

“Toward a Standardized Developmental Staging Framework for Pediatric Artificial Intelligence”

Esanika Chauhan

Year 13 | Charterhouse school, United Kingdom

ABSTRACT

Artificial intelligence systems are increasingly used by and applied to children, in clinical care and in general use. How these systems represent developmental stage determines whether evidence of their safety in one group transfers to another. This review aims to examine how developmental stage is currently represented across clinical, regulatory, and conversational applications of pediatric artificial intelligence, identifies gaps arising from those representations, and proposes requirements for a staging standard.

Literature was identified through documented searches of PubMed, covering 2015 to 2026, supplemented by citation chaining to computer science venues not indexed by pub Med. Developmental stage is represented differently by each research community and consistently inadequately. Adult trained segmentation models underperform in children, achieving a mean Dice similarity coefficient of 0.73 compared with 0.81 in adults ($p < 0.001$). With the largest deficits occurring in the smallest anatomical structures period of 952 artificial intelligence enabled device submissions, 4.4% included pediatric age ranges. Among 101 high risk devices, 25.7% specified no age range, and only 7.9% used recommended pediatric age ranges, which are non-binding and historically rather than developmentally grounded. Evaluations of large language models for pediatric use do not stratify findings by developmental stage, while research defining developmental stages for artificial intelligence systems relies on simulated rather than real children. Five gaps emerge: the absence of a binding definition of developmental stage, evidence drawn predominantly from sources other than children, the lack of a validated method for determining developmental stage, evaluation periods shorter than the development tool changes covered by validity claims, and separation between the clinical and conversational literatures.

Four requirements are proposed, none of which has been tested yet: representing developmental state as a continuous covariate while preserving discrete bands for reporting; justifying stage boundaries through measured differences in outcomes; establishing an explicit rule that defaults to the least mature possible developmental state under uncertainty while minimizing the information used to determine that state; and requiring validity claims to specify whether they are based on single occasion or developmentally extended evaluation. Quantified pediatric performance deficits and proposed developmental stages to exist in the literature but have not been brought together. Establishing whether matching system behavior developmental stage improves outcomes for real children is a prerequisite for validating the proposed framework.

INTRODUCTION

Artificial Intelligence (AI) is now embedded in systems that children encounter in clinics, classrooms, and at home. In pediatric medicine, AI supports diagnosis, treatment planning, and prognosis [1]. Large language models (LLMs) are entering clinical documentation and decision support. Outside the clinic, LLMs are being evaluated as sources of medical information for families managing common pediatric conditions such as asthma [2]. Hence, the central questions are no longer whether children will interact with these systems, but whether the systems are built to account for who children are.

Drug disposition and drug action change as organ systems mature, so pediatric dosing is governed by developmental physiology instead of treating children as scaled-down adults [3]. This same reasoning also applies to AI. Children differ from adults in anatomy, disease distribution, language, and reasoning. These differences change across childhood itself. Reviews of pediatric AI identify variable developmental status, alongside limited pediatric data and ethical concerns, as a distinct obstacle to deploying AI in children [1].

Empirical evidence indicates that models developed on adult data do not transfer reliably to pediatric populations. A computerized tomography (CT) organ segmentation model trained on adults achieved a significantly lower mean Dice score on pediatric scans than on adult scans (0.73 vs 0.81, $p < 0.001$), with the largest deficits in small structures such as the adrenal glands [4]. Training on pediatric data or fine-tuning the adult model recovered much of that performance.

Segmentation frameworks that explicitly condition on demographic attributes, including age, report reduced disparities between pediatric and adult groups [5]. This implies that age-blind architectures carry avoidable errors.

Regulatory evidence suggests that, in a cross-sectional analysis of 952 AI-enabled device submissions reviewed by the United States Food and Drug Administration (FDA) between November 1995 and June 2024, only 42 of 952 (4.4%) included pediatric age ranges of 0 to 17 years [6]. Furthermore, only five exclusively pediatric technologies were introduced between 2020 and 2024 [6]. Pediatric submissions also had a significantly longer median review time, 162 days, than non-pediatric submissions, 134 days ($p = 0.049$) [6]. The authors of this analysis conclude that the FDA should standardize age labeling and validation requirements for AI-enabled technologies [6]. A parallel analysis of FDA authorization of pediatric AI-enabled devices also reaches related conclusions [7].

Pediatric status is represented as a label or an age range rather than as a developmental state, and the age ranges themselves are inconsistent across devices, benchmarks, and regulatory frameworks. A single category spanning birth to seventeen years groups a preverbal infant with an adolescent, obscuring the variation that drives model error and interaction risk. Shared developmental staging standards that make results comparable across clinical AI, conversational AI, and policy are currently limited.

This review argues that pediatric AI requires such a standard. This review examines three implicit assumptions in current practice: (1) chronological age is an adequate proxy for developmental capacity (2) models trained on adult data transfer to children (3) a single set of pediatric safeguards suffices across childhood. It then analyzes gaps in current standards, including inconsistent stage definitions, reliance on simulated rather than real child participants, and the absence of validated methods for inferring developmental stage, limited longitudinal studies, and the separation between clinical and conversational AI research. Finally, the review proposes a banded developmental staging framework with an explicit rule for handling stage uncertainty and outlines testable studies to evaluate it.

METHODS

2.1 Review Design

This article presents a narrative review supported by a documented search strategy. A narrative approach was chosen over a systematic or scoping review for two reasons.

First, relevant research spans multiple fields with distinct publication and indexing practices. Clinical applications of pediatric AI are primarily indexed in PubMed, whereas studies of conversational AI used by children are often published in computer science venues and preprint servers. Relying on a single database would therefore exclude substantial portions of the literature.

Second, terminology differs across these fields. Preliminary searches found that “development stage” in biomedical literature primarily refers to cellular or neurological development rather than cognitive development of child users. Among the ten most recent records retrieved using a targeted search combining AI, development stage, and pediatric population terms, none addressed the developmental staging of child users. Because existing controlled vocabularies do not distinguish these meanings, database search alone cannot reliably achieve both high recall and high precision.

Given these limitations, reporting screening results as a complete selection process could overstate the comprehensiveness of the search. The review therefore reports its search strategy, sources, and retrieval counts, explicitly distinguishing evidence identified through database searches from evidence identified through other methods. Reporting follows the PRISMA 2020 recommendations for transparent description of search methods, without claiming adherence to the PRISMA-ScR checklist intended for reviews with a documented selection process [8, 9, 10].

2.2 Information Sources and Search Execution

PubMed was searched systematically using the strings reported in Table 1. All searches were restricted to the period January 2015 to October 2026. The lower bound was set at 2015 because it precedes the first clearance by the United States FDA of an AI-enabled device whose labeling included pediatric age ranges [6].

PubMed does not index computer science or human-computer interaction venues, so evidence from those fields was identified by targeted retrieval rather than by systematic search. Evidence from those fields was identified through forward and backward citation chaining from records retrieved in PubMed, and through direct retrieval of works cited in the policy and review literature.

This is an acknowledged asymmetry between the two bodies of evidence and is treated as a limitation in Section 8 rather than presented as equivalent coverage.

Table 1. Search strings executed in PubMed with records retrieved. All searches were limited to 2015-2026. Searches were conducted in October 2026.

#	String (concept blocks combined with AND)	Records
1	(child OR children OR pediatric OR paediatric OR adolescent OR infant) AND (artificial intelligence OR machine learning OR deep learning OR large language model OR chatbot) AND (developmental stage OR age-stratified OR age-specific OR age-appropriate)	298
2	(child OR children OR pediatric OR paediatric OR adolescent) AND (artificial intelligence OR machine learning OR deep learning OR large language model OR chatbot) AND (developmental stages OR cognitive development OR Piaget)	328
3	(pediatric OR paediatric OR children) AND (artificial intelligence OR machine learning OR deep learning) AND (age-stratified OR age labeling OR age labelling OR pediatric indication OR age-specific validation)	282
4	(artificial intelligence OR machine learning OR deep learning OR large language model OR chatbot OR conversational agent) AND (developmental stage OR developmental stages) AND (child OR children OR pediatric OR adolescent)	102
5	Title-field only: (artificial intelligence OR machine learning OR deep learning OR large language model OR chatbot) AND (child OR children OR pediatric OR paediatric OR adolescent OR adolescents OR youth OR kids)	3583

The difference between searches four and five illustrates the limitations of the available search terminology. Requiring explicit developmental stage terms returned 102 records, none of the sample records addressing developmental staging of child users. Removing this requirement and searching for artificial intelligence and pediatric terms in the title field returned 3,583 records, focused on clinical prediction and medical imaging without distinguishing between developmental stages. These results suggest that the relevant literature is difficult to identify through standard indexing terms and remains scattered across research areas.

2.3 Eligibility Criteria

Works were considered eligible for discussion if they concerned an AI or machine learning (ML) system used by, or applied to, individuals under 18 years of age, and if they reported empirical results, design requirements, or regulatory characteristics disaggregated by chronological age or by developmental stage.

Works were excluded if (1) studies reported only pooled pediatric results, with no breakdown by age or stage as they cannot inform an analysis of how stage is operationalized (2) developmental stages were described as a computational model rather than a human child, such as work treating training-set size as an analogue of cognitive development, and (3) the works presented neither original data nor an explicit framework.

2.4 Treatment of Evidence Quality

Peer-reviewed publication and preprints are distinguished throughout. Peer-reviewed sources carry the review's substantive claims. Preprints are cited only where no peer-reviewed equivalent exists, and every such citation is identified as a preprint at the point of use, so that the reader can weight it accordingly. All numerical values reported in this review were taken from the primary publication and verified.

2.5 Approach to Synthesis

The included studies use different outcome measures, including segmentation accuracy, classification performance, safety rating, and regulatory labeling counts. Hence, their results cannot be directly combined. Therefore, no meta-analysis was performed. The evidence was synthesized through three comparisons that do not require directly comparable metrics.

- (1) Definitions of developmental stages. Age boundaries specified by studies are recorded in years and compared to assess how definitions vary across studies.
- (2) Evidentiary basis of the study. The review records whether the conclusions are based on real child participants, simulated agents, retrospective records, or no empirical participants. It also documents whether each study examines a narrow or broad range within childhood.
- (3) Performance rankings across age groups. Groups are ranked from worst to best performance when outcomes are reported for three or more age groups. These rankings are compared across studies to assess whether the youngest groups consistently perform worst across domains. No claims of statistical significance are made because few studies meet this criterion.

Assumptions

3.1 Chronological age as an adequate proxy for developmental capacity

An assumption made by pediatric AI-enabled technologies is that a child's age in years is a sufficient index of the child's developmental state. This assumption is embedded when a pediatric population is defined by an age band, including regulatory labeling covering ages 0 to 17 years, as well as where a system's behavior is conditioned on an age value rather than on assessment of capacity.

However, different biological systems mature at different rates within the same individual, suggesting that no single measure can fully capture developmental state. An analysis of 951 participants aged 3 – 21 years from the Pediatric Imaging, Neurocognition, and Genetics study found that developmental trajectories of cortical measures varied by brain region [11]. Prefrontal regions showed more prolonged developmental changes than primary sensory areas [11]. This limitation reinforces that if developmental maturity cannot be readily summarized even within neuroimaging, chronological age cannot fully represent it.

Pediatric medicine has long recognized the limitations of using age as a proxy for biological development. Although drug dosing based on age, body weight, or body surface area provides a useful starting point, these measures may not capture developmental differences in drug metabolism and response. As a result, children receiving similar doses adjusted by weight may still experience differences in treatment efficacy and toxicity [3, 12]. In pharmacology, age remains a practical predictor, but its limitations are acknowledged. However, pediatric AI often uses age without explicitly accounting for these limitations.

3.2 Models developed on adults transfer to children

Systems validated in adults are often used in pediatric populations without child-specific evaluations. This assumes that children differ from adults only in degree, rather than in kind, and that performance estimates from adults remain broadly applicable to children.

An organ segmentation model trained on adult CT scans achieved a significantly lower mean Dice similarity coefficient on pediatric scans than on adult scans (0.73 vs 0.81, $p < 0.001$) [4]. The largest performance gaps occurred in smaller anatomical structures. Training on pediatric data or fine-tuning the adult model with pediatric images substantially improved performance. Similarly, segmentation frameworks that explicitly account for demographic characteristics, including age, report smaller performance disparities between adult and pediatric groups [5]. This suggests that some limitations associated with age-blind models can be avoided.

Reviews of pediatric AI also identify developmental variability, limited availability of pediatric data, and ethical concerns about bias and transparency as challenges that distinguish pediatric applications from adult applications [1]. Zapotoczny et al. accordingly recommend that age labeling and validation requirements for these technologies be standardized [6].

Additionally, segmentation errors are greatest for the smallest anatomical structures. Because these structures are generally smaller in younger children, the youngest patients may be particularly vulnerable to such errors. This is concerning in pediatric settings because patients may have limited ability to recognize or communicate the consequences of inaccurate model outputs.

3.3 Single set of pediatric safeguards is sufficient across childhood

The third assumption is that childhood can be treated as a single regulatory and design category, meaning that a system considered safe for one child is assumed to be safe for all children.

This assumption is reflected in current regulatory labeling. Among 952 AI-enabled device submissions reviewed by the United States FDA between November 1995 and June 2024, 4.4% included pediatric age ranges spanning 0 – 17 years [6]. A single label covering this range places infants who cannot speak and 17-year-olds within the same indication, despite their substantial developmental differences. Hence, standardizing age-labeling practices and validation requirements for these technologies is imperative.

This assumption extends beyond clinical regulation. When families use general-purpose language models to seek information about pediatric conditions, identified risks include hallucinations, bias, and data privacy concerns [2]. These risks may vary across childhood. A user's ability to recognize misleading information depends partly on their developmental stage and capacity to critically evaluate the model's output. Safeguards designed for an undifferentiated pediatric population may be calibrated to older children, who may be better able to evaluate information. Younger children, who may face greater risks, could therefore receive inadequate protection.

How Developmental Stage is Currently Represented

4.1 Evaluation of Language Models for Pediatric Use

General-purpose language models are frequently used for pediatric populations. In an evaluation of ChatGPT-3.5 as an information source for adolescent kidney transplant recipients and their caregivers, 37 questions were assessed by five pediatric nephrologists across accuracy, relevance, comprehensiveness, understandability, readability, and safety. Mean accuracy, relevance, and comprehensive scores were 4.51, 4.56, and 4.55 respectively, while 12 outputs were judged potentially risky, of which 3 carried a risk grade of moderate or higher [13]. The study concludes that human oversight remains necessary. Reviews of language models in childhood asthma identify hallucination, bias, and privacy as principal risks [2]. None of these evaluations stratify findings by developmental stage. Readability and understandability are assessed as properties of the output rather than as relations between the output and a particular reader's developmental capacity, with the consequence that a result obtained with adolescent users carries no stated implication for younger children.

4.2 Developmental Science

Developmental science supplies the evidence that stage is multidimensional but supplies it in a way that may not be directly usable by system designers. Developmental trajectories differ systematically between cortical regions within the same individual, with prefrontal regions showing protracted change relative to primary sensory areas across 951 participants aged 3 to 21 years [11]. This finding does not yield stage boundaries but establishes an approximation as a constraint on a staging standard rather than a component of one.

Gap Analysis

5.1 No Binding Definition of Developmental Stage

Stage boundaries are drawn independently in each body of work, with no mechanism requiring correspondence between them. Among 101 high-risk devices approved between 2008 and 2017, 26 (25.7%) specified no age range, only 4 used fully structured age ranges, and only 8 (7.9%) followed the age ranges recommended by the United States FDA [14]. This lack of adherence reflects inconsistent implementation rather than minor deviations from an established standard. Reviews of AI-enabled devices similarly call for standardized labeling and validation requirements [6, 14].

This gap does not arise from a lack of proposed classifications. A framework exists, but its use is not required. Without mandatory requirements, pediatric claims need not clearly specify the age groups for which a system has been evaluated. Consequently, two systems labeled as validated for pediatric use may have been tested on entirely different age groups, with no way for users to determine whether their validation populations overlap.

5.2 Evidence Drawn Predominantly From Sources Other Than Children

Across the bodies of work surveyed, conclusions about children rest largely on retrospective records, expert judgement, or simulated users rather than on children interacting with the systems in question.

Studies with real child participants exist but are not configured to close this gap. A qualitative focus group study conducted at a German University children's hospital and surrounding schools recruited 58 participants across three age bands, 10 aged 8 to 11 years, 23 aged 12 to 14 years, and 25 aged 15 to 17 years, in 18 focus groups examining views on AI in pediatric healthcare. Participants raised data privacy concerns, suspected lack of empathy, and risk of false diagnoses as barriers to acceptance [15]. Authors conclude that age-appropriate education is needed for acceptance and that minors

should be involved in development [15]. This is direct evidence from children, and it is evidence about attitudes rather than about system performance across developmental stages [15]. The gap is therefore not that children are never studied, but that where children are studied the outcome measured is not the one that stage-dependent system behavior would require.

5.3 Lack of Validated Method for Determining a Child's Developmental Stage

Any system designed to adapt its behavior to a child's developmental stage must first determine which stage applies. However, the reviewed literature does not establish a validated method for making this determination.

As discussed in Section 3.1, developmental trajectories vary across biological systems within the same individual, and developmental differences in drug disposition and response persist among children of similar weight [11, 12]. A method that infers developmental stage from a single observable measure may therefore fail to capture this variability, reflecting the same limitations that make chronological age an imperfect proxy for development.

A further challenge is the trade-off between accurate stage identification and data privacy. More accurate developmental assessments may require collecting additional information about the child, yet privacy is among the concerns children themselves raise about AI in healthcare [15]. A practical staging framework must therefore define not only the developmental stages but also the minimum information required to assign them.

5.4 Evaluation Horizons Shorter than Developmental Change

Because developmental stage changes as a child grows, and prefrontal development continues through adolescence, evaluations conducted at a single time point may not capture how a system performs as the child develops [11].

This limitation has two important implications. First, a system validated for a child at one developmental stage may continue to be used as the child progresses to another stage, without requiring reassessment of its performance. Second, the effects of sustained interaction remain largely unmeasured in peer-reviewed research because single session evaluations cannot capture cumulative effects. The only identified attempt to address the latter uses longitudinal simulations with simulated rather than real users.

This gap is relevant given that pediatric device labels commonly cover ages 0-17 years [6]. Such a broad age range implies that a system is suitable across multiple developmental stages. However, evaluations conducted at a single time point cannot establish whether its performance and safety remain consistent throughout childhood and adolescence.

5.5 Separation of the Clinical and Conversational Literatures

The two fields, clinical and conversational, most relevant to pediatric AI remain disconnected because they are published in different research venues and indexed in different databases.

The searches described in Section 2 highlight this separation. When the search required terms related to developmental stages alongside AI and pediatrics, it returned 102 records in PubMed. None of the sampled records addressed how developmental stages are defined for children interacting with AI systems. Removing the developmental stage requirements increased the results to 3,583 records, most of which focused on clinical prediction and medical imaging without distinguishing between developmental stages. Research that explicitly defines developmental stages for AI systems is published in venues that PubMed does not index.

As a result, each field lacks evidence that the other could contribute. Clinical research has quantified performance gaps in pediatric applications, shown that these gaps are greatest for the smallest anatomical structures, and demonstrated that accounting for age can reduce disparities [4, 5]. However, this work does not define developmental stages. Research on conversational AI, by contrast, proposes developmental stages but has not yet validated them against data from real children. Neither field has established what a developmental standard would require, namely clearly defined stages with boundaries supported by measured differences in outcomes.

A Proposed Developmental Staging Standard

6.1 Why Existing Age Bands Are an Inadequate Starting Point

One approach is to adopt the pediatric age bands already recommended by regulators, on the grounds that the problem identified in Section 5.1 is a lack of enforcement and not a lack of standards. However, evidence from pharmacology challenges this approach.

The age categories commonly used in regulatory guidelines, including neonates, infants and toddlers, children, and adolescents have more historical than physiological or developmental justification [16]. Regulatory guidelines acknowledge

that dividing the pediatric population into age groups is somewhat arbitrary, and that these categories should serve only as a guide [16]. The boundaries defining the neonatal and adolescent periods are particularly problematic [16]. The use of 12 years to distinguish children from adolescents is also questionable along with the upper age limit for adolescents, which varies across regions [16].

Adopting these categories would therefore standardize an existing convention without establishing its developmental validity. The boundaries of a useful staging standard must be supported by empirical evidence.

6.2 Principle (1): Stage as a Continuous Covariate, with bands as a reporting convention

The proposed framework's first principle is to represent developmental state as a continuous variable whenever the system allows it, using discrete age bands only for reporting and labeling.

This approach reflects developments in pediatric pharmacology, where modeling and simulation increasingly support the use of growth and development as continuous variables in drug development [16]. It also aligns with evidence that developmental trajectories differ across systems within the same individual (Section 3.1) and bad conditioning models and age can reduce performance disparities between pediatric and adult populations (Section 3.2) [5]. Unlike discrete bands, a continuous representation does not assume that children within the same age band are developmentally equivalent, or that meaningful differences emerge abruptly at band boundaries.

Discrete bands nevertheless remain useful for communication. Labels must specify which children the system's claims apply to, a function that a continuous variable alone cannot fulfill. The framework therefore treats age bands as a reporting layer built on top of a continuous representation, rather than as the primary representation of developmental state.

6.3 Principle (2): Boundaries justified by measured outcome differences

The second principle is that developmental stage boundaries should be determined by meaningful changes in system performance or risk, with the supporting evidence reported alongside each boundary.

This reverses the usual approach. Rather than adopting existing age bands and reporting performance within them, systems should be evaluated across the pediatric age range, with boundaries defined by the observed results. If performance remains stable across a range of ages, that range can be treated as a single band, regardless of conventional age categories. In contrast, if performance changes substantially at a particular age, a boundary should be placed there even if existing classifications do not recognize it.

Evidence from medical imaging suggests that such boundaries can be identified empirically. Adult trained segmentation models showed the largest performance deficits for the smallest anatomical structures, suggesting a relationship between anatomical size, developmental state, and model error [4]. Requiring boundaries to be supported by measured changes in performance would make this relationship an explicit part of the reporting process rather than leaving it as an observation.

6.4 Principle (3): An explicit rule for stage uncertainty

Because no validated method for determining a child's developmental stage was identified, the framework must specify how systems should behave when that stage is unknown or uncertain.

The proposed rule has two parts. First, when a child's developmental state cannot be established with sufficient confidence, the system should default to the behavior appropriate for the least mature developmental stage that remains possible, rather than relying on a population average or an adult default. Second, systems should disclose which information they use to estimate developmental state and what the minimum information required for that estimate is. This would limit the collection and use of personal data to what is necessary for developmental staging.

Children themselves frequently identify privacy as a concern when discussing AI in their care. A framework that improves safety by encouraging more extensive monitoring of children could address one problem while creating another [15]. Limiting the information used for developmental staging is therefore a core requirement of the proposed standard.

6.5 Principle (4): Validity claims bounded in time

Claims that a system is valid for a particular development mental band should specify the time frame of the supporting evidence, distinguishing between evaluation at a single point in time and evaluation across developmental changes. This does not mean that every system must undergo longitudinal evaluation. The type of evidence supporting a claim should be made explicit. For example, a claim of validity from ages 0 to 17 based on testing at a single point in time provides weaker

evidence than one supported by evaluations across developmental transitions. Currently, labels do not distinguish between these two levels of evidence.

Research Trajectory

7.1 Establishing that stage dependent behavior changes outcomes

The framework rests on the premise that matching system behavior to developmental state improves outcomes compared with using the same approach for all children. This has been demonstrated in clinical imaging but has not been established for conversational AI systems interacting with real children [4, 5].

Testing this premise would require presenting children across a broad age range with system outputs under matched and unmatched conditions. Outcomes would include comprehension and the ability to recognize incorrect information. The key question is whether the difference in outcomes between the two conditions varies with developmental state. A null result would also be valuable, suggesting that developmental matching does not improve the outcomes measured and directing future research toward other outcomes where it might.

Two features are essential to the study. First, participants must be real children rather than simulated agents, directly addressing the gap identified in Section 5.2. Second, the age range must be broad enough to capture potential developmental transitions. Studies involving only a narrow age range may miss meaningful differences simply because they do not include the ages at which those changes occur.

7.2 Locating Boundaries Empirically

Section 6.3 proposes that developmental stage boundaries should be based on measured differences in outcomes. Testing this requires evaluating performance across the pediatric age range using a consistent task and modeling outcome as a continuous function of developmental state, rather than comparing predefined age groups.

The analysis would determine whether performance changes gradually or shows distinct transitions. Gradual changes would support the continuous representation proposed in Section 6.2, suggesting that discrete reporting bands lack clear empirical boundaries. Distinct transitions, by contrast, would provide evidence for where boundaries should be placed and allow them to be compared with conventional age categories.

This study could begin with retrospective clinical data, which are available at the required scale in medical imaging archives, before moving to prospective evaluation [4].

7.3 Developing and Validating Stage Determination

Addressing the lack of a validated method for determining developmental requires a measurement study rather than a system evaluation. Candidate indicators would be compared against reference developmental measures collected from the same children, with each method's accuracy and sources of error reported.

Two requirements follow from Section 6.4. First, accuracy must be reported across different ages, since a method that performs well on average may be unreliable at particular developmental stages, where misclassification could have greater consequences. Second, methods should be compared according to the amount and sensitivity of the information they require. This would make it possible to identify approaches that offer only marginal improvements in accuracy at the cost of collecting substantially more personal data [15].

7.4 Evaluating across developmental change

Section 5.4 concerns the mismatch between the duration of an evaluation and the developmental changes covered by the systems validity claims. Addressing this gap requires following the same children through a developmental transition and applying the same evaluation measures at each time point. The central question is whether a system validated for a child at one point in time remains valid as that child develops. Until longitudinal evidence becomes available, validity claims must explicitly state when they are based on evaluations conducted at a single point in time.

7.5 A shared reporting specification

the studies outlined above would produce comparable evidence only if their findings are reported consistently. Establishing a shared reporting specification therefore requires agreement on the information every pediatric AI evaluation should include: the ages of participants, the developmental measures used, performance across the age range rather than only aggregated results within predefined bands, and the time frame of the evidence supporting each validity claim.

Regulatory analyses reviewed in this article have called for greater standardization of age labeling and validation requirements [6, 14]. A shared reporting specification would translate this need into research practice, establishing consistent expectations for reporting pediatric AI evaluations before any corresponding regulatory requirements are introduced.

LIMITATIONS

The search strategy differed across the fields covered by this review. PubMed was searched systematically, whereas literature from computer science and human-computer interaction venues was identified through citation chaining and targeted searches. As a result, the clinical literature is more comprehensively represented than the literature on conversational AI, and the discussion in Section 4 should be interpreted with this limitation in mind. This imbalance also reflects the separation between research fields identified in Section 5.5.

Screening and extraction were done by one individual with no duplicate assessment, so no agreement statistic available. Furthermore, no component of Section 6 has been validated.

CONCLUSION

Children encounter AI systems in both clinical care and everyday life, yet the evidence reviewed here suggests that these systems do not adequately account for developmental differences. Adult-trained models show measurable performance deficits in pediatric populations, with the largest gaps occurring for the smallest anatomical structures [4]. Pediatric indications remain uncommon among regulated AI devices, accounting for just 42 of 952 submissions [6]. Age labeling is also inconsistent, with only 8 of 101 high-risk devices using recommended age ranges [14]. Moreover, the age categories used in existing labeling reflect historical conventions rather than established developmental differences [16].

The main contribution of this review is not to present these findings as new, but to show how they relate to one another. Research documenting pediatric performance deficits has not established developmental stages, while research proposing developmental stages for AI systems has not validated them with real children. Neither body of work provides what a developmental standard requires: clearly defined stages with boundaries supported by measured differences in outcomes. The framework proposed here does not resolve these gaps. Instead, it outlines that addressing them would require representing developmental state continuously where possible, defining stage boundaries according to observed changes and outcomes, specifying how systems should behave when developmental state is uncertain, and limiting validity claims to what the available evidence supports. Each principle requires empirical testing through the studies proposed in Section 7, beginning with the fundamental question of whether matching system behavior to developmental state improves outcomes for children.

The case for pursuing this research does not depend on the framework being correct. It rests on the need to evaluate systems that are already used by children across different stages of development, despite labeling practices that often fail to distinguish between them. Where developmental differences have been measured, the youngest children have experienced the greatest performance deficits. Establishing whether and how AI systems should adapt to these differences is therefore an empirical question requiring research investigation.

REFERENCES

1. Park, T., Lee, I.-H., Lee, S. W., & Kong, S. W. (2025). Artificial Intelligence in Pediatric Healthcare: Current applications, potential, and implementation considerations. *Clinical and Experimental Pediatrics*, 68(9), 641–651. <https://doi.org/10.3345/cep.2025.00962>
2. Drummond, D., Girault, A., & Gonsard, A. (2026). CHATGPT and other large language models for childhood asthma. *Paediatric Respiratory Reviews*, 59, 57–64. <https://doi.org/10.1016/j.prrv.2025.09.002>
3. Kearns, G. L., Abdel-Rahman, S. M., Alander, S. W., Blowey, D. L., Leeder, J. S., & Kauffman, R. E. (2003). Developmental pharmacology — drug disposition, action, and therapy in infants and children. *New England Journal of Medicine*, 349(12), 1157–1167. <https://doi.org/10.1056/nejmra035092>
4. Chatterjee, D., Kanhere, A., Doo, F. X., Zhao, J., Chan, A., Welsh, A., Kulkarni, P., Trang, A., Parekh, V. S., & Yi, P. H. (2024). Children are not small adults: Addressing limited generalizability of an adult deep learning CT organ segmentation model to the pediatric population. *Journal of Imaging Informatics in Medicine*, 38(3), 1628–1641. <https://doi.org/10.1007/s10278-024-01273-w>
5. Ma, J., Tan, T., Jia, D., & Sun, Y. (2026). Demographic-Aware Deep learning for multi-organ segmentation: Mitigating gender and age biases in CT images. *Medical Physics*, 53(2). <https://doi.org/10.1002/mp.70322>

6. Zapotoczny, G., Goyal, A., Christmas, M., Qazi, S., Carroll, M., & Espinoza, J. (2026). FDA-regulated AI-enabled medical devices with pediatric indications. *JAMA Network Open*, 9(3). <https://doi.org/10.1001/jamanetworkopen.2026.2636>
7. Brewster, R. C., Nagy, M., Wunnava, S., & Bourgeois, F. T. (2025). US FDA approval of Pediatric Artificial Intelligence and machine learning-enabled medical devices. *JAMA Pediatrics*, 179(2), 212. <https://doi.org/10.1001/jamapediatrics.2024.5437>
8. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). Prisma extension for scoping reviews (PRISMA-SCR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/m18-0850>
9. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The prisma 2020 statement: An updated guideline for Reporting Systematic Reviews. *BMJ*. <https://doi.org/10.1136/bmj.n71>
10. Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... McKenzie, J. E. (2021). Prisma 2020 explanation and elaboration: Updated guidance and Exemplars for Reporting Systematic Reviews. *BMJ*. <https://doi.org/10.1136/bmj.n160>
11. Hagler, D. J., Thompson, W. K., Chen, C.-H., Reuter, C., Akshoomoff, N., & Brown, T. T. (2022). Do aggregate, multimodal structural neuroimaging measures replicate regional developmental differences observed in highly cited cellular histological studies? *Developmental Cognitive Neuroscience*, 54, 101086. <https://doi.org/10.1016/j.dcn.2022.101086>
12. Satyam, S. M., Prabhakar, S., Densil, T., Mohammed, H. T., Kumari, R., El-Tanani, M., Rehman, A., Kharoufeh, A., Dalbah, M., & Eltrabishi, M. T. (2026). Pharmacogenomics and epigenetic regulation transforming pediatric precision therapeutics. *Journal of Personalized Medicine*, 16(6), 329. <https://doi.org/10.3390/jpm16060329>
13. Demirbaş, K. C., Saygılı, S., Yılmaz, E. K., Gülmez, R., Ağbaş, A., Taşdemir, M., & Canpolat, N. (2025). The potential of chatgpt as a source of information for kidney transplant recipients and their caregivers. *Pediatric Transplantation*, 29(3). <https://doi.org/10.1111/ptr.70068>
14. Zapotoczny, G., Snitman, A., Shah, P., Bent, M., Glicksberg, B. S., Bourgeois, F., Bar-Cohen, Y., & Espinoza, J. (2025). Describing inconsistencies in pediatric labeling of medical devices. *JAMA Network Open*, 8(12). <https://doi.org/10.1001/jamanetworkopen.2025.46517>
15. Reinhart, L., Kerth, J.-L., Bischops, A. C., Hagemester, M., Krassuski, L., Mayatepek, E., & Meissner, T. (2026). Children and adolescents' views on artificial intelligence in pediatric healthcare: A qualitative focus group study. *BMC Pediatrics*, 26(1). <https://doi.org/10.1186/s12887-026-07121-w>
16. Hoppu, K., & Fonseca, H. (2021). Why are certain age bands used for Children in Paediatric Studies of medicines? *Archives of Disease in Childhood*, 106(7), 631–635. <https://doi.org/10.1136/archdischild-2020-319019>