informed by that cohort, and the study was extended with 213 participants in 2017.

The evidence should be interpreted as progressive development within one system lineage, rather than as evidence from two unrelated multimodal systems.

The pivotal Canvas Dx evaluation added prospective multisite evidence to this lineage. Megerian et al. [12] conducted a double-blind, multisite, prospective active-comparator study in which 711 children were enrolled and 425 completed all study procedures. The device fused a caregiver questionnaire, analyst ratings of two home videos, and a healthcare provider questionnaire. It returned a determinate result for 31.8% of completed cases; among those cases, sensitivity was 98.4%, specificity 78.9%, positive predictive value 80.8%, and negative predictive value 98.3%. These estimates could not be interpreted separately from coverage; 68.2% of completed cases were indeterminate, and most indeterminate cases had at least one neurodevelopmental condition. The English-speaking sample with smartphone access and the underpowered subgroup comparisons also constrained generalizability.

Two 2025 reports extended this lineage with postmarket and implementation evidence. In 254 completed postmarket prescriptions, Canvas Dx produced determinate results for 63.0%; sensitivity was 99.1%, specificity 81.6%, positive predictive value 92.4%, and negative predictive value 97.6% among determinate cases [13]. The data were manufacturer-held aggregates from completed prescriptions, and information on race, ethnicity, socioeconomic status, and longitudinal diagnostic stability was unavailable. In a prospective observational integration study, 80 children were evaluated by seven ECHO-trained primary care clinicians [14]. Mean time from concern at enrollment to final determination was 39.22 days, compared with reported 180–264-day waits at adjacent specialist centers; 52.5% received a determinate device output, all of which agreed with the final clinical determination, and 93% of responding caregivers reported satisfaction. The nonrandomized design, small sample of trained clinicians, and contextual rather than concurrent wait-time comparison limited causal inference. Clinicians also continued to use professional judgment and additional structured observation when needed.

SenseToKnow represented a second multimodal lineage. In a prospective multiclinic well-child study of 475 children aged 17–36 months, camera-derived digital behavioral phenotypes and touch interaction were combined for autism screening [15]. The app alone achieved AUROC 0.90, sensitivity 87.8%, specificity 80.8%, positive predictive value 40.6%, and negative predictive value 97.8%; combining the app with M-CHAT-R/F increased AUROC to 0.97, sensitivity to 92.1%, and specificity to 91.8%. Interpretation should account for the incomplete comprehensive evaluation of some apparently neurotypical children and the lower specificity observed among Black children. A later remote study involved 620 toddlers aged 16–40 months whose caregivers administered the app at home using their own iPhone or iPad [16]. The study reported AUROC 0.92, sensitivity 83.0%, specificity 93.3%, positive predictive value 84.3%, and negative predictive value 92.6%, and provided direct evidence of caregiver-led multimodal acquisition outside a clinic.

Eye-tracking studies provided a different form of objective behavioral measurement. In a double-blind, six-site study, eye-tracking-based social visual engagement, measurements were successfully obtained for 475 of 499 enrolled children and achieved sensitivity of 71.0% and specificity of 80.7% relative to expert clinical diagnosis [27]. A related discovery-and-replication program in 1089 children reported sensitivity near 81% and specificity between 82% and 90% across cohorts [28]. These studies provided multisite and replication evidence, but the tools measured gaze behavior from a single patient-level signal family. The use of multiple visual stimuli and eye-tracking features did not make the model multimodal.

In addition, Bae et al. [8] evaluated a third multimodal lineage. A mobile application collected parent–child interaction audio and standardized screening instruments from 1242 children aged 18–48 months across nine Korean hospitals. A two-stage approach

distinguished typically developing children from high-risk or ASD groups (AUROC 0.942) and then separated high-risk from ASD groups (AUROC 0.91). Reported ablation results showed that the combined audio-plus-instrument model outperformed the audio-only model and modestly improved over the instrument-only model. Isotonic calibration was also evaluated. The inclusion of matched modality comparisons and calibration made this study particularly informative for assessing multimodal contribution. However, the exclusion of cases with missing audio, limited evidence outside the source population, and the absence of prospective referral outcomes prevented an inference of universal deployment readiness.

Intended use and evidentiary independence remain important when interpreting this evidence. The FDA De Novo decision for the Cognoa ASD Diagnosis Aid defines the device as an aid for qualified clinicians within specified ages, inputs, contraindications, and review requirements; it does not authorize autonomous diagnosis [29]. Five reports in the charted set concerned versions or deployments of the same lineage, and several included company-affiliated authors or manufacturer-controlled data. Their prospective, postmarket, and implementation contributions were valuable, but the number of reports should not be interpreted as the number of independent model replications. A digital biomarker review also found substantial variation in app characteristics, outcomes, and accuracy metrics [30]. Screening, diagnostic support, and risk stratification should remain distinct intended uses, with positive, negative, failed, and indeterminate outputs linked explicitly to referral, reassessment, or clinician review.

### *4.3. Cry and Sleep Monitoring*

Infant cry studies illustrate the gap between signal classification and clinical validity. Ashwini et al. [31] transformed audio into spectrograms and classified pain, hunger, and sleepiness, with a best reported test accuracy of 88.89%. Similarly, Hammoud et al. [32] used acoustic features to classify five cry-reason categories, and reported 96.39% test accuracy. Both were classified as unimodal audio studies under the definition used in this review. More importantly, high accuracy on short labeled clips does not establish that a model can distinguish clinical states in noisy home environments, across infant ages and microphones, or when several needs coexist. A review of infant cry analysis identified limited shared datasets, inconsistent labels, and weak cross-study comparability as persistent constraints [33].

As contextual evidence outside the 22-report analytic set, language-environment studies provided a related example of longitudinal acoustic monitoring. A meta-analysis of Language Environment Analysis measures found associations with child language outcomes, but also substantial variation across measures and study contexts [34]. Such tools may support monitoring or research, yet an observed predictive association does not establish that automated feedback improves language development. Evaluation should consider recording completeness, household language, socioeconomic context, consent of bystanders, and whether the resulting output changes a beneficial caregiver or clinician action.

Sleep monitoring raises similar evidentiary boundaries. Video-based actigraphy achieved 92.0% mean accuracy in 11.6 h of recordings from ten healthy infants [35], while an actigraphy review emphasized the need for standardized scoring and reporting in young children [36]. A related contextual study outside the 22-report analytic set compared non-contact radar with video-reviewed neonatal sleep/wake states [37]. In that design, video served as a reference rather than as a fused model input. It was not classified as multimodal inference. Future home studies should evaluate motion artifacts, occlusion, co-sleeping,

device placement, false alarms, and whether monitoring improves a prespecified clinical or caregiving outcome.

## 5. Multimodal Modeling, Robustness, and Decision Evaluation

### 5.1. Fusion Is an Empirical Claim, Not a System Label

Multimodal fusion can occur at the input, representation, prediction, or decision level [5–7]. The fusion level should be reported because it affects both potential failure modes and how the contribution of each modality can be evaluated. Early fusion may capture fine cross-modal interactions, but generally requires aligned and jointly available inputs. Intermediate fusion combines modality-specific representations after separate feature extraction. Late or decision-level fusion can preserve separate models and may accommodate asynchronous acquisition, but it can also obscure whether one modality dominates the final output. None of these strategies is inherently superior; the appropriate design depends on the clinical task, acquisition timing, missingness pattern, and burden of obtaining each input.

Nine charted reports used patient-level multimodal fusion [8,10–17], but these reports represented only four system lineages. The Cognoa/Canvas Dx lineage progressed from questionnaire-plus-video fusion to a three-input device, and was subsequently evaluated in pivotal, postmarket, and primary care implementation studies. SenseToKnow combined camera-derived behavior with touch or inertial interaction in clinic-based and remote-home evaluations. In addition, Bae et al. [8] combined interaction audio with standardized screening instruments. The BiliSG secondary analysis combined skin image features with structured clinical data, and directly compared the fused model with single-modality models [17]. The evidence shows that selected multimodal systems—particularly in autism-related applications—had progressed beyond initial proof-of-concept evaluation, although independent replication across technologies, populations, and care systems remains limited.

A claim of multimodal benefit requires comparison with the constituent modalities rather than an evaluation of the fused model alone. At minimum, studies should compare the fused model with each modality under the same data split, preprocessing procedure, and tuning budget, and should report modality ablation where applicable. Incremental value should not be expressed solely as a change in AUROC. Calibration, threshold-specific missed cases and false referrals, acquisition failure, determinate result coverage, and additional time or caregiver burden may also affect the clinical relevance of fusion. A small gain in discrimination may not justify an additional recording task, whereas a modality with little effect on average AUROC could still contribute to robustness in a clinically relevant subgroup. However, such subgroup benefit requires direct evaluation rather than assumption.

### 5.2. Temporal Alignment and Missing Modalities

Early-childhood modalities often operate on different time scales. A caregiver questionnaire may summarize behavior over several weeks, an interaction recording may capture only minutes, a bilirubin image reflects a single acquisition time, and a sensor stream may extend over hours. Fusion without a prespecified temporal relationship can combine clinically discordant states or introduce leakage if information acquired after an intended decision point contributes to an earlier prediction. Studies should specify the index time, allowable acquisition window, direction of prediction, and whether each modality was available before the intended clinical action.

Missingness can have both technical and clinical meaning. A failed audio recording may reflect household noise, language, child distress, device permissions, or caregiver capacity, and some of these factors may also be associated with the outcome or with

access to care. Complete-case evaluation can produce optimistic performance estimates and may conceal unequal input availability. In Bae et al. [8], cases without the required audio were excluded. SenseToKnow incorporated quality and missing-variable handling within its digital phenotype pipeline [15,16], whereas the Cognoa/Canvas Dx lineage used inconclusive or indeterminate outputs when the available evidence was insufficient [11–14].

These approaches address different aspects of missingness, but the retained studies do not provide a comprehensive head-to-head evaluation of all missing-input patterns, fallback strategies, and affected subgroups. Future multimodal evaluations should examine clinically plausible missing-input scenarios explicitly. Relevant analyses include modality dropout during development, prespecified fallback to an independently validated unimodal model where appropriate, explicit abstention, missingness indicators, and analyses of which participants are unable to provide each input. These approaches should be evaluated rather than assumed to preserve performance.

Robustness should be assessed under realistic degradation rather than random input deletion alone. Audio performance may vary with background noise, microphones, language, and recording duration; image models may be affected by phone hardware, illumination, skin appearance, blur, and body site; video-based systems may encounter occlusion, framing differences, and rater disagreement; and sensors may be affected by placement, motion, or data loss. A deployed system should specify whether failed quality control triggers reacquisition, a lower-confidence or indeterminate output, use of an alternative validated model, or human review.

### *5.3. Calibration, Abstention, and Clinical Action*

AUROC measures discrimination across thresholds; it does not establish that a predicted probability is numerically reliable or that a selected threshold improves clinical decisions. Calibration plots, calibration intercept and slope, Brier score, and threshold-specific observed risk can provide complementary information. Calibration should preferably be assessed in temporally or geographically distinct data and reconsidered when prevalence or case mix changes. Post-hoc methods can improve calibration on held-out data [38], as illustrated by isotonic calibration in the multimodal ASD study by Bae et al. [8]. However, recalibration does not correct spectrum bias, outcome misclassification, or missing subgroups.

An indeterminate or abstaining output may be preferable to a forced prediction when acquisition quality is inadequate or available evidence is insufficient. Inconclusive or indeterminate outputs were reported across the Cognoa/Canvas Dx lineage [10–14]. Determinate-result coverage varied substantially: 31.8% in the pivotal study [12], 63.0% in the postmarket analysis [13], and 52.5% in the primary-care implementation [14]. Sensitivity and specificity among determinate cases should be interpreted together with the proportion receiving a determinate result and, where available, the characteristics and outcomes of abstained cases. Conformal prediction can generate prediction sets with stated coverage properties under specified assumptions [39], but it should not be treated as a universal safety mechanism. Its clinical use still depends on the validity of assumptions such as exchangeability, the definition of the target label, and the action taken when prediction sets are broad, ambiguous, or otherwise non-actionable.

This review does not combine discrimination, calibration, latency, privacy, and usability into a single "clinical decision utility" score. These dimensions have different units and involve different assumptions about benefit and harm. A weighted composite can conceal an unacceptable weakness—for example, high AUROC alongside poor calibration in a clinically important subgroup. Aggregation would be more defensible only within a prespecified decision analysis that defines stakeholder utilities, harm trade-offs, and the

intended use. These dimensions are kept separate throughout the synthesis so that strong performance in one area does not obscure an important weakness in another.

### 5.4. External Validity, Equity, and Reporting

Young children should not be treated as a homogeneous subgroup. Age in months, gestational age, developmental stage, language, disability, skin appearance, clinical risk, and caregiver access may affect both the input distribution and the consequences of an error. Device and site can also introduce sources of distribution shift: a model may partly learn a particular camera pipeline, clinical workflow, or referral pattern that does not transfer to another setting. Evaluation design should prevent repeated recordings from the same child from being distributed across development and test sets and, where relevant, should distinguish participant-, household-, site-, and device-level separation.

Reporting guidance can help make these limitations explicit. STARD-AI adds AI-specific items for diagnostic accuracy studies, including dataset practices, evaluation, bias, fairness, and generalizability [40]. TRIPOD+AI addresses the reporting of prediction model studies [41], DECIDE-AI focuses on early-stage live clinical evaluation [42], and CONSORT-AI addresses clinical trials of AI interventions [43]. These guidelines serve different study designs and should be selected according to the intended evaluation. However, none substitute for pediatric representation, an appropriate clinical reference standard, or direct validation in the population and pathway for which the system is intended.

## 6. Deployment Evidence and Child-Centered Governance

### 6.1. Implementation Architecture and Operational Reliability

The charted pediatric studies rarely compared local-only, cloud-only, and hybrid processing as alternative implementation strategies. Several systems acquired data through smartphones or tablets and relied on remote processing, whereas others incorporated device-based interaction or quality control. These reports established the technical feasibility of different processing arrangements but did not show that any particular architecture is clinically superior. Implementation studies should report end-to-end latency, connectivity failures, retry behavior, model versioning, the transmission of raw and derived data, local storage, and the output provided when processing cannot be completed.

The clinically relevant unit is the complete action pathway rather than the physical location of computation. Acquisition, quality assessment, inference, abstention, communication, confirmatory assessment, and follow-up must operate as a coherent workflow. Local processing may support immediate quality checks, whereas remote processing may facilitate centralized model management. Hybrid designs may reduce some data transmission, but this does not in itself establish greater privacy, security, or auditability. Comparative pediatric evidence is needed before one processing architecture can be considered preferable for a given clinical pathway.

### 6.2. Privacy, Caregiver Roles, and Pediatric Rights

Early-childhood systems may collect faces, voices, home environments, routines, and caregiver responses. Such data can reveal information about family members and bystanders as well as the child. WHO guidance emphasizes human autonomy, safety, transparency, accountability, equity, and sustainability across the AI lifecycle [44]. Its guidance on large multimodal models further addresses concerns related to general-purpose systems, secondary uses, and governance at scale [45]. These principles provide an important governance context, but pediatric deployment also requires developmentally appropriate communication, parental permission, age-appropriate assent when relevant, and the explicit consideration of data that may persist beyond childhood.

Child-specific frameworks further define these responsibilities. PEARL-AI organizes ethical recommendations for child-centered medical AI, while explicitly noting that its recommendations are based on Level C evidence and expert opinion rather than large clinical trials [46]. A 2025 pediatric governance review emphasized developmental trajectories, stakeholder engagement, consent and assent, equitable data practices, accountability, and post-deployment monitoring [47]. ACCEPT-AI addresses age, communication, consent and assent, equity, data protection, and technological considerations relevant to pediatric data [48]. These sources can inform study design, documentation, and governance questions, but they should not be interpreted as empirical evidence that a particular pediatric AI system is safe or effective.

Legal requirements also vary by jurisdiction and intended use. Article 8 of the European Union's General Data Protection Regulation addresses a child's consent in relation to information society services [49], while the U.S. Children's Online Privacy Protection Rule governs specified forms of online data collection from children [50]. Medical, research, educational, and consumer uses may be subject to different legal obligations. Manuscripts should describe data flows, users, permissions, retention, deletion, and secondary-use policies, and should avoid inferring legal compliance from technical design alone. Jurisdiction-specific legal review may be required for actual deployment.

### *6.3. Regulation and Readiness*

Regulatory evidence further supports the need for pediatric-specific validation. The limited number of pediatric indications and the sparse explicit representation of pediatric patients in validation data for FDA-authorized AI devices indicate that regulatory labeling alone should not be treated as evidence of validity in early childhood populations [2,3]. For models expected to change after authorization, FDA guidance on predetermined change control plans requires a bounded description of planned modifications, a modification protocol, and an impact assessment [51]. For pediatric use, lifecycle planning should additionally consider age-related distribution shift, the introduction of new device models, changes in language or demographic composition, and the possibility that caregiver or family behavior changes after automated feedback is introduced.

This review does not propose a competing maturity ladder. FUTURE-AI provides international lifecycle guidance for fair, universal, traceable, usable, robust, and explainable healthcare AI [52]. The 2026 Pediatric AI Readiness framework organizes evidence-to-practice considerations specifically for pediatric AI [53]. Together with study design reporting guidance and child-specific ethics frameworks, these resources provide a more established basis for evaluating readiness than a new author-derived maturity scale. Instead, readiness should be supported by linked evidence across technical validity, external clinical validity, human factors testing, workflow effects, safety monitoring, and governance, rather than inferred from the presence of particular architectural components.

## 7. Critical Synthesis by Research Question

### *7.1. RQ1—Tasks, Populations, and Settings*

The charted evidence was concentrated in three non-interchangeable task families. Neonatal jaundice studies estimated or screened bilirubin from optical data during the first days of life. Autism-related studies evaluated behavioral risk, diagnostic support, or objective social visual engagement in toddlers and preschool-aged children. Cry and sleep studies classified short acoustic events or behavioral states. Although many acquisition methods were compatible with decentralized use, much of the validation still occurred under supervised clinical or laboratory conditions. Within the retained evidence

set, caregiver-led home use and downstream referral or monitoring outcomes were less extensively evaluated than model discrimination.

The principal differences in intended decision, available evidence, characteristic failure modes, and near-term evaluation needs across these task families are summarized in Table 4.

**Table 4.** Critical comparison of the principal early childhood task families.

| Task | Intended Decision | Current Evidence | Failure Modes | Near-Term Study |
| --- | --- | --- | --- | --- |
| Neonatal jaundice | • Trigger confirmatory bilirubin testing or follow-up<br>• Not treatment from app output alone | • Prospective TSB comparisons<br>• Threshold-specific metrics<br>• Multicountry or temporal evaluation<br>• One matched multimodal feature analysis | • Lighting variation<br>• Phone-camera variation<br>• Skin appearance<br>• Operator error<br>• Underestimation at high bilirubin | • Caregiver/community acquisition<br>• TSB reference<br>• Threshold-specific harms<br>• Device/site external validation |
| Autism risk or assessment | • Prioritize comprehensive evaluation<br>• Support diagnostic assessment<br>• Quantify behavior | • 11 retained reports<br>• 8 multimodal reports across 3 lineages<br>• Prospective and remote evaluation<br>• Postmarket and implementation evidence | • Lineage dependence<br>• Variable abstention coverage<br>• Language/cultural shift<br>• Missing recordings<br>• Spectrum and referral bias | • Independent prospective pathway comparison<br>• Time to evaluation<br>• Missed cases and false referrals<br>• Coverage and equity |
| Cry classification | • Describe a likely infant state<br>• Flag a recording for review | • Proof-of-concept signal classification<br>• Short labeled clips | • Weak labels<br>• Shared-infant leakage<br>• Noise<br>• Overlapping causes<br>• Microphone and age shift | • Patient-level external cohort<br>• Clinically meaningful labels<br>• Abstention<br>• Caregiver-safety outcomes |
| Sleep/activity monitoring | • Estimate state or trajectory<br>• Potentially flag a change | • Small video/sensor comparisons<br>• Manual-state reference | • Occlusion<br>• Motion<br>• Device placement<br>• Co-sleeping<br>• State-definition disagreement<br>• False alarms | • Longitudinal home evaluation<br>• Prespecified clinical action<br>• Device reliability<br>• User burden |

### *7.2. RQ2—Modalities, Fusion, Alignment, and Missingness*

Images, video or gaze, audio, questionnaires, touch or inertial interaction, and non-contact sensors were represented across the retained studies. Of the 22 reports, 9 jointly modeled distinct patient-level modalities [8,10–17], but these reports represented only four system lineages. The central interpretation is not simply that multimodal evidence was present, but that it was concentrated within a small number of technological lineages. Compared with the earlier seed corpus, the targeted update identified additional direct multimodal evidence, while independent replication across technologies, populations, and care systems remained limited. Fusion level, temporal alignment, determinate result coverage, and missing-modality behavior were not consistently evaluated together.

### *7.3. RQ3—Discrimination, Calibration, Subgroups, and Validity*

Discrimination and threshold-specific sensitivity and specificity dominated performance reporting. Some studies used multicenter, temporal, multicountry, independent, or replication cohorts, but these validation approaches were not consistently applied across

the selected domains. Calibration was explicitly evaluated in the recent multimodal ASD study [8] but was otherwise uncommon. Subgroup and transportability concerns included skin-tone and site variation in bilirubin estimation, age and language dependence in developmental assessment, and device or recording variation across modalities. Few retained studies linked uncertainty or subgroup-specific performance to a prespecified change in referral, reacquisition, abstention, or human review.

### 7.4. RQ4—Implementation, Privacy, Caregiver Use, and Post-Deployment Evidence

Caregiver-led acquisition was directly evaluated for home video, remote multimodal app use, and selected image- or questionnaire-based workflows, but failed capture and unequal input availability were not consistently analyzed. Abstention was particularly consequential in the Cognoa/Canvas Dx lineage, where determinate result coverage varied across validation, postmarket, and implementation reports [11–14]. Prospective workflow evidence within the retained set was concentrated in one small implementation study involving trained primary care clinicians. Measures of privacy exposure, clinician workload, false referral burden, referral completion, and post-deployment drift were uncommon, and processing location was generally described rather than compared. Child-centered governance guidance increasingly specifies relevant design and oversight considerations [46–48], but prospective pediatric evidence that particular governance controls improve clinical or workflow outcomes remains limited.

### 7.5. RQ5—Translation Gaps

Five linked gaps continue to limit translation from model performance to safe screening, assessment, and monitoring practice: insufficient head-to-head comparison of fused and matched unimodal models; incomplete external and prospective validation; limited evaluation of calibration, missing modalities, and subgroups; weak linkage between model uncertainty and prespecified clinical action; and sparse implementation evidence concerning caregivers, workflow, privacy, cost, and post-deployment drift. Within the selected application domains, addressing these gaps is more directly relevant to evidence-to-practice translation than proposing another general architecture or composite readiness score.

### 7.6. Recent Related Literature on Pediatric and Clinical AI

To provide a broader background while keeping the 22-report direct empirical analytic set unchanged, recent related literature was also reviewed. Recent pediatric AI reviews emphasized developmental heterogeneity, limited external validation, provider-facing interpretability, and the need to connect technical performance with pediatric workflow and decision requirements [54,55]. A recent review of affect-aware digital childcare systems similarly linked multimodal sensing, machine learning inference, caregiver or clinician interpretation, and deployment robustness [56]. A broader review of neonatal AI also identified data scarcity, maturational variability, generalizability, and clinical integration as persistent constraints, despite promising technical applications [57].

Recent autism-focused reviews described a shift from conventional screening toward digital phenotyping, multimodal sensing, voice and behavioral analysis, sensor-based assessment, and more clinically interpretable AI [58–61]. Related pediatric ASD studies showed the potential of machine learning and explainable AI while also pointing to continuing concerns about demographic representation, interpretability, and external clinical validation [62,63]. Several of these papers included wider pediatric age ranges or technologies outside the early childhood acquisition criteria used for the direct evidence set. They are used to inform the discussion rather than being added to the 22 retained empirical reports.

Reviews of medical AI also showed why having several types of data is not the same as fusing them in a model. Recent papers described early, intermediate, late, and decision-level fusion [64,65], and stressed the need to explain how information is combined, compare the fused model with appropriate unimodal alternatives, and evaluate whether any improvement matters for the intended clinical decision. Recent studies and reviews of diagnostic AI, explainability, caregiver views, human factors, privacy, and post-deployment monitoring also supported the reporting and evaluation themes developed in the preceding sections [66–71].

Across this literature, common concerns included performance in new settings, human oversight, caregiver understanding and burden, the protection of sensitive data, reliability after deployment, and monitoring over time. These sources are used as contextual support for the discussion and study design recommendations, not as direct evidence that a particular early childhood multimodal system is clinically effective.

## 8. Research Priorities and Decision-Oriented Evaluation

Future studies should begin with an explicit decision statement specifying who uses the output, at what point in the care pathway, and what action may change. Development and evaluation can then be aligned with the consequences of false negatives, false positives, indeterminate results, failed acquisition, and delayed follow-up. Technical performance, operational performance, and downstream consequences should be reported together. For example, an improvement in AUROC may have limited clinical value if it is accompanied by substantially more failed home acquisitions or specialist referrals.

Figure 3 summarizes the decision-oriented logic underlying the research priorities. Multimodal value is not established by fusion or discrimination alone; evidence must connect modality contribution to calibrated and robust outputs, a prespecified clinical action, and observable workflow consequences.

- Priority 1—Matched modality comparisons. Compare the fused model with each constituent modality under identical data splits, preprocessing, and tuning conditions, and quantify both incremental value and additional acquisition burden.
- Priority 2—Prospective multisite pathways. Recruit from the intended screening or assessment population across relevant sites, devices, languages, and socioeconomic contexts, and measure referral completion and time to appropriate assessment.
- Priority 3—Calibration and action thresholds. Predefine clinically relevant thresholds, report calibration and determinate result coverage, and specify the action taken after positive, negative, failed, or indeterminate outputs.
- Priority 4—Missingness and degradation. Evaluate realistic modality absence, poor-quality inputs, network loss, and fallback behavior, including whether failures are concentrated in particular participant or caregiver groups.
- Priority 5—Child and caregiver outcomes. Co-design instructions, consent processes, feedback, and escalation pathways with families and clinicians, and evaluate comprehension, burden, trust, anxiety, and access.
- Priority 6—Lifecycle evidence. Lock or version models during evaluation, document updates, monitor drift and adverse events, and prespecify triggers for recalibration, rollback, or revalidation.

Building on the preceding discussion of modality contribution, calibration, external validity, robustness, human oversight, caregiver factors, privacy, workflow, and lifecycle monitoring, Table 5 brings these reporting and evaluation considerations together in a non-aggregated format. Supporting references are placed directly beside the relevant items so that the basis of each recommendation can be identified without separating the recommendation from its source.

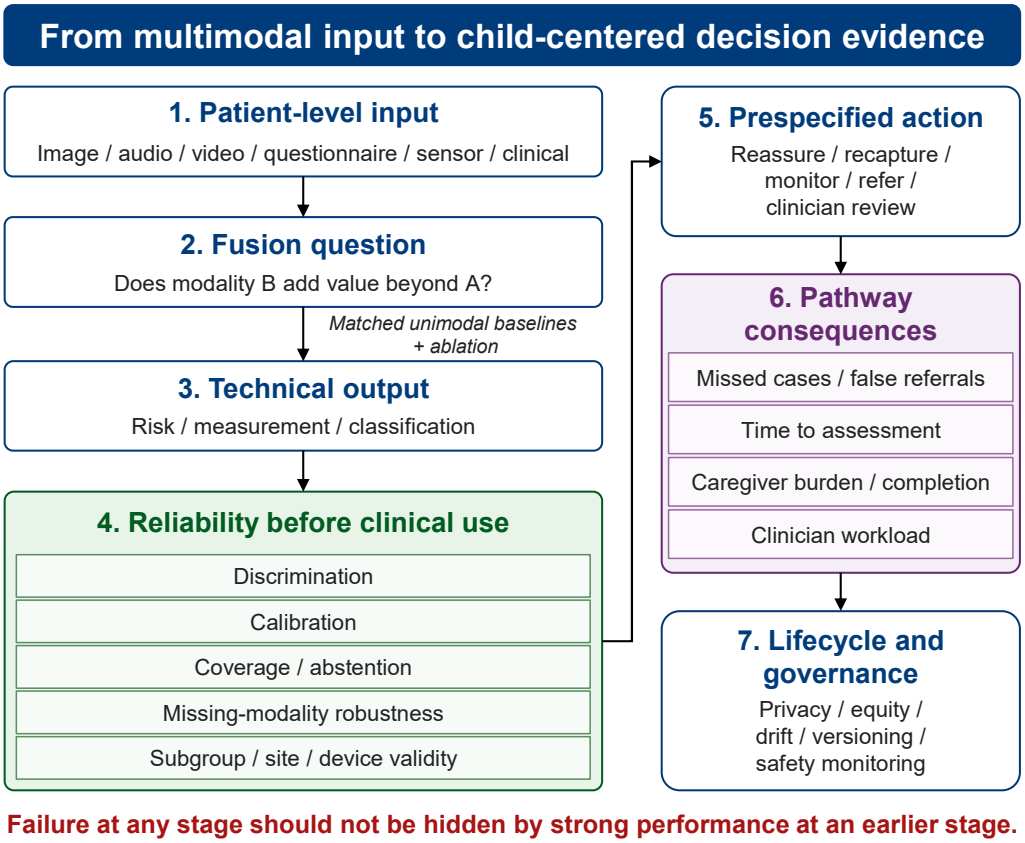



**Figure 3.** Decision-oriented pathway for evaluating multimodal early-childhood AI. Multimodal value should first be demonstrated against matched unimodal alternatives and then followed through reliability, clinical action, workflow consequences, and lifecycle governance. Strong discrimination at an earlier stage does not compensate for poor calibration, low determinate-result coverage, missing-modality failure, subgroup transportability, or an unsafe downstream pathway. The figure is a synthesis and study design aid, not a validated readiness score or certification framework.

**Table 5.** Reporting and deployment considerations for early-childhood multimodal AI. Supporting references are shown directly beside the reporting items and evaluation approaches they inform.

| Dimension | Minimum Report | Decision Relevance | Suggested Evaluation |
|---|---|---|---|
| Discrimination [40,41] | • AUROC/AUPRC with confidence intervals<br>• Threshold sensitivity and specificity | • Separates higher- and lower-risk cases | • Participant-level test set<br>• Prevalence-appropriate precision–recall analysis |
| Calibration [38,41] | • Calibration plot<br>• Intercept/slope<br>• Brier score<br>• Recalibration procedure | • Assesses whether stated probabilities support threshold-based decisions | • Temporal or external cohort<br>• Report before and after recalibration |
| Clinical consequences [41–43,53] | • Missed cases<br>• False referrals<br>• Indeterminate results<br>• Confirmatory tests | • Expresses benefits and harms in pathway units | • Decision curve or prospective pathway analysis<br>• Prespecified utilities |
| External validity [40,41,52,53] | • Sites/countries<br>• Devices/operators<br>• Languages<br>• Time periods | • Assesses transport to intended use | • Geographic and temporal validation<br>• No model retuning where appropriate<br>• Targeted local validation |

**Table 5.** *Cont.*

| Dimension | Minimum Report | Decision Relevance | Suggested Evaluation |
| --- | --- | --- | --- |
| Modality contribution [5–8,17,64,65] | • Matched unimodal baselines<br>• Ablation of each modality | • Tests whether fusion justifies added burden | • Same cohort and data split<br>• Same preprocessing and tuning budget |
| Missing-modality robustness [8,10–16,64,65] | • Availability rate<br>• Reason for absence<br>• Fallback or abstention behavior<br>• Coverage | • Identifies complete case optimism and hidden exclusion | • Realistic modality dropout<br>• Quality degradation<br>• Subgroup-specific failure analysis |
| Subgroup performance [20–24,40,46,48] | • Age/developmental stage<br>• Sex<br>• Skin appearance<br>• Language/disability<br>• Access variables, as relevant | • Identifies unequal error and acquisition burden | • Prespecified stratified or interaction analysis<br>• Report uncertainty<br>• Avoid unsupported ranking |
| Operational reliability [42,51–53,71] | • Capture failures<br>• Latency<br>• Data volume<br>• Battery/compute demand<br>• Network loss behavior<br>• Model version | • Assesses whether the system completes the intended action pathway | • Bench and field testing<br>• Representative devices and connectivity |
| Privacy and security [44–50,52,70] | • Raw/derived data flow<br>• Retention/access<br>• Secondary use<br>• Incident response | • Addresses risks to children, families, and bystanders | • Threat modeling<br>• Privacy review<br>• Security testing<br>• Jurisdiction-specific legal review |
| Human oversight [42,44–46,52,67,69] | • Intended users<br>• Recapture/review triggers<br>• Override/escalation<br>• Accountability | • Links uncertainty and error to a defined human response | • Human factors simulation<br>• Prospective clinical evaluation |
| Caregiver usability [46,53,68,69] | • Completion/time<br>• Comprehension<br>• Burden/anxiety<br>• Accessibility<br>• Digital exclusion | • Affects data quality, adherence, and equity | • Co-design<br>• Usability testing<br>• Qualitative analysis<br>• Representative field study |
| Workflow and safety [42,43,51–53,71] | • Referral completion<br>• Workload<br>• Time to assessment<br>• Adverse events<br>• Drift/update plan | • Evaluates effects beyond model performance | • Implementation study<br>• Continuous monitoring<br>• Specified rollback/revalidation triggers |

AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve. The dimensions are kept separate rather than combined into one weighted score. Any combined score would require a prospective decision analysis with explicit stakeholder utilities and prespecified harm trade-offs. References beside individual items identify the literature used to support those specific reporting or evaluation considerations; they do not imply that every item has already been prospectively validated in early-childhood care.

Table 5 is intended as a practical guide to reporting and study design, not as a validated scale, certification checklist, or clinical readiness score. It draws on established guidance, including STARD-AI [40], TRIPOD+AI [41], DECIDE-AI [42], CONSORT-AI [43], ACCEPT-

AI [48], FUTURE-AI [52], and the Pediatric AI Readiness framework [53], together with direct pediatric evidence and multimodal methods literature. Keeping the dimensions separate helps make important gaps visible instead of allowing strong performance in one area to hide a weakness in another.

## 9. Limitations

This review has several limitations. First, it is a structured narrative review with a documented targeted search update, not a systematic or scoping review. The complete original search record was unavailable, and the update did not include a preregistered protocol, exhaustive database export, duplicate independent screening, a PRISMA flow, or formal risk-of-bias assessment. The 22-report evidence map addressed material omissions identified in the seed corpus but does not establish field-wide prevalence or demonstrate exhaustive retrieval. The modality classifications of all 22 retained reports were also checked again against the criteria in Table 1 and the source articles. This additional check improved consistency, but it cannot remove selection and interpretation bias from the original single-reviewer process. The reported counts should be interpreted as the analytic set used in this review rather than as independently screened prevalence estimates. A future scoping review could prospectively register its protocol, preserve complete search and deduplication records, use duplicate study selection and charting where feasible, and report a complete record of included and excluded studies.

Second, the retained studies span heterogeneous conditions, populations, reference standards, thresholds, and intended uses. Quantitative pooling and direct performance ranking were inappropriate. Neonatal jaundice screening, autism risk stratification, clinician-supervised diagnostic support, cry-reason classification, and sleep state monitoring should not be interpreted as a single clinical performance category. Some audio studies also lacked a clearly reported infant-level sampling frame, which limits the clinical interpretation of their reported performance.

Third, the definition of modality is intentionally conservative. It classifies conceptually distinct patient-level information channels rather than file formats, multiple views, or derived representations. Some authors may describe questionnaire subscales, multiple image regions, or spectrograms as separate modalities; the definition used here avoids inflating multimodal counts but must be applied consistently when interpreting or replicating this review.

Fourth, within the retained evidence set, direct pediatric empirical evidence was more developed for task-specific performance than for comparative architecture, missing-modality management, privacy engineering, or prospective workflow evaluation. These latter topics were informed partly by methodological, reporting, regulatory, and policy sources. Their implications are treated as design or evaluation considerations rather than as observed pediatric clinical effects.

Finally, several reports evaluated successive versions or deployments of the same system lineage. Although report and system-lineage counts were presented separately, partial cohort reuse, shared development history, company-affiliated authorship, and manufacturer-held postmarket data can reduce evidentiary independence. The search cutoff of 29 July 2026 captured recent empirical, reporting, governance, readiness, and regulatory sources available within the targeted update, but subsequent publications may alter the balance of evidence. Product characteristics, legal requirements, and regulatory information may also change and should be checked against the applicable version of record at the time of submission or deployment.

## 10. Conclusions

Within the selected early-childhood application domains reviewed here, the evidence remains task-specific. Smartphone imaging for neonatal jaundice, video- or eye-tracking-based approaches for autism-related screening or assessment, and audio- or sensor-based monitoring address different clinical or caregiving decisions, and should not be combined into a single performance category. Under the patient-level definition used in this review, 9 of the 22 retained reports used multimodal fusion, representing four system lineages—one in neonatal jaundice and three in autism-related applications. Within the retained autism-related reports, evidence included prospective controlled evaluation, primary-care implementation, remote caregiver administration, postmarket analysis, and workflow evaluation. These reports also illustrated why determinate result coverage and evidentiary independence matter. High sensitivity or specificity among determinate cases can coexist with substantial abstention, and multiple publications from the same system lineage do not substitute for independent replication. Across the selected domains, key translational needs include matched comparisons of fused and unimodal models, independent multisite validation, calibration and abstention linked to prespecified actions, realistic missing modality and subgroup evaluation, caregiver and workflow outcomes, and lifecycle monitoring. Local, cloud, and hybrid processing should be evaluated as implementation strategies rather than assumed to constitute validated pediatric architectures.

Near-term progress may depend more on generating child-centered evidence that connects model outputs to clinically or operationally meaningful decisions than on proposing additional general frameworks. Established reporting, ethics, governance, and readiness guidance can help structure this work. The reporting matrix in Table 5 keeps discrimination, calibration, external validity, robustness, equity, privacy, caregiver usability, workflow, and safety as separate considerations so that strength in one area does not hide an important limitation in another. The practical implication is not that multimodal systems are inherently better, but that each added modality should show useful added value, remain reliable across relevant settings, and contribute to a meaningful clinical or caregiving decision.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16189128/s1, Table S1: Targeted search update and source-retention log; Table S2: Report-level characteristics and decision-relevant evidence appraisal of the analytically charted pediatric empirical studies.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Acknowledgments:** During the preparation of this manuscript, the author used OpenAI's ChatGPT (GPT-5.6 Sol) to assist with developing preliminary concepts and layouts for the figures. All final figures were manually created and edited by the author using Microsoft PowerPoint. The author reviewed and revised all AI-assisted outputs and takes full responsibility for the content of this publication. The author sincerely thanks Seungmin Rho of Chung-Ang University and Hyeonwoo Kim of Soonchunhyang University for their critical review of the manuscript and constructive comments that helped improve its clarity and scholarly presentation.

**Conflicts of Interest:** The author declares no conflicts of interest.

## References


1. Lipkin, P.H.; Macias, M.M.; Council on Children with Disabilities; Section on Developmental and Behavioral Pediatrics. Promoting optimal development: Identifying infants and young children with developmental disorders through developmental surveillance and screening. *Pediatrics* **2020**, *145*, e20193449. [CrossRef] [PubMed]
2. Zapotoczny, G.; Goyal, A.; Christmas, M.; Qazi, S.; Carroll, M.; Espinoza, J. FDA-regulated AI-enabled medical devices with pediatric indications. *JAMA Netw. Open* **2026**, *9*, e262636. [CrossRef] [PubMed]
3. Brewster, R.C.L.; Nagy, M.; Wunnava, S.; Bourgeois, F.T. US FDA approval of pediatric artificial intelligence and machine learning-enabled medical devices. *JAMA Pediatr.* **2025**, *179*, 212–214. [CrossRef] [PubMed]
4. Wang, Q.; Yin, J.; Zhang, X.; Ou, H.; Li, F.; Zhang, Y.; Wan, W.; Guo, C.; Cao, Y.; Luo, T.; et al. Applications of artificial intelligence in early childhood health management: A systematic review from fetal to pediatric periods. *Front. Pediatr.* **2025**, *13*, 1613150. [CrossRef] [PubMed]
5. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
6. Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; Luo, Y. Multimodal machine learning in precision health: A scoping review. *npj Digit. Med.* **2022**, *5*, 171. [CrossRef] [PubMed]
7. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E.J. Multimodal biomedical AI. *Nat. Med.* **2022**, *28*, 1773–1784. [CrossRef] [PubMed]
8. Bae, S.; Hong, J.; Ha, S.; Moon, J.; Yu, J.; Choi, H.; Lee, J.; Do, R.; Sim, H.; Kim, H.; et al. Multimodal AI for risk stratification in autism spectrum disorder: Integrating voice and screening tools. *npj Digit. Med.* **2025**, *8*, 538. [CrossRef] [PubMed]
9. Baethge, C.; Goldbeck-Wood, S.; Mertens, S. SANRA—A scale for the quality assessment of narrative review articles. *Res. Integr. Peer Rev.* **2019**, *4*, 5. [CrossRef] [PubMed]
10. Abbas, H.; Garberson, F.; Glover, E.; Wall, D.P. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1000–1007. [CrossRef] [PubMed]
11. Abbas, H.; Garberson, F.; Liu-Mayo, S.; Glover, E.; Wall, D.P. Multi-modular AI approach to streamline autism diagnosis in young children. *Sci. Rep.* **2020**, *10*, 5014. [CrossRef] [PubMed]
12. Megerian, J.T.; Dey, S.; Melmed, R.D.; Coury, D.L.; Lerner, M.; Nicholls, C.J.; Sohl, K.; Rouhbakhsh, R.; Narasimhan, A.; Romain, J.; et al. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. *npj Digit. Med.* **2022**, *5*, 57. [CrossRef] [PubMed]
13. Salomon, C.; Heinz, K.; Aronson-Ramos, J.; Wall, D.P. An analysis of the real world performance of an artificial intelligence based autism diagnostic. *Sci. Rep.* **2025**, *15*, 29503. [CrossRef] [PubMed]
14. Sohl, K.; Linstead, E.; Heinz, K.; Eiroa Lledo, E.; Brewer Curran, A.; Mahurin, M.; Nanclares-Nogués, V.; Salomon, C.; Seal, M.; Taraman, S. Integration of an artificial intelligence-based autism diagnostic device into the ECHO Autism primary care workflow: Prospective observational study. *JMIR Form. Res.* **2025**, *9*, e80733. [CrossRef] [PubMed]
15. Perochon, S.; Di Martino, J.M.; Carpenter, K.L.H.; Compton, S.; Davis, N.; Eichner, B.; Espinosa, S.; Franz, L.; Krishnappa Babu, P.R.; Sapiro, G.; et al. Early detection of autism using digital behavioral phenotyping. *Nat. Med.* **2023**, *29*, 2489–2497. [CrossRef] [PubMed]
16. Krishnappa Babu, P.R.; Di Martino, J.M.; Aiello, R.; Eichner, B.; Espinosa, S.; Green, J.; Howard, J.; Perochon, S.; Spanos, M.; Vermeer, S.; et al. Validation of a mobile app for remote autism screening in toddlers. *NEJM AI* **2024**, *1*, AIcs2400510. [CrossRef] [PubMed]
17. Liang, Y.; Zou, L.; Goh, M.M.R.; Ngeow, A.J.H.; Tan, N.C.; Ta, A.W.A.; Goh, H.L. Multimodal feature analysis for automated neonatal jaundice assessment using machine learning. *JAMIA Open* **2025**, *8*, ooaf165. [CrossRef] [PubMed]
18. Taylor, J.A.; Stout, J.W.; de Greef, L.; Goel, M.; Patel, S.; Chung, E.K.; Koduri, A.; McMahon, S.; Dickerson, J.; Simpson, E.A.; et al. Use of a smartphone app to assess neonatal jaundice. *Pediatrics* **2017**, *140*, e20170312. [CrossRef] [PubMed]
19. Ngeow, A.J.H.; Tan, M.G.; Dong, X.; Tagamolila, V.; Ereno, I.; Tay, Y.Y.; Xin, X.; Poon, W.B.; Yeo, C.L. Validation of a smartphone-based screening tool (Biliscan) for neonatal jaundice in a multi-ethnic neonatal population. *J. Paediatr. Child Health* **2023**, *59*, 288–297. [CrossRef] [PubMed]
20. Ngeow, A.J.H.; Moosa, A.S.; Tan, M.G.; Zou, L.; Goh, M.M.R.; Lim, G.H.; Tagamolila, V.; Ereno, I.; Durnford, J.R.; Cheung, S.K.H.; et al. Development and validation of a smartphone application for neonatal jaundice screening. *JAMA Netw. Open* **2024**, *7*, e2450260. [CrossRef] [PubMed]
21. Enweronu-Laryea, C.; Leung, T.; Outlaw, F.; Brako, N.O.; Insaidoo, G.; Hagan-Seneadza, N.A.; Ani-Amponsah, M.; Nixon-Hill, M.; Meek, J. Validating a sclera-based smartphone application for screening jaundiced newborns in Ghana. *Pediatrics* **2022**, *150*, e2021053600. [CrossRef] [PubMed]
22. Jiménez-Díaz, G.; Gierman, L.M.; Keitsch, M.; Elizarrarás-Rivas, J.; Silva-Méndez, R.M.; Infanti, J.J.; Marcuzzi, A. Validation of a mobile health device for neonatal jaundice screening: A cross-sectional study in a resource-limited setting in Mexico. *SAGE Open Pediatr.* **2025**, *12*, 30502225251320544. [CrossRef] [PubMed]

23. Aune, A.; Jiménez-Díaz, G.; Gierman, L.M.; Vartdal, G.; Reyes Berlanga, M.; Tusoy, J.; Bergseng, H.; Shakya, S.; Darj, E. Smartphone-based screening of neonatal jaundice in three populations in low- and middle-income countries: A cross-sectional study. *BMJ Paediatr. Open* **2025**, *9*, e002242. [CrossRef] [PubMed]
24. Gülcan Kersin, S.; Ünal, İ.; Toprak Sevinç, S.; Ozdemir, H.; Memişoğlu, A.; Bilgen, H.S. Diagnostic accuracy of the Picterus jaundice Pro app for neonatal jaundice: A single-center study from Turkiye. *Sci. Rep.* **2026**, *16*, 10205. [CrossRef] [PubMed]
25. Hegde, D.; Rath, C.; Amarasekara, S.; Saraswati, C.; Patole, S.; Rao, S. Performance of smartphone application to accurately quantify hyperbilirubinemia in neonates: A systematic review with meta-analysis. *Eur. J. Pediatr.* **2023**, *182*, 3957–3971. [CrossRef] [PubMed]
26. Tariq, Q.; Daniels, J.; Schwartz, J.N.; Washington, P.; Kalantarian, H.; Wall, D.P. Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS Med.* **2018**, *15*, e1002705. [CrossRef] [PubMed]
27. Jones, W.; Klaiman, C.; Richardson, S.; Aoki, C.; Smith, C.; Minjarez, M.; Bernier, R.; Pedapati, E.; Bishop, S.; Ence, W.; et al. Eye-tracking-based measurement of social visual engagement compared with expert clinical diagnosis of autism. *JAMA* **2023**, *330*, 854–865. [CrossRef] [PubMed]
28. Jones, W.; Klaiman, C.; Richardson, S.; Lambha, M.; Reid, M.; Hamner, T.; Beacham, C.; Lewis, P.; Paredes, J.; Edwards, L.; et al. Development and replication of objective measurements of social visual engagement to aid in early diagnosis and assessment of autism. *JAMA Netw. Open* **2023**, *6*, e2330145. [CrossRef] [PubMed]
29. U.S. Food and Drug Administration. *De Novo Classification Request for Cognoa ASD Diagnosis Aid (DEN200069): Decision Summary*; U.S. Food and Drug Administration: Silver Spring, MD, USA, 2021. Available online: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN200069.pdf (accessed on 29 July 2026).
30. Ponzo, S.; May, M.; Tamayo-Elizalde, M.; Bailey, K.; Shand, A.J.; Bamford, R.; Multmeier, J.; Griessel, I.; Szulyovszky, B.; Blakey, W.; et al. App characteristics and accuracy metrics of available digital biomarkers for autism: Scoping review. *JMIR mHealth uHealth* **2023**, *11*, e52377. [CrossRef] [PubMed]
31. Ashwini, K.; Vincent, P.M.D.R.; Srinivasan, K.; Chang, C.-Y. Deep learning assisted neonatal cry classification via support vector machine models. *Front. Public Health* **2021**, *9*, 670352. [CrossRef] [PubMed]
32. Hammoud, M.; Getahun, M.N.; Baldycheva, A.; Somov, A. Machine learning-based infant crying interpretation. *Front. Artif. Intell.* **2024**, *7*, 1337356. [CrossRef] [PubMed]
33. Ji, C.; Mudiyanselage, T.B.; Gao, Y.; Pan, Y. A review of infant cry analysis and classification. *EURASIP J. Audio Speech Music Process.* **2021**, *2021*, 8. [CrossRef]
34. Wang, Y.; Williams, R.; Dilley, L.C.; Houston, D.M. A meta-analysis of the predictability of LENA™ automated measures for child language development. *Dev. Rev.* **2020**, *57*, 100921. [CrossRef] [PubMed]
35. Long, X.; Otte, R.; van der Sanden, E.; Werth, J.; Tan, T. Video-based actigraphy for monitoring wake and sleep in healthy infants: A laboratory study. *Sensors* **2019**, *19*, 1075. [CrossRef] [PubMed]
36. Schoch, S.F.; Kurth, S.; Werner, H. Actigraphy in sleep research with infants and young children: Current practices and future benefits of standardized reporting. *J. Sleep Res.* **2021**, *30*, e13134. [CrossRef] [PubMed]
37. Lee, W.H.; Kim, S.H.; Na, J.Y.; Lim, Y.-H.; Cho, S.H.; Cho, S.H.; Park, H.-K. Non-contact sleep/wake monitoring using impulse-radio ultrawideband radar in neonates. *Front. Pediatr.* **2021**, *9*, 782623. [CrossRef] [PubMed]
38. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Proceedings of Machine Learning Research; Volume 70, pp. 1321–1330.
39. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [CrossRef]
40. Sounderajah, V.; Guni, A.; Liu, X.; Collins, G.S.; Karthikesalingam, A.; Markar, S.R.; Golub, R.M.; Denniston, A.K.; Shetty, S.; Moher, D.; et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **2025**, *31*, 3283–3289. [CrossRef] [PubMed]
41. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]
42. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [CrossRef] [PubMed]
43. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef] [PubMed]
44. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health*; World Health Organization: Geneva, Switzerland, 2021. Available online: https://www.who.int/publications/i/item/9789240029200 (accessed on 29 July 2026).

45. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*; World Health Organization: Geneva, Switzerland, 2024. Available online: https://www.who.int/publications/i/item/9789240084759 (accessed on 29 July 2026).
46. Chng, S.Y.; Tern, M.J.W.; Lee, Y.S.; Cheng, L.T.-E.; Kapur, J.; Eriksson, J.G.; Chong, Y.S.; Savulescu, J. Ethical considerations in AI for child health and recommendations for child-centered medical AI. *npj Digit. Med.* **2025**, *8*, 152. [CrossRef] [PubMed]
47. Richter, F.; Holmes, E.; Richter, F.; Guttmann, K.; Duong, S.Q.; Gangadharan, S.; Schadt, E.E.; Salmasian, H.; Gelb, B.D.; Glicksberg, B.S. Toward governance of artificial intelligence in pediatric healthcare. *npj Digit. Med.* **2025**, *8*, 636. [CrossRef] [PubMed]
48. European Parliament; Council of the European Union. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *L119*, 1–88. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 (accessed on 29 July 2026).
49. U.S. Federal Trade Commission. Children's Online Privacy Protection Rule ("COPPA"). Available online: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa (accessed on 29 July 2026).
50. Muralidharan, V.; Burgart, A.; Daneshjou, R.; Rose, S. Recommendations for the use of pediatric data in artificial intelligence and machine learning: ACCEPT-AI. *npj Digit. Med.* **2023**, *6*, 166. [CrossRef] [PubMed]
51. U.S. Food and Drug Administration. *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff*; U.S. Food and Drug Administration: Silver Spring, MD, USA, 2025. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence (accessed on 29 July 2026).
52. Lekadir, K.; Frangi, A.F.; Porras, A.R.; Glocker, B.; Cintas, C.; Langlotz, C.P.; Weicken, E.; Asselbergs, F.W.; Prior, F.; Collins, G.S.; et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **2025**, *388*, e081554. [CrossRef] [PubMed]
53. Sezgin, E.; Boch, S. The pediatric AI readiness framework: Bridging evidence to practice in pediatric artificial intelligence. *Front. Artif. Intell.* **2026**, *9*, 1800047. [CrossRef] [PubMed]
54. Ganatra, H.A. Machine Learning in Pediatric Healthcare: Current Trends, Challenges, and Future Directions. *J. Clin. Med.* **2025**, *14*, 807. [CrossRef] [PubMed]
55. Mendoza Bolivar, L.E.; Satzer, M. Essential Concepts in Artificial Intelligence: A Guide for Pediatric Providers. *Children* **2025**, *12*, 1386. [CrossRef] [PubMed]
56. Moon, J. A Review of Affect-Aware Digital Childcare Systems for Early Screening and Personalized Intervention. *J. Platf. Technol.* **2026**, *14*, 3–20. [CrossRef]
57. Chioma, R.; Sbordone, A.; Patti, M.L.; Perri, A.; Vento, G.; Nobile, S. Applications of Artificial Intelligence in Neonatology. *Appl. Sci.* **2023**, *13*, 3211. [CrossRef]
58. Luo, W.; Yin, Z.; Dai, J. Early Detection Methods for Autism Spectrum Disorder: From Clinical Screening to Multimodal AI. *Diagnostics* **2026**, *16*, 2376. [CrossRef] [PubMed]
59. Zgheib, R.; El Naggar, A.; Kermani Kolankeh, A.; Takshe, A.A. Toward Trustworthy AI for Autism Spectrum Disorder: A Systematic Review of Multimodal Systems, Knowledge Representation, and Clinical Integration. *Information* **2026**, *17*, 802. [CrossRef]
60. Rakotomanana, H.; Rouhafzay, G. A Scoping Review of AI-Based Approaches for Detecting Autism Traits Using Voice and Behavioral Data. *Bioengineering* **2025**, *12*, 1136. [CrossRef] [PubMed]
61. Bouchouras, G.; Kotis, K. Integrating Artificial Intelligence, Internet of Things, and Sensor-Based Technologies: A Systematic Review of Methodologies in Autism Spectrum Disorder Detection. *Algorithms* **2025**, *18*, 34. [CrossRef]
62. Jeon, I.; Kim, M.; So, D.; Kim, E.Y.; Nam, Y.; Kim, S.; Shim, S.; Kim, J.; Moon, J. Reliable Autism Spectrum Disorder Diagnosis for Pediatrics Using Machine Learning and Explainable AI. *Diagnostics* **2024**, *14*, 2504. [CrossRef] [PubMed]
63. Moon, J.; Woo, J. CARE: Comprehensive Artificial Intelligence Techniques for Reliable Autism Evaluation in Pediatric Care. *Comput. Mater. Contin.* **2025**, *85*, 1383–1425. [CrossRef]
64. Xu, X.; Li, J.; Zhu, Z.; Zhao, L.; Wang, H.; Song, C.; Chen, Y.; Zhao, Q.; Yang, J.; Pei, Y. A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering* **2024**, *11*, 219. [CrossRef] [PubMed]
65. Kumar, S.; Rani, S.; Sharma, S.; Min, H. Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review. *Bioengineering* **2024**, *11*, 1233. [CrossRef] [PubMed]
66. Bartusik-Aebisher, D.; Justin Raj, D.R.; Aebisher, D. Artificial Intelligence in Medical Diagnostics: Foundations, Clinical Applications, and Future Directions. *Appl. Sci.* **2026**, *16*, 728. [CrossRef]
67. Eje, O.A.; Azim, S.M.; Dehzangi, I. Explainable AI Applications in Healthcare: A Systematic Review. *Algorithms* **2026**, *19*, 488. [CrossRef]

68. Berghea, E.C.; Ionescu, M.D.; Gheorghiu, R.M.; Tincu, I.F.; Cobilinschi, C.O.; Craiu, M.; Bălgrădean, M.; Berghea, F. Integrating Artificial Intelligence in Pediatric Healthcare: Parental Perceptions and Ethical Implications. *Children* **2024**, *11*, 240. [CrossRef] [PubMed]
69. Alharbi, O.M. Human Factors in Technology-Mediated Autism Intervention: A Caregiver-Centered Framework. *Appl. Sci.* **2026**, *16*, 4218. [CrossRef]
70. Croitoru, I.; Turcu, C.E.; Turcu, C.O. Privacy-by-Design in AI-Assisted Systems for Caregivers of Children with Autism: A Secure Multi-Agent Architecture. *Appl. Sci.* **2026**, *16*, 2157. [CrossRef]
71. El Arab, R.A.; Hussein Mustafa, M.; Almagharbeh, W.T.; Ayoub, M.Y.; Alsanawi, F.; Almathen, F.; Almosabeh, R.; Al Talaq, M. Clinical AI Beyond Development: A Scoping Review of Deployment-Related Robustness, Algorithmovigilance, and Lifecycle Oversight. *Healthcare* **2026**, *14*, 2052. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.