Accuracy and Reliability of Artificial Intelligence-Generated Medical Learning Capsules as a Tool for Continuing Education in Residency

Evan Dimentberg, MD, MSc[1]; Lotfi Djani, MD[1]; Nathanel Harell[2]; Louis-Philippe Thibault, MD, MMSc, BBA[1,2, 3, 5]; Olivier Drouin, MD, MSc, MPH[1,2,3,4]

[1] Department of Pediatrics, CHU Sainte-Justine, Montréal, Québec, Canada

[2] Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada

[3] Azrieli Research Centre, CHU Sainte-Justine

[4]Department of Social and Preventive Medicine, School of Public Health, Université de Montréal

[5] Health Professions Education Research Center (CRPS), Université de Montréal

Corresponding author: Evan Dimentberg — evan.dimentberg@umontreal.ca

Keywords: large language models; artificial intelligence; medical education; continuing professional development; pediatrics

Funding: OD is supported by a Chercheur Boursier Clinicien Award, from the Fonds de recherche du Québec – Santé.

Conflicts of interest: E.D. developed the content-generation tool evaluated in this study. The authors report no financial conflicts of interest.

Word count (main text):

Word count (abstract):

Data availability: All data, including Claude Prompt, sampled editions, and auditor scores, remains available upon request. The algorithm will be published online following publication.

Abstract

Background. Large language models can encode clinical knowledge and generate fluent content. They are increasingly used to produce educational material in medicine. The same unsupervised models, however, can state incorrect facts and cite sources that do not support — or do not exist to support — the statement being made. Before AI-generated educational content is distributed at scale, its factual accuracy and source reliability must be validated.

Objective. To evaluate the factual accuracy of an AI-generated pediatric educational tool.

Methods. The authors designed the Daily Pediatric Learning algorithm, a locked and identical large language model, to produce daily 5-minute clinical briefs for pediatrics residents based on the Royal College of Physicians and Surgeons of Canada objectives. A sentence-level accuracy audit of editions was performed. A test-run of 5 samples was analyzed first to refine the tool, followed by twenty editions sampled at random. Editions were decomposed into individual statements and reviewed by two independent auditors. Primary outcome was accuracy, defined as the proportion of AI-generated clinical statements that are supported by medical literature (gold-standard).

Results: Initial tests highlighted a statement accuracy of 89.6% (121/135) and source-fidelity of 76.3% (103/135). Upon revisions, on a new dataset, accuracy improved to 99.6% (709/712; 95% confidence interval 98.8 to 99.9) and source-fidelity to 99.4% (708/712; 98.6 to 99.8). All 20 editions exceeded 92% accuracy and three errors with potential clinical impact were detected.

Conclusions: The Daily Pediatric Learning algorithm was found to be accurate and reliable at producing clinical briefs. Large language models can be tailored to ensure safety and effectiveness in medical education. With proper guardrails and precise instructions, AI-generated

content can serve as a convenient, cost-friendly supplement to current techniques in residency and medical school.

1. Introduction

Large language models (LLMs) are advancing rapidly in medicine. Non-medical, unvalidated systems trained on broad corpora now encode substantial clinical knowledge and answer medical questions at a level approaching that of clinicians.[1] Recent reviews describe expanding applications across diagnosis, documentation, decision support, and education.[2] On standardized assessments, LLMs have performed at or near the passing threshold of the United States Medical Licensing Examination,[3,4] demonstrating a capacity to reproduce and apply medical facts.

These newfound capabilities have driven their rapid adoption for clinical communication and teaching. LLM-generated responses to real patient questions have been rated higher in quality and empathy than physician responses in a blinded comparison,[5] yet commentators have detailed both the promise and the hazards of conversational models such as GPT-4 in medicine.[6] In medical education specifically, systematic and narrative reviews describe potential roles in tutoring, content generation, assessment, and self-directed learning.[7,8,9] Brief, recurring formats also share features with gamified e-learning, which commonly relies on regular challenges and feedback to sustain learner engagement.[14]

The exciting advancements are not without risk. Independent evaluation of LLM answers to physician questions has found variable accuracy.[10] LLMs can generate fabricated or inaccurate citations: studies of model-generated medical and scientific content have reported that a substantial proportion of references are erroneous or entirely invented,[11,12] such that a statement may appear well-supported while its cited source does not exist or does not contain the stated finding. This is particularly dangerous when designing tools intended for medical education. As clinical reasoning is increasingly delegated to AI, users may accept model output uncritically: a behaviour described as “cognitive surrender.”[13] Medical educators have also warned of

deskilling, the erosion of skills a trainee has already acquired, and never-skilling, the failure to develop them, and recommend that trainees verify AI output rather than accept it.[15–17]

The aim of this study is to evaluate the factual accuracy of an AI-generated pediatric educational tool for the purpose of providing residents with reliable and accessible medical content.

2. Methods

2.1 Context

The first author (ED) designed a LLM specifically for continuing education in pediatric residency, with content derived from the Royal College of Physicians and Surgeons of Canada's list of pediatric final exam topics.[18] The output is a 5-minute read clinical brief designed as an email, entitled 'Daily Peds Learning' (DPL), with a randomly selected topic daily. A statement - level accuracy and reliability audit was performed on the LLM's daily outputs.

2.2 Objectives

The primary objective of this study was to evaluate the factual accuracy of the Daily Peds Learning's AI-generated pediatric educational content, defined as the proportion of clinical statements supported by medical literature (gold standard). Secondary objectives include the sourced-rate, inter-rater reliability, the number and clinical severity of inaccuracies, the quality of sources the model selects, and any difference between English- and French-language editions.

2.3 Study design and setting

This is a statement-level pre-deployment accuracy audit of an AI-generated daily pediatric educational tool in the Department of Pediatrics, CHU Sainte-Justine, Université de Montréal. As there are no human participants or use of patient records, Institutional Ethics Review Board

approval was not required. An initial test of the original tool version informed a revision of the tool, which was then evaluated in a full run of 20 editions (Table 1).

2.4 System under evaluation: the content-generation tool

The material audited was produced by an automated content-generation tool (DPL) built by the primary author (E.D.). The tool runs once daily as an automated scheduled task and produces a 5-minute read pediatric educational brief; editions alternate between English and French to match local language contexts. Each edition follows fixed literature-review, in-depth-review or case-based formats (clinical question; context; design, population, and intervention; results with figures; interpretation; limitations; practical conclusion; key take-home points; sources), similar to a journal club.

The configuration under study is frozen as tool version DPL-v3.1. Editions are generated by a single, pinned model — Anthropic Claude Opus 5.5 (model identifier claude-opus-5-5), released September 22, 2026, so that every audited edition is produced by the same model version. The tool operates under fixed constraints: every clinical fact must carry a citable source listed in an explicit “Sources” section; acceptable sources are restricted to consensus guidelines (Canadian Paediatric Society, American Academy of Pediatrics), systematic and Cochrane reviews, and peer-reviewed journals from MEDLINE; numerical results must be transcribed as the actual figures from the source rather than approximated; the model must not fabricate facts or citations and must flag uncertainty where evidence is limited. Generation is source-grounded, with sources retrieved and verified at generation time. The process is run daily as to ensure sources are up to date. The complete frozen prompt is provided in Appendix A.

2.5 Editions, eligibility, and sampling

Twenty editions were selected at random with a 1:1 English to French ratio. The subject is the individual edition; the unit of analysis for scoring is the number of individual statements, of which there are approximately 30–50 per edition.

2.6 Audit procedure

Each edition was evaluated independently by two auditors, with a third to resolve discrepancies between the evaluators (ED, LD, NH). For every edition, the auditors commenced by parsing the original sentences into individual statements, with exclusions performed as necessary. Statements were excluded when they did not assert a checkable fact — clinical-case narratives, duplicate statements, rhetorical questions and section headers. All exclusions were performed manually by the auditors and the reason documented.

For each included statement, the auditors: (1) located the cited source and judged whether it supports the statement (sourced: yes or no); (2) judged the factual accuracy of the statement against gold-standard medical literature (true: yes or no); (3) for any inaccurate statement, assigned a clinical severity — potential clinical impact (errors in dose, treatment indications, symptoms, or diagnostic testing) versus no clinical impact (e.g., errors in epidemiology); (4) recorded verbatim the relevant cited text and the source used; and (5) characterised the type and tier of each source selected by the tool.

2.7 Reference standard and source hierarchy

Statements were assessed against the best available literature (gold standard), ranked by a predefined source hierarchy: Tier 1 consists of clinical practice guidelines (e.g., Canadian Paediatric Society, American Academy of Pediatrics, World Health Organization); Tier 2, systematic reviews, Cochrane reviews, and major-journal primary studies; Tier 3, narrative

reviews, textbooks, and point-of-care references. Blogs, unsourced opinion, and AI-generated summaries were not deemed acceptable reference sources.

2.8 Outcomes

The primary outcome is the proportion of clinical statements determined to be accurate relative to gold-standard medical literature. This is expressed as the percentage of individual statements adequately fact-checked against the medical literature. Secondary outcomes include: the sourced-rate, defined as the percentage of statements fact-checked directly in the listed citations, inter-rater reliability of the instrument, the number and severity distribution of inaccuracies; the proportion of editions achieving ≥90% accurate statements; the mean per-edition accuracy with its standard deviation; the distribution of source types across hierarchy tiers; and the comparison between English and French editions.

3. Results

3.1 Initial test

The five initial-test editions yielded 182 statements, of which 135 were included (21 to 33 per edition). After final rating, 121 were accurate (accuracy 89.6%; [95% CI 83.3 to 93.7]) and 103 were supported by their cited source (76.3%; [95% CI 68.5 to 82.7]) (Table 2). Per-edition accuracy ranged from 76.0% to 100% (mean 89.7%, SD 9.6), and three of five editions had accuracy >90%. Of the thirty-two statements that were inaccurate or unsourced, 10 had potential clinical impact. Agreement between auditors was 80.7% for accuracy (kappa 0.11; PABAK 0.61) and 79.8% for sourcing (kappa 0.44; PABAK 0.60).

3.2 Complete evaluation

The 20 full-run editions contained 991 statements. After 146 structural exclusions (107 duplicates, 14 questions, and 12 other), 712 statements were scored (mean 35.6 per edition, range 25 to 42).

After adjudication, 709 statements were accurate (accuracy of 99.6%; [95% CI 98.8 to 99.9]) and 708 were supported by their cited source (99.4%; [95% CI 98.6 to 99.8]), both significantly higher than in the initial test ($p<0.001$; Table 2). Mean per-edition accuracy was 99.6% (SD 1.8), and all 20 editions reached 90%; 19 were completely accurate. Accuracy was 99.2% (356/359) in English and 100% (353/353) in French editions, and 98.7%, 100% and 100% in literature-review, in-depth-review and case-based formats, respectively.

Four statements failed review (Table 3). Three, with potential clinical impact, came from a literature review on febrile infants aged 60 days or younger: the edition gave a single empiric anti-infectious regimen without the separate regimens the cited guideline recommends for infants by age range. The fourth, with no direct clinical impact, was a narrative opinion drafted by the LLM, not mentioned in the two sources.

Agreement between auditors on the 712 statements both included was 97.3% for accuracy (kappa 0.00; PABAK 0.95) and 95.0% for sourcing (kappa 0.04; PABAK 0.90).

The tool selected 36 sources (one to three per edition, mean 1.8): 22 clinical practice guidelines (Tier 1), 11 systematic reviews or major-journal studies (Tier 2) and 3 narrative reviews (Tier 3); none were unacceptable.

4. Discussion

Revising the generation rules of an automated pediatric teaching tool, so that every statement had to be drawn from, and cited to, a verified high-quality source, raised statement-level accuracy

from 89.6% to 99.6% and source fidelity from 76.3% to 99.4%. Every edition in the full run exceeded 90% accuracy, guidelines made up most of the sources, and performance was similar in English and French.

These results contrast with reports of variable accuracy and frequent fabricated references when general-purpose chatbots answer medical questions or produce medical text.[10–12] This difference likely reflects design rather than model capability. With the initial version, which listed sources only at the end and allowed the model to draw on its own knowledge, all 32 statements that failed review were unsupported by the listed sources, and 14 were also inaccurate. Requiring the tool to open each source, cite it for every statement and avoid any content it could not attribute converts the task from recall to summarization of a known document.

Few studies have audited AI-generated teaching material for trainees, and none at the statement level. When two anesthesiologists rated chatbot responses on entry-level residency topics, both judged only 55% of responses completely accurate, and requested references were frequently wrong.[21] Statement-level audits of source support, closer to our design, found that about 30% of statements from a web-enabled model were not supported by the sources it cited, and that 50% to 90% of responses from seven widely used models were not fully supported.[22] Our initial-test source fidelity of 76.3% falls within that range. Conversely, grounding a model in clinical guidelines has produced gains similar to ours: in hepatitis C management, accuracy rose from 43% to 99% when the guideline was supplied as context.[23] To our knowledge, this is the first study to audit a recurring, automated educational product with a pinned model at the statement level, in two languages and in pediatrics.

The residual errors are instructive. The clinically relevant error was not a fabrication but an omission: a dosing recommendation was presented without the age stratification in the source, so

a resident applying it to a neonate could choose the wrong regimen. Errors of this type, in which every element is individually sourced, but the summary is incomplete, are unlikely to disappear with stricter citation rules alone. Low error rates therefore do not remove the need for expert oversight of content used in clinical teaching.

5. Limitations

This study has some limitations. First, the tool could only use sources it could open; several runs rejected higher-tier sources behind paywalls or blocked by publisher websites, which may bias source selection toward openly available documents. Second, the adjudicating investigator developed the tool, which may introduce bias; this was mitigated by triple review. Third, auditors could not be blinded to the tool version, and statement-level scoring may understate how many distinct facts were checked in compound sentences. Finally, results apply to one tool, one model version and general pediatrics; model updates or adaptations to new specialties require revalidation, which the version lock is designed to trigger.

6. Implications and future directions

Pinning the model, freezing the prompt and recording both in every edition allow results like these to be tied to a defined configuration and repeated, despite advancements and model improvements from corporate LLM developers. Using an identical design, the next phases will test the tool in other specialties, beginning with pediatric emergency medicine and pediatric cardiology, to assess whether its accuracy can be generalized to any specialty. DPL will then be deployed at a Canadian academic site, where residents' appreciation will be studied. The tool will also be freely accessible online for medical trainees according to results of subsequent validations.

7. Conclusion

A frozen, source-grounded LLM configuration produced pediatric educational briefs in which more than 99% of statements were accurate and supported by their cited source, with high agreement between independent auditors. Strict citation rules, rather than model capability alone, appear to account for DPL's improved reliability. Large language models can be tailored to ensure safety and effectiveness in medical education. With proper guardrails and precise instructions, AI-generated content can serve as a convenient, cost-friendly supplement to current techniques in residency and medical school. The few remaining errors were omissions of context rather than fabrications, which supports diversification of learning strategies and continued expert review of AI-generated content used in clinical teaching.

References

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–180.

2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–1940.

3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.

4. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.

5. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–596.

6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233–1239.

7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887.

8. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. 2024;99(1):22–27.

9. Abd-alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9:e48291.

10. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483.

11. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus. 2023;15(5):e39238.

12. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13:14045.

13. Shaw SD, Nave G. Thinking — fast, slow, and artificial: how AI is reshaping human reasoning and the rise of cognitive surrender. Working paper. Philadelphia (PA): The Wharton School, University of Pennsylvania; 2026.

14. Khaldi A, Bouzidi R, Nader F. Gamification of e-learning in higher education: a systematic literature review. Smart Learn Environ. 2023;10:10.

15. Abdulnour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med. 2025;393(8):786–797.

16. Chow M, Li AY, Quek DYJ, Fan EB, Sadasiv MS. The cognitive ecology of medicine: generative AI, clinical reasoning, and the preservation of adaptive expertise. Postgrad Med J. 2026;102(1205):qgag068.

17. Abernethy J, Shah A, Chen B, Reynolds S, Wright SM, O'Rourke P. Integrating AI scribes into medical education: guardrails for preserving clinical reasoning. J Gen Intern Med. 2026;41(9):2598–2602.

18. Royal College of Physicians and Surgeons of Canada. Pediatrics competencies [Internet]. Version 1.0. Ottawa (ON): Royal College of Physicians and Surgeons of Canada; 2021 [cited 2026 Sep 27]. Available from:

https://www.royalcollege.ca/content/dam/documents/ibd/pediatrics/pediatrics-competencies-e.pdf

19. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–549.

20. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423–429.

21. Jin Z, Abola R, Bargnes V, Tsivitis A, Rahman S, Schwartz J, et al. The utility of generative artificial intelligence chatbot (ChatGPT) in generating teaching and learning material for anesthesiology residents. Front Artif Intell. 2025;8:1582096.

22. Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nat Commun. 2025;16:3615.

23. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med. 2024;7:102.

Table 1. Tool versions and audit procedures in the two stages

	Initial test	Full run
Tool version	Original version	Frozen revised version (DPL-v3.1)
Model	Pinned: Claude Opus 5.5 (claude-opus-5-5), checked at each run	Pinned: Claude Opus 5.5 (claude-opus-5-5), checked at each run
Editions audited	5	20
Language	French only	10 English, 10 French
Citations	Source list at end of edition	Inline citation required for every statement
Content rule	None	No content from the model's own knowledge (except a labelled fictional vignette)
Sources	Not restricted	1 to 3 per edition; predefined tier hierarchy; guideline required when one exists; open full text verified at generation; published within 8 years
Final rating	Statements passed by both auditors accepted; third investigator adjudicated all others	Statements passed by both auditors accepted; third investigator adjudicated all others

Table 2. Accuracy, source fidelity and agreement in the two stages

	Initial test	Complete evaluation
Editions	5	20
Statements scored	135	712
Statements per edition, mean (range)	27.0 (21–33)	35.6 (25–42)
Accurate, n (%; 95% CI)	121 (89.6; 83.3–93.7)	709 (99.6; 98.8–99.9)*
Supported by cited source, n (%; 95% CI)	103 (76.3; 68.5–82.7)	708 (99.4; 98.6–99.8)*
Per-edition accuracy, mean (SD); range	89.7 (9.6); 76.0–100	99.6 (1.8); 92.1–100
Editions with ≥90% accuracy, n (%)	3 (60)	20 (100)
Statements failing review, n	32	4
Potential clinical impact	10	3
No direct clinical impact	22	1
Auditor agreement, accuracy: % (kappa; PABAK)	80.7 (0.11; 0.61)	97.3 (0.00; 0.95)
Auditor agreement, sourcing: % (kappa; PABAK)	79.8 (0.44; 0.60)	95.0 (0.04; 0.90)
Sources, Tier 1 / Tier 2 / Tier 3/ unacceptable	Not assessed	22 / 11 / 3 / 0

CI, confidence interval; PABAK, prevalence-adjusted bias-adjusted kappa; SD, standard deviation. *P<0.001 versus initial test (Fisher’s exact test).

Table 3. Statements that failed review in the full run

Edition (language, format)	Statement (abridged)	Failure	Severity	Reviewer finding
13 (English, literature review)	High risk for invasive bacterial infection: IV ceftriaxone 100 mg/kg/dose every 24 h	Not true; not sourced	Potential clinical impact	Cited guideline recommends a different empiric regimen for infants aged 28 days or younger
13	If meningitis is suspected, add IV vancomycin 15 mg/kg/dose every 6 h	Not true; not sourced	Potential clinical impact	Age-specific regimen omitted (cefotaxime added for meningitis at 28 days or younger)
13	If herpes simplex virus risk factors are present, add IV acyclovir 20 mg/kg/dose every 8 h	Not true; not sourced	Potential clinical impact	Recommendation presented without the age range to which it applies
18 (French, case-based review)	Pearl: the two documents differ on the timing of ultrasound and the prophylactic antibiotic	Not sourced	No direct clinical impact	Comparison of the two sources stated by neither source

Appendix

Appendix A: Frozen re-usable prompt (tool version DPL-v3.1)

Daily Peds Learning — tool version DPL-v3.1

Frozen configuration. Any edit to this text, the model, the schedule or the tools = a new tool version.

Effective from the first run on 2026-09-24. Required model: claude-opus-5-5.

You generate one daily pediatric learning brief for a cohort of pediatric residents at CHU Sainte-Justine (Université de Montréal). The reader is a busy, senior resident (R2/R3) with solid baseline knowledge. Topics are drawn from the Medical Expert content of the Royal College of Physicians and Surgeons of Canada *Pediatrics Competencies* (2021, version 1.0), organized into the 23 rotating domains in Step 2.

The only deliverable is a Google Calendar invitation with the complete brief in its description. Do NOT create HTML files. Do NOT upload anything to Google Drive.

Step 0 — Model lock (RUN THIS FIRST)

Your system prompt states the exact model ID you are running on. This tool version requires the model **`claude-opus-5-5`**.

- If your model ID is anything other than exactly `claude-opus-5-5`: **STOP IMMEDIATELY.** Do not search, write or create anything. End the task with:

`⛔ DPL-v3.1 model lock: expected claude-opus-5-5, got [your model ID]. No edition generated.`

- If it matches, continue.

Step 1 — Already sent today? (DUPLICATE CHECK)

The task runs at 06:35, 10:35 and 14:35. The later two runs are catch-up attempts if the morning run failed.

List today's events on the primary calendar. Use `pageSize` 50 and follow `pageToken` until done. Read only the event titles. If any title starts with "📚 Pédiatrie du jour" or "📚 Peds of the Day", the edition was already sent. **STOP IMMEDIATELY** and end with: "✅ Today's edition has already been sent. No action needed." Otherwise continue.

Step 2 — Language, domain and format of the day

Get today's date in America/Toronto.

Language (alternates daily): compute N = number of days between 2026-09-23 and today.

- N even → **English** edition.

- N odd → **French** edition (Canadian French, standard Québec medical terminology; drug names may stay in INN form where that is Québec clinical practice, e.g. « omalizumab », « épinéphrine »).

The edition is written entirely in that language: title, headings, body, key points. Sources may be in either language.

Domain: day-of-year (1–365) mod 23. Each domain lists the Royal College competency items it covers (Royal College *Pediatrics Competencies* 2021, Medical Expert 1.3). Today's subtopic must fall under one of today's items.

- **0 Behaviour and development:** common behavioural problems (anger/aggression, bullying, crying, defiance); cognitive, motor, social, speech/language delays; alcohol-related neurodevelopmental disorder; ADHD; autism spectrum disorder; communication disorders; developmental coordination disorder; global developmental delay; intellectual disability; learning disabilities; sleep problems.
- **1 Cardiology:** arrhythmias; murmurs; chest pain; syncope; cyanotic and acyanotic congenital heart disease before and after repair; syndromes with cardiac defects; acute and chronic heart failure; cor pulmonale and pulmonary hypertension; endocarditis, myocarditis, pericarditis; cardiac manifestations of Kawasaki disease; rheumatic fever.

- **2 Nutrition and growth:** failure to thrive; nutritional deficiencies and excesses; obesity and metabolic syndromes; infant/child/adolescent nutritional requirements; human milk and breastfeeding; enteral and parenteral nutrition; restricted and fad diets.

- **3 Ear, nose, throat and upper airway:** ear deformities; epistaxis; feeding/swallowing issues; hearing loss and deafness; nasal/nasopharyngeal obstruction; neck masses; stridor, hoarseness, voice abnormalities; dentition, caries, dental trauma; cleft lip and palate; mastoiditis and sinusitis; otitis media and externa; pharyngitis; retropharyngeal abscess, supraglottitis, epiglottitis; tonsillectomy and its complications; upper airway abnormalities.

- **4 Endocrinology and metabolism:** ambiguous genitalia and disorders of sexual development; growth failure and short stature; hypo-/hypercalcemia; hypo-/hyperglycemia; precocious and delayed puberty; thyroid nodules, goiter and thyroid disorders; weight gain and loss; adrenal disease; type 1 and type 2 diabetes; hyperlipidemia; hypothalamic and pituitary disorders; rickets, osteoporosis and metabolic bone disease.

- **5 Ophthalmology:** change in vision; ocular discharge; painful or swollen eye; papilledema; red eye; retinal hemorrhage; strabismus; amblyopia; conjunctivitis; nasolacrimal duct obstruction; ocular injuries; ocular manifestations of systemic disease; orbital cellulitis; structural eye abnormalities.

- **6 Gastroenterology, hepatology and biliary:** abdominal masses; acute and chronic abdominal pain; constipation and encopresis; diarrhea; dysphagia; feeding difficulties; hepatomegaly; intestinal bleeding; intestinal obstruction; jaundice; vomiting and regurgitation; congenital GI malformations and repair complications; motility disorders; functional disorders; esophagitis, GERD, peptic ulcer; enteropathies, colitis, IBD; celiac disease and malabsorption;

pyloric stenosis; short gut and intestinal failure; liver and biliary disorders; liver transplantation; pancreatic disorders.

- **7 Genetics, teratology and metabolic disease:** dysmorphic features; common genetic syndromes; inborn errors of metabolism; teratogenesis; genetic testing indications.

- **8 Hematology and oncology:** common CBC and blood-smear abnormalities; abnormal bleeding or bruising; thrombosis; lymphadenopathy; hepatosplenomegaly; bleeding and clotting disorders; cytopenias; hemoglobin and red-cell disorders including nutritional anemias; platelet disorders; bone marrow failure; acute leukemias; common childhood malignancies; side effects of chemotherapy and radiotherapy; late effects of cancer therapy; transfusion of blood products.

- **9 Immunology and allergy:** anaphylaxis (IgE and non-IgE); angioedema and urticaria; prolonged fever; recurrent infections; allergic rhinosinusitis; autoimmune disorders; FPIES and non-IgE food allergy; hypersensitivity to drugs, biologics, vaccines, food, insect stings; immunodeficiencies (B-cell, T-cell, combined, phagocytic, complement, innate, immune dysregulation); care of the immunocompromised patient.

- **10 Infectious diseases:** diarrhea; fever; lymphadenopathy; rash; common viral, bacterial, fungal, parasitic and spirochetal infections; fever of unknown origin; fever without a focus; occult bacteremia; HIV; infection in the immunocompromised host; travel- and immigration-related infections; perinatal and congenital infections; antimicrobial stewardship and infection prevention.

- **11 Mental health:** anxiety; fatigue; sadness; school avoidance; self-harm; suicidal ideation or attempt; violent behaviour; anxiety disorders; conduct and oppositional defiant disorders; conversion and somatic symptom disorders; mood disorders; OCD; personality disorders; psychotic disorders; substance use disorders; involuntary treatment, capacity and consent.

- **12 Musculoskeletal and rheumatology:** joint and limb pain; joint swelling or deformity; spinal deformity and scoliosis; acute and chronic arthritis; bone tumours; common fractures, dislocations, injuries and osteochondroses; gait disorders; congenital MSK abnormalities; non-inflammatory connective tissue disease; septic arthritis and osteomyelitis; SLE, JIA, juvenile dermatomyositis, periodic fever syndromes; vasculitides including IgA vasculitis (HSP) and Kawasaki disease.

- **13 Neurology:** altered consciousness; developmental regression; headache; hypertonia; hypotonia and weakness; nystagmus, dizziness, vertigo; plagiocephaly and torticollis; raised intracranial pressure; seizures, seizure disorders and seizure mimics (including breath-holding spells); tics; autoimmune neurologic disorders; cerebral palsy; cerebrovascular disease; CNS malformations; demyelinating disorders; muscular dystrophies and myopathies; peripheral nerve disorders; head trauma and concussion; CNS infections; movement disorders; neurocutaneous syndromes.

- **14 Nephrology and genitourinary:** enuresis and incontinence; hematuria; hypertension; proteinuria; retroperitoneal and pelvic masses; scrotal and inguinal swelling or pain; hydronephrosis; urinary tract anomalies; external genital disorders including undescended testes; acid–base, electrolyte and water balance; nephritic and nephrotic syndromes; renal calculi; acute and chronic kidney dysfunction and dialysis indications; tubular disorders; urinary tract infection; vesicoureteral reflux and obstructive uropathy.

- **15 Respiratory:** acute and chronic cough; dyspnea; hemoptysis; hypoxemia; mediastinal masses; wheezing; acute lung injury and ARDS; acute and chronic aspiration; asthma; bronchiolitis; cystic fibrosis; pleural effusion; pneumothorax; sleep apnea; invasive and non-invasive ventilation.

- **16 Sexual and reproductive health:** dysmenorrhea; menorrhagia and menstrual disorders; urethral or vaginal discharge; gynecologic problems; adolescent pregnancy; sexually transmitted infections; contraceptive counselling.

- **17 Dermatology:** bullae; nevi; pruritus; rash; acne; autoimmune skin conditions; hair and nail conditions; papulosquamous eruptions; pigmentary and vascular congenital lesions (including hemangiomas); skin infections and infestations; cutaneous signs of systemic disease; eczema and dermatitis; burns and chemical/physical exposures; vesiculobullous diseases.

- **18 Life-threatening conditions and resuscitation:** respiratory distress and failure; acute upper airway obstruction; BRUE; cardiorespiratory arrest and CPR; coma; drowning; accidental and intentional overdoses; poisonings; hepatic failure; hypo- and hyperthermia; metabolic crises (DKA, hyperammonemia, metabolic acidosis); renal failure; sepsis; shock; severe burns; severe electrolyte and acid–base disturbance; status epilepticus; multiple trauma and traumatic brain injury; inter- and intrahospital transport; procedural sedation.

- **19 Neonatal-perinatal health:** neonatal adaptation; newborn care (temperature, feeding, fluids, glucose, pain); newborn screening; abnormal fetal growth; asphyxia and therapeutic hypothermia; birth trauma; chronic lung disease; cyanosis; neonatal drug withdrawal; electrolyte and glucose abnormalities; feeding problems; neonatal hematologic conditions; hyperbilirubinemia; intraventricular hemorrhage; neonatal neurologic issues; prematurity and post-discharge follow-up; respiratory distress; retinopathy of prematurity; neonatal sepsis; surgical newborn conditions (anorectal malformation, diaphragmatic hernia, esophageal atresia/TEF, gastroschisis, omphalocele, intestinal atresia or obstruction, necrotizing enterocolitis); neonatal resuscitation.

- **20 Adolescent health:** developmentally appropriate adolescent care and HEADSS assessment; confidentiality and its limits; lifestyle anticipatory guidance; self-management and adherence in chronic disease; transition to adult care; sexual orientation and gender identity; chronic fatigue; chronic pain; feeding and eating disorders; substance use.

- **21 Child maltreatment:** social risk factors; child-protection law and mandatory reporting; sentinel findings; abusive head trauma; emotional abuse and neglect; physical abuse; sexual abuse; adverse childhood experiences.

- **22 Prevention, hospital care principles and special populations:** anticipatory guidance (sleep, feeding, infant colic, injury prevention); vision, hearing and developmental screening; routine and targeted immunization, the under-immunized child, vaccine hesitancy; pain assessment and management; palliative and end-of-life care; perioperative medical care; complex care and technology dependence; medication dosing in renal or hepatic dysfunction; radiation dose and imaging choice (ALARA); social determinants of health; health of Indigenous children; children in care; international adoptees; refugees and global health.

Format: day-of-year mod 3:

- 0 🔮 In-depth clinical review

- 1 🩺 Mini-case with full review

- 2 📊 Literature review

Step 3 — Exclusion list: last 120 days (NO REPEATS)

No subtopic may repeat within 120 days.

1. List events on the primary calendar from 120 days ago to today. Run two searches, one with `fullText` "Pédiatrie du jour" and one with "Peds of the Day". Use `pageSize` 50 and follow every `pageToken`.

2. Keep only titles starting with "📚". Extract the subtopic after the last "|".

3. The chosen subtopic must not:

   - match any title word for word,

   - cover the same clinical concept in other words (e.g. "Bronchiolitis management" = "Treating infant bronchiolitis"),

   - substantially overlap a recent topic (e.g. no "Acute asthma in the ED" if "Pediatric asthma exacerbation" ran in the window).

Step 4 — Choose the subtopic

Within today's domain, choose a specific, high-impact subtopic. It must be:

- **within one of today's Royal College competency items listed in Step 2** (name the matching item in the end-of-run report),

- **not on the exclusion list** (critical),

- frequently met in daily practice (wards, ED, clinic, consults),

- high-stakes (errors have real consequences for patients),

- often misunderstood, under-recognized, or changing in the literature,

- actionable at the bedside the next day.

Avoid rare syndromes, zebra diagnoses, overly subspecialized content, and topics too basic for a senior resident.

The subtopic is only final once Step 5 finds qualifying sources for it. If no qualifying source exists, choose another subtopic rather than lowering the source standard.

Step 5 — Source selection (STRICT)

Sources determine the content. Be critical and conservative.

Number: a minimum of 1 and a **maximum of 3** sources per edition.

Eligibility — every source must meet ALL of the following:

1. **Open full text online.** Use WebFetch to open the full article or guideline yourself and confirm the complete text is accessible (not only the abstract, a preview, or a paywall page). Free-to-read versions (journal open access, PMC, or the society's own website) qualify. If you cannot read the full text, the source is ineligible.

2. **Published within the last 8 years.** The publication date of the version you are using must be on or after the date 8 years before today (for today's run: today's date minus 8 years). For guidelines and position statements, use the date of the current version. A "reaffirmed" date with no revision of content does **not** reset the date. Older landmark studies are ineligible, even if the recent sources discuss them. You may report what an eligible source says about an older study, citing the eligible source.

3. **Acceptable source type, chosen according to this hierarchy:**

- **Tier 1:** consensus guidelines (CPS, AAP, WHO, national guidelines).
- **Tier 2:** systematic reviews / Cochrane / major-journal primary studies.
- **Tier 3:** narrative reviews, textbooks, point-of-care references.
- **Not acceptable:** blogs, unsourced opinion, AI-generated summaries. This also excludes news articles, press releases, conference abstracts, preprints, and secondary websites that summarize a source. Always go to the original document.

4. **Real and verified.** Confirm the authors, title, journal, year and DOI/URL from the page you opened. Never cite a source you have not opened.

Hierarchy rule: always use the highest available tier. If a current Tier 1 guideline covers the subtopic, it must be one of the sources. Use Tier 3 only when no Tier 1 or Tier 2 source covers the content, and name the reason in your end-of-run report.

Internal record (not in the edition): before writing, list for each candidate: tier, publication date, full-text URL confirmed (yes/no), and accept/reject with the reason.

Step 6 — Write the brief: sourced content ONLY

Core rule: never create your own content. Every factual statement in the edition must come directly from one of the 1–3 selected sources. Do not use your own background knowledge to add facts, figures, mechanisms, doses, thresholds, recommendations, or "pearls".

- **Cite every statement.** End each factual sentence (or dosing/criteria block) with its source number: [1], [2] or [3]. A sentence without a citation must not state a fact.
- **Transcribe numbers exactly.** Doses, thresholds, percentages, effect sizes, confidence intervals and sample sizes are copied exactly as in the source (translated into the edition's language where needed, with no rounding or approximation).
- **Stay close to the source.** Paraphrase faithfully or translate. Do not extrapolate, generalize beyond the study population, combine sources into a new conclusion that neither states, or turn a weak recommendation into a strong one. Keep the source's strength-of-recommendation and certainty wording.
- **Gaps stay gaps.** If a template section cannot be filled from the sources, shorten it or omit it. Do not fill it from memory. If the sources disagree, state the disagreement and cite both.
- **Synthesis sections too.** "Pearls and pitfalls" and "Key take-home points" must also be cited statements taken from the sources, not new advice.
- **Mini-case vignette (format 1 only):** the vignette is the only permitted non-sourced text. Start it with "Illustrative case (fictional):" (FR: « Cas illustratif (fictif) : »). Its clinical findings

must be consistent with the presentation described in the sources. The answer and the entire review that follow must be fully cited.

- Spell out every acronym in full the first time it is used, with the abbreviation in parentheses (e.g. "vesicoureteral reflux (VUR)"; FR: « reflux vésico-urétéral (RVU) »).

Length: **700–900 words** of body text (a ~5-minute read).

Structure (translate headings into the edition's language):

📚 Daily Pediatric Learning — [full date] / Apprentissage pédiatrique du jour — [date complète]

Topic / Sujet: [subtopic]

Domain / Domaine: [domain]

Format: [format]

———

[CONTENT: see formats below, with [n] citations throughout]

———

🎯 Key take-home points / Points clés à retenir: 3–5 points, each cited

📖 Sources:

[1] Authors. Title. Journal/Organization. Year;vol:pages. Tier [1/2/3]. Full text: [DOI or short URL]

[2] …

[3] …

———

DPL-v3.1 · claude-opus-5-5 · [EN/FR]

Formats:

- **🔮 In-depth clinical review:** 1) Why it matters · 2) Background and pathophysiology · 3) Presentation and diagnosis (red flags, investigations, thresholds) · 4) Management (evidence-based approach, doses, escalation, controversies) · 5) Pearls and pitfalls.

- **🩺 Mini-case with full review:** fictional vignette (4–6 sentences, labelled as above) → the key clinical question → the answer → the same 5 sections, fully cited.

- **📊 Literature review:** Clinical question → Context → Design/population/intervention → Results with the actual figures → Interpretation → Limitations → Practical conclusion. The study being reviewed must itself meet every rule in Step 5.

Step 7 — Create the Google Calendar invitation

Create one 5-minute event:

- **summary:**
  - EN: `📚 Peds of the Day [EN] – [Ddd Mmm DD] | [subtopic]`
  - FR: `📚 Pédiatrie du jour [FR] – [jjj JJ mois] | [sous-sujet]`
- **start/end:** 06:35–06:40 America/Toronto, today's date

- **notifications:** send to all attendees

- **attendees (12):**

- **reminders:** popup at 0 minutes, email at 0 minutes

- **description:** the complete brief.

Description formatting (Google Calendar limit ~8,000 characters):

- Compact layout. No heavy separator lines (no ━━ or ═══). Use a short "——" sparingly. Short section headings. No blank-line padding.

- Target ≤ 7,500 characters. Count the characters before creating the event. If the count is over 7,800, shorten the prose while keeping every citation. **Never drop the full-text links from the Sources section. Never truncate the clinical content mid-section. Never drop the version footer.**

Step 8 — Quality checks before creating the event

1. The model is `claude-opus-5-5` (Step 0).

2. The subtopic falls under one of today's Royal College competency items and is not on the 120-day exclusion list.

3. There are 1–3 sources. Each has open full text you opened yourself, is dated within the last 8 years, and is an acceptable type. The highest available tier was used.

4. Every factual sentence carries a [n] citation that points to a listed source and is supported by that source's text. Re-read each one against the source. Remove any sentence you cannot match.

5. Numbers match the sources exactly.

6. The whole edition is in today's language (EN or FR) and acronyms are spelled out on first use.

7. The description is under 8,000 characters and includes the version footer.

8. All 12 attendees are on the invitation.

End-of-run report (in the task output, not in the invitation)

Report: model ID · language · domain · format · subtopic · matching Royal College competency item · the table of candidate sources with accept/reject reasons · the final sources with their tiers · the description's character count · the event link.

Appendix B: Sample edition

📚 Daily Pediatric Learning — Friday, September 25, 2026

Topic: Recurrent infections: when to suspect an inborn error of immunity

Domain: Immunology & allergy

Format: 🩺 Mini-case with full review

——

CASE

Illustrative case (fictional): A 20-month-old boy is admitted with his second pneumonia in 10 months; the first required intravenous (IV) antibiotics. His weight has crossed down from the 50th to the 15th percentile. On examination his tonsils are barely visible and you cannot palpate any cervical lymph nodes.

QUESTION

Which features should make you suspect an inborn error of immunity (IEI), what is the first-line workup, and who needs referral?

ANSWER

IEIs should be suspected at any age in patients with recurrent, severe and/or unusual infection(s), poor response to appropriate antibiotics, failure to thrive, persistent thrush or skin abscesses, or a family history of IEI [1]. A male presenting within the first 2 years of life with recurrent sinopulmonary infections and absent lymph nodes and tonsils matches the source's description of X-linked agammaglobulinemia (XLA) [1]. Start with a complete blood count (CBC) and smear and serum IgG, IgA and IgM, and involve a clinical immunologist: patients meeting warning-sign criteria should be referred immediately [1].

1) Why it matters

IEIs, formerly primary immunodeficiencies, are a large heterogeneous group of disorders caused by deficiencies in immune system development and/or function [1]. They are more frequent than previously believed, with an estimated prevalence of 1 in 1200 live births [1]. Many present as "routine" infections, often of the sinuses, ears and lungs [1]. Timely diagnosis and treatment are imperative to prevent significant disease-associated morbidity and mortality [1].

2) Background and pathophysiology

- Antibody deficiency: recurrent, often severe sinopulmonary infections with encapsulated bacteria, particularly Streptococcus pneumoniae; presentation after 6 months of age follows loss of transplacentally transferred maternal IgG [1].

- XLA: mutation in the Bruton's tyrosine kinase (BTK) gene; markedly reduced circulating B cells and serum IgG, IgA and IgM [1].

- Common variable immunodeficiency (CVID): markedly reduced IgG, low IgA and/or IgM, and poor or absent vaccine responses; usually later onset; associated with autoimmune and granulomatous disease [1].

- Selective IgA deficiency: about 1 in 300 to 1 in 500 persons; most are asymptomatic, but among symptomatic patients up to one-third have recurrent infections and one-third present with autoimmunity [1].

- Severe combined immunodeficiency (SCID): may present in the first year with chronic diarrhea, failure to thrive, opportunistic infections (Candida albicans, Pneumocystis jiroveci pneumonia [PJP], cytomegalovirus [CMV]) and skin rashes [1].

3) Presentation and diagnosis

How good are warning signs? A systematic review of 12 studies (~7000 participants, 8 countries; MEDLINE and EMBASE to May 2024) assessed their accuracy in children [2]. Performance was highly heterogeneous and pooled estimates are exploratory [2]. Most signs had high specificity but low sensitivity [2]. Pooled sensitivity / specificity, % (95% confidence interval) [2]:

- Need for IV antibiotics: 71.6 (41.3-90.1) / 64.4 (44.5-80.2)

- ≥2 pneumonias: 50.8 (32.5-68.8) / 77.8 (61.1-88.6)

- Failure in weight gain: 36.5 (25.7-48.7) / 89.8 (74.2-96.4)

- Persistent thrush: 28.5 (13.4-50.8) / 94 (86.8-97.4)

- ≥2 months of antibiotics: 25 (11.8-45.4) / 97.3 (93.9-98.8)

- Family history: 19 (10-33.3) / 95.1 (86.9-98.3)

Up to one-third of children with IEI met none of the Jeffrey Modell Foundation (JMF) warning-sign criteria [2]. Beyond infection, excessive inflammatory responses and autoimmunity are common, and multiple hypersensitivities and atopy may also signal an IEI [1].

Initial workup [1]:

- CBC and smear: significant lymphopenia may be the first indication of T-cell immunodeficiency.

- Serum IgG, IgA, IgM: levels clearly below age-appropriate reference values may indicate B-cell immunodeficiency.

- Specific antibody titers to vaccine antigens, protein (e.g., tetanus toxoid) and polysaccharide (e.g., pneumococcus): the best way to confirm an antibody deficiency.

- Flow cytometry for B, T and natural killer (NK) cells: abnormal in most SCID cases.

- Targeted tests: dihydrorhodamine (DHR) neutrophil function assay, complement studies, genetic testing (gene panels or whole exome sequencing).

4) Management

Consultation with a clinical immunologist is required to confirm the diagnosis and establish a treatment plan [1]. Immunoglobulin (Ig) replacement is the mainstay for antibody deficiency; IV and subcutaneous formulations are considered equally effective, and dosing must be carefully monitored (the source gives no mg/kg dose) [1]. Commonly used prophylactic antibiotics include amoxicillin, trimethoprim-sulfamethoxazole and azithromycin [1].

SCID is a pediatric emergency: it is fatal unless corrected, hematopoietic stem cell transplantation (HSCT) should start as quickly as possible, and outcomes are dramatically improved when it is early [1]. Patients with SCID should start PJP prophylaxis, receive irradiated CMV-negative blood products, and must not receive live attenuated vaccines (measles/mumps/rubella/varicella, bacillus Calmette-Guérin, rotavirus, oral polio) [1].

5) Pearls and pitfalls

- Warning signs have good rule-in but poor rule-out value; relying on them alone risks delayed or missed diagnoses, particularly for noninfectious phenotypes [2].

- Recurrent pneumonia or need for IV antibiotics gave a more balanced but still imperfect profile [2].

- The T-cell receptor excision circle (TREC) newborn screen is effective for some types of SCID but not other IEIs [1].

———

🎯 Key take-home points

- IEIs affect about 1 in 1200 live births and often masquerade as routine sinus, ear and lung infections [1].

- A warning sign raises suspicion, but its absence does not exclude IEI: up to one-third of affected children meet no JMF criteria [2].

- First-line tests: CBC with smear, IgG/IgA/IgM against age-appropriate norms, and vaccine antibody titers [1].

- Refer early to a clinical immunologist; suspected SCID is an emergency requiring prompt HSCT and no live vaccines [1].

📖 Sources:

[1] Kim VHD, Upton JEM, Derfalvi B, Hildebrand KJ, McCusker C. Inborn errors of immunity (primary immunodeficiencies). Allergy Asthma Clin Immunol. 2024;20:76 (published 2025). Tier 3. Full text: https://doi.org/10.1186/s13223-024-00938-z

[2] Aden A, Jemide I, Islam T, et al. Clinical warning signs in detecting inborn errors of immunity in children: a diagnostic accuracy systematic review. J Allergy Clin Immunol Pract. 2026;14(4):920-936.e1. Tier 2. Full text: https://doi.org/10.1016/j.jaip.2026.01.026

———

DPL-v3.1 · claude-opus-5-5 · EN