Ethiopian Medical Journal 2026, 64 (18) *ISSN 0014-1755* *eISSN 2415-2420* *Leulseged et al*

Original Article

Accuracy and Reliability of MedBrain in Assisting Triage and Diagnosis of Common Acute Pediatric Conditions in Ethiopia

Tigist Workneh Leulseged[1,2*], Tadele Hailu[3], Fitsum Libeyesus[4], Seyoum Berihun Derbew[5], Teferi Gebrewahd Gebreslassie[6], Kirubel Tesfaye Hailu[1,7], Bezawit Woldaregay Wagaye[4], Thomas Shimelis[8], Tsegay G/anenia Hagos[9], Firaol Mame Abdi [10], Betelhem Tiruneh Gebremedhin[4], Solomon Worku Beza [11]

[1] Medical Research Lounge (MRL), Addis Ababa, Ethiopia
[2]Department of Internal Medicine, St. Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia
[3]Department of Pediatrics and Child Health, St. Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia
[4]Department of Pediatrics, Alert Comprehensive specialized Hospital, Addis Ababa, Ethiopia
[5]Department of Orthopedics, Onandjokwe Intermediate Hospital, Oniipa, Namibia
[6]Department of Radiology, Manhal Specialty Hospital, Hargeisa, Somaliland
[7]School of Public Health, University College Cork, Cork, Ireland
[8]Innovation and Strategic Operations Directorate, St. Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia
[9]Emergency and Critical Care Medicine, Alert Comprehensive specialized Hospital, Addis Ababa, Ethiopia
[10]Primary Health Care, Emirates Health Service, Fujairah, United Arab Emirates
[11] Federal Ministry of Health, Addis Ababa, Ethiopia

*Corresponding author: tigdolly@gmail.com

Abstract

Background: *Childhood illnesses are a leading cause of morbidity and mortality in Sub-Saharan Africa, where healthcare infrastructure and trained personnel are limited. MedBrain, a digital decision support system (DDSS), aims to enhance pediatric emergency care by supporting mid-level healthcare workers in low-resource settings.*

Objective: *To evaluate MedBrain's triage and diagnostic performance among children presenting with common acute conditions to the emergency departments of two large hospitals in Ethiopia.*

Methods: *A prospective observational diagnostic accuracy study was conducted between July 2024 and April 2025 at St. Paul's Hospital Millennium Medical College and Alert Comprehensive Specialized Hospital. MedBrain's triage and diagnostic performance were compared against healthcare professionals' triage and pediatricians' top presumptive diagnoses as gold standards. Performance metrics included accuracy, sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), likelihood ratios (LR+ and LR–), and Cohen's Kappa for reliability. Diagnostic performance was assessed for MedBrain's top three ranked diagnoses (Top 1: highest probability diagnosis; Top 2: top two diagnoses; Top 3: top three diagnoses).*

Results: *Of 1,204 patients screened, 274 were excluded for conditions not yet represented in MedBrain's database (including malaria), leaving 930 participants. Of which, most were infants (33.8%) and children under 5 (31.9%), with pneumonia (20.4%) the most common diagnosis. MedBrain achieved 72.2% triage agreement, with 3.7% over-triage and 24.2% under-triage. Total diagnostic accuracy was 84.1% (Top 1), 91.5% (Top 2), and 93.3% (Top 3), with Sn of 93.3% and PPV of 100%. For prevalent conditions (pneumonia, acute bronchitis, late-onset neonatal sepsis, acute gastroenteritis, bronchiolitis, and meningitis), accuracy exceeded 97.4%, and Sp and PPV were consistently perfect. Sn increased from 73.0–98.6% (Top 1) to ≥90–100% (Top 2–3).NPV increased from 97.1–99.9% (Top 1) to 98.9–100% (Top 2) and 99.2–100% (Top 3). LR– improved from 0.014–0.270 (Top 1) to 0–0.100 (Top 2) and 0–0.079 (Top 3). Similarly, Cohen's Kappa rose from 0.830–0.993 (Top 1) to 0.943–1.000 (Top 2) and 0.955–1.000 (Top 3). Diagnostic failures were rare, highest for late-onset neonatal sepsis (0.8%), bronchiolitis (0.5%), and pneumonia (0.4%), and none for gastroenteritis.*

Conclusion: *MedBrain demonstrated high diagnostic accuracy and reliability in Ethiopian pediatric emergency settings. Under-triage and limited disease coverage remain challenges, warranting further validation with expanded* disease libraries and in diverse settings.

Conclusion: *MedBrain demonstrated high diagnostic accuracy and reliability in Ethiopian pediatric emergency settings. Under-triage and limited disease coverage remain challenges, warranting further validation with expanded* disease libraries and in diverse settings.

Key words: *MedBrain, digital decision support system, pediatric emergency care, diagnostic accuracy, Ethiopia*
Citation: *Leulseged TW, Hailu T, Libeyesus F et al .Accuracy and Reliability of MedBrain in Assisting Triage and Diagnosis of Common Acute Pediatric Conditions in Ethiopia. Ethiop Med J 64 (18) 1-12*
Submission date : *3 October2025 Accepted: 11 December 2025 Published: 14 January 2026*

Childhood illnesses remain a major cause of morbidity and mortality worldwide, with a disproportionate burden on low- and middle-income countries (1, 2). In Sub-Saharan Africa, pneumonia, diarrheal diseases, and neonatal sepsis continue to be leading causes of under-five mortality, despite significant progress in child health interventions (1–3). In Ethiopia, these conditions are also among the leading causes of emergency visits and hospital admissions (3, 4), and national data revealed persistently high mortality rates, with 33 neonatal deaths, 47 infant deaths, and 59 under-five deaths per 1,000 live births (5).

This challenge is compounded by limited healthcare infrastructure and an inadequate health workforce, hindering the achievement of Universal Health Coverage by 2030, a key Sustainable Development Goal (2). Sub-Saharan Africa is reported to have the lowest health worker density, with nurses and midwives providing 90% of patient contact (7, 8). Ethiopia, in particular, has a health professional density far below national and WHO targets, with general practitioners and specialists accounting for only a small share of health professionals (3, 4), resulting in reliance on mid-level professionals to provide frontline care in most settings.In such contexts, clinical decision support tools to enhance the quality of care provided by mid-level professionalsmay help improve diagnostic accuracyand support timely intervention.

In addition, limitations in providers' clinical decision-making, particularly in accurately diagnosing and appropriately triaging patients, often contribute to misclassification and delays in care, which can lead to poor patient outcomes (37–39). These further underscores the need for decision-making tools that can strengthen provider competency and improve patient safety (13–16, 40).

Digital decision support systems (DDSSs) have shown promise in enhancing clinician performance and patient outcomes. However, existing tools have demonstrated limited diagnostic accuracy and poor triage performance(6). Symptom checkers for self-diagnosis generally show variable triage accuracy (49 –90%) and low diagnostic accuracy (19–38%) (6, 9). Other DDSSs have reported better results, with diagnostic accuracy up to 75% (10, 11), and one pediatric -focused tool reaching 95% accuracy (12). While some studies suggest DDSSs can improve professional performance and reduce costs (13–16), others report inconsistent results (17-19). Recent efforts, including AI-based DDSSs, aim to improve accuracy and usability (20–22). However, concerns about applicability and generalizability persist due to limitations in study design, such as small sample sizes (6), reliance on retrospective data or case vignettes (23–27), use of less representative populations (28–31),inclusion of highly diverse patient groups that limit understanding of applicability in specific fields (32, 33),and reporting accuracy based on a narrow set of diagnoses (11, 12). Moreover, the utility of DDSSs in low- and middle-income countries, where they could offer substantial benefits, remains largely understudied.

MedBrain is a newly developed DDSS specifically designed for pediatric emergency care in low-resource settings, with the primary goal of supporting mid-level healthcare professionals in triaging and diagnosing common pediatric conditions (34).MedBrain is a web-based platform that guides healthcare professionals through a dynamic, stepwise interview. It conducts a dynamic interview through iterative steps, asking questions and prompting physical examinations, then provides a diagnostic prediction once a probabilistic threshold is met.MedBrain applies a novel approach by assigning diagnostic weights (likelihood ratios) to symptoms and signs, emphasizing combinations of findings (clinical patterns) rather than static decision trees. Patient data are then matched against the database to generate a ranked list of potential diagnoses and recommended next steps.

In an internal, pre-clinical validation using 250 externally sourced cases from the British Medical Journal (BMJ), MedBrain achieved a diagnostic accuracy of 83% for the top-ranked diagnosis, 93% for the top two, and 98% for the top threediagnoses. Although this validation was limited by its retrospective, internal, and non-clinical design, the results indicated the tool's diagnostic potential. Building on this, the current study aimed to evaluate MedBrain's diagnostic and triage performance in real-world clinical settings among children presenting to the emergency departments of two large hospitals in Ethiopia with common acute conditions.

Methodology

Study Design and Setting

A prospective observational study of diagnostic accuracy was conducted from July 2024 to April 2025, adhering to the Standards for Reporting Diagnostic Accuracy Studies (STARD) guideline (35). The study was carried out at the pediatric emergency departments of two hospitals in Ethiopia: Alert Comprehensive Specialized Hospital (ACSH) and St. Paul's Hospital Millennium Medical College (SPHMMC), a tertiary referral teaching hospital. These sites were selected to ensure the presence of pediatricians who could consistently provide a reliable gold standard for comparison, while also representing different patient flows and case profiles. To capture seasonal variation in pediatric illnesses, data were collected in two phases: July to August 2024 (the main rainy season) and March to April2025 (the hot, dry season).

Population and Eligibility

The source population comprised all pediatric patients aged 18 years or younger who presented to the emergency departments of the hospitals during the study period. From this population, all eligible patients were included in the study. Patients were considered eligible if they presented with common acute pediatric conditions and visited the emergency department for a new complaint, regardless of urgency. After screening, patients were further excluded if they had a condition not yet included in MedBrain's database.

Sample Size Determination

Sample size was determined to estimate MedBrain's diagnostic performance for common pediatric conditions using a formula for diagnostic accuracy estimation with known disease prevalence (35).

The sample size required to estimate MedBrain's sensitivity was calculated considering the following parameters: an estimated prevalence of common pediatric conditions of 60% (based on the expected occurrence rate of various common conditions), an expected sensitivity of 60% (based on similar prior studies) (7-10), a 5% margin of error, and a 5% level of significance. This yielded a sample size of 616.

Similarly, the sample size required to estimate MedBrain's specificity was calculated under the same prevalence assumption (60%),an expected specificity of 60% (based on similar prior studies) (7-10), a 5% margin of error, and a 5% level of significance. This yielded a sample size of 924.

Taking the larger of the two calculated sample sizes (924) and adding a 30% non-response rate, the final sample size required was 1201.A 30% non-response rate was applied to maintain the study's statistical power, anticipating high exclusion rates from diagnoses not present in the MedBrain database during its initial real-time testing. The final sample size was distributed proportionally across the two study hospitals basedon their patient flow in the previous year: 240 patients at ACSH and 961 at SPHMMC. Consecutive pediatric emergency patients were recruited until the target was achieved, with equal allocation across the two data collection phases.

Outcome Measurements

The study assessed the following outcomes:

- **Triage Accuracy**: Agreement between MedBrain's urgency classification and the triage staff's assessment, based on the **Integrated Interagency Triage Tool (IITT)**, which categorizes patients as high urgency, priority, or non-urgent (36).
- **Diagnostic Accuracy**: Proportion of cases in which MedBrain's suggested diagnoses matched the pediatrician's presumptive diagnosis. Accuracy was evaluated at three ranked levels: **Top 1** (correct diagnosis listed first), **Top 2** (within the first two outputs), and **Top 3**(within the first three outputs).
- **Diagnostic Failure**:Cases without a correct diagnosis in the top three, or with no diagnosis generated, were classified as failures. A "no diagnosis" outcome indicated insufficient matches between patient data and MedBrain's library, even when the disease was present in its database.

Reliability: Degree of agreement between MedBrain and pediatricians' presumptive diagnoses, measured using Cohen's Kappa statistic.

Data Collection Procedure and Quality Assurance

Data on socio-demographics and clinical presentation were collected using the MedBrain software through structured interviews administered by BSc nurses trained as independent data collectors.

Upon arrival at the emergency department, patients first underwent routine triage performed by the assigned hospital staff (nurse, general practitioner, or a resident) using the hospital's standard criteria. After the patient was triaged and while waiting for medical assessment/intervention in the ER, the independent data collector initiated a parallel assessment using MedBrain, starting from the patient's chief presenting complaint. After a few questions, the application triaged the patient. It displayed that triage was complete, without showing the score in the data collector's window to avoid influencing routine medical care or biasing the collected data. The professional triage classification was subsequently entered into the software to serve as the gold standard.

Following triage and emergency evaluation, the data collector continued to gather clinical data in parallel with routine care through direct patient interview and physical examination. This infor-

mation was entered into MedBrain to generate a presumptive diagnosis, independent of the assessment and care provided by the patients' physicians. The tool then produced a ranked list of the top three potential diagnoses, which were stored in the application's database but concealed from the data collector's display to prevent bias. Subsequently, the pediatrician's top presumptive diagnosis, documented in the medical record, was extracted and entered into MedBrain as the diagnostic gold standard. Presumptive diagnoses were used instead of definitive diagnoses because MedBrain is designed for rural settings where laboratory and imaging modalities are limited, and management decisions rely primarily on clinical judgment.

MedBrain's interface consisted of structured, dropdown-based questions with text and image options. At the end of each encounter, an open-text field allowed entry of additional relevant clinical or investigative information not captured by the predefined options, supporting iterative refinement of the tool.

To ensure data quality, data collectors were trained in using MedBrain and the study procedures. To further improve data quality, MedBrain was designed to offer predefined options and to allow only the required fields, thereby avoiding missing values, inconsistencies, and numerical errors. Furthermore, whenever there was an error or discrepancy in the collected data from patient medical records, the data were verified with the primary data source.

Statistical analysis

In preparation for statistical analysis, the collected data underwent data management procedures, including cleaning, transformation, and creation of new variables, to ensure data suitability. All data management and analysis were performed using SPSS software version 23.0. Descriptive statistics were used to summarize sociodemographic and clinical characteristics, presented as frequencies and percentages with 95% confidence intervals (CIs) where appropriate.

Triage accuracy of MedBrain was evaluated using healthcare professionals' triage as the gold standard. Patient urgency was classified into high urgency, priority, and non-urgent categories. The overall triage accuracy was determined by calculating the percentage of cases in which MedBrain's triage matched healthcare professionals' triage. Cases of disagreement were further categorized as over-triage (MedBrain assigning a higher urgency than staff) and under-triage (MedBrain assigning lower urgency than staff).

Diagnostic performance of MedBrain was assessed by measuring its accuracy and reliability with the pediatrician's presumptive diagnosis serving as the gold standard. Accuracy and reliability were evaluated for the total study population and for prevalent disease conditions, according to MedBrain's diagnostic ranks: Top 1, Top 2, and Top 3. Diagnostic accuracy was quantified using overall accuracy, sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), and negative likelihood ratio (LR–).Finally, reliability was assessed using Cohen's Kappa statistic with 95% CIs and interpreted according to conventional thresholds (≤0 = no agreement; 0.01–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–1.00 = almost perfect agreement).

Ethical Considerations

The study was conducted after obtaining ethical clearance from the AHRI/ALERT Ethics Review Committee (AAERC) (Protocol no: PO-05-24, 29/02/24, renewed on 28/03/25) and the Institutional Review Board of SPHMMC (Ref. no: pm23/386, 25/12/23). The study adhered to the ethical principles of the Declaration of Helsinki and conformed to Good Clinical Practice (GCP) guidelines.

For patients younger than 8 years, written informed consent was obtained from their legal guardian. For patients aged 8 years and above, both written informed consent from the legal guardian and assent from the patient were obtained. Consent and assent were sought after routine triage and stabilization, while patients awaited further medical assessment or intervention, by an independent data collector.

The study posed no major risks or negative consequences to participants. While there were no direct benefits to individual patients during the study period, the findings are expected to inform improvements in clinical practice in similar resource-limited settings, thereby enhancing pediatric care and outcomes. Anonymity was ensured byusing medical record numbers, with no personal identifiers included in the research report. Access to collected data was restricted to the investigators, and confidentiality was maintained throughout the study.

Result

Participant Recruitment and Profile of Excluded cases

A total of 1,247 patients were initially assessed for eligibility. Of these, 20 were excluded because they presented with severe medical emergencies requiring immediate care, five were excluded due to parental/guardian refusal, and 18 were excluded because the study team was unable to complete MedBrain assessments before clinical care proceeded during peak emergency department workload. This left 1,204 patients who were initially included in the study. Following pediatrician assessment, 274 were excluded because their diagnoses were not yet in the MedBrain database, result-

ing in a final sample of 930 eligible participants and a response rate of 77.4%. Of these, 450 were enrolled during the first phase of data collection and 480 during the second phase**(Figure 1).**

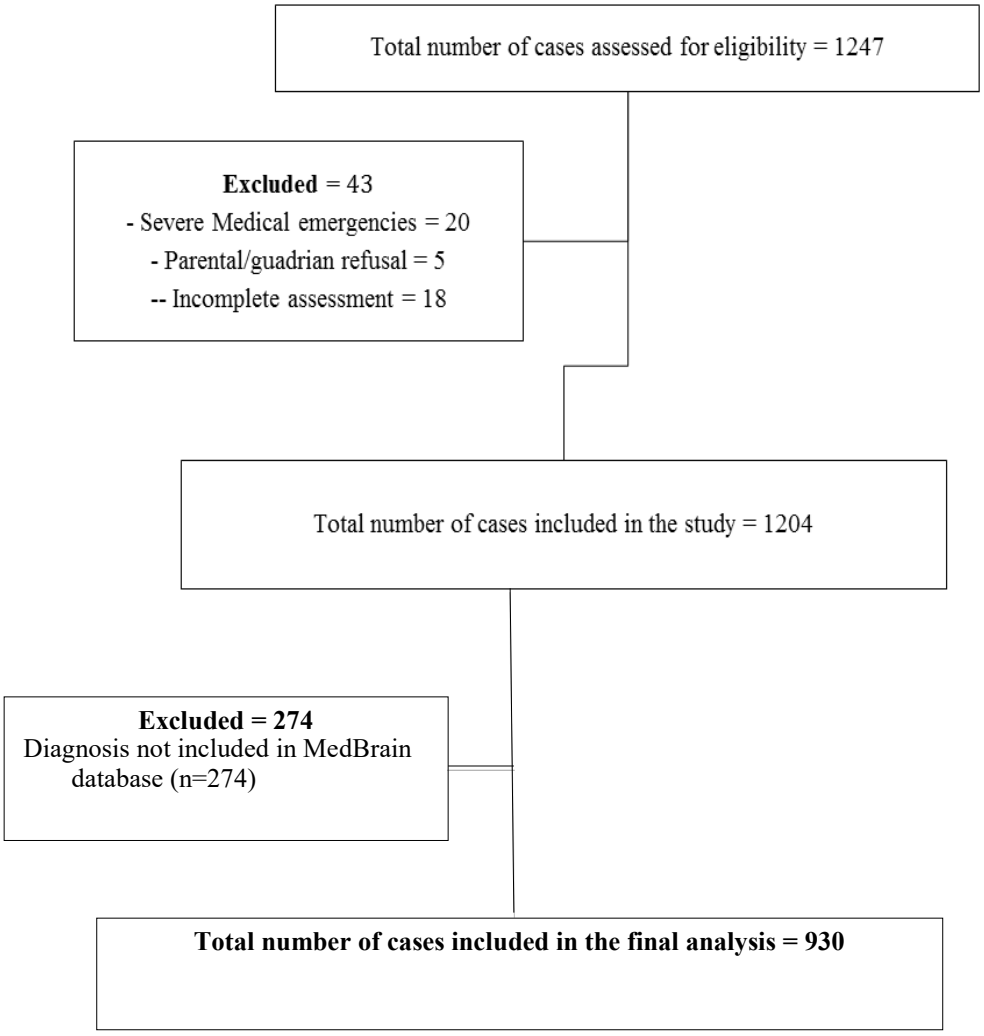

Figure 1: Flow chart showing disposition of pediatric patients included in the final analysis from the emergency departments of the two hospitals in Ethiopia.

Among the 274 excluded cases, several commonly observed conditions were reported, including hematologic malignancy (13.1%), Hirschsprung's disease (8.4%), acute hepatitis (5.1%), febrile seizure (5.1%), cerebral palsy/developmental delay (4.7%), obstructive sleep apnea (due to nasal polyps or tonsillar hypertrophy) (4.0%), retinoblastoma (4.0%), malaria (3.6%), acute kidney injury (3.3%), and cellulitis (3.3%). **(Figure 2)**

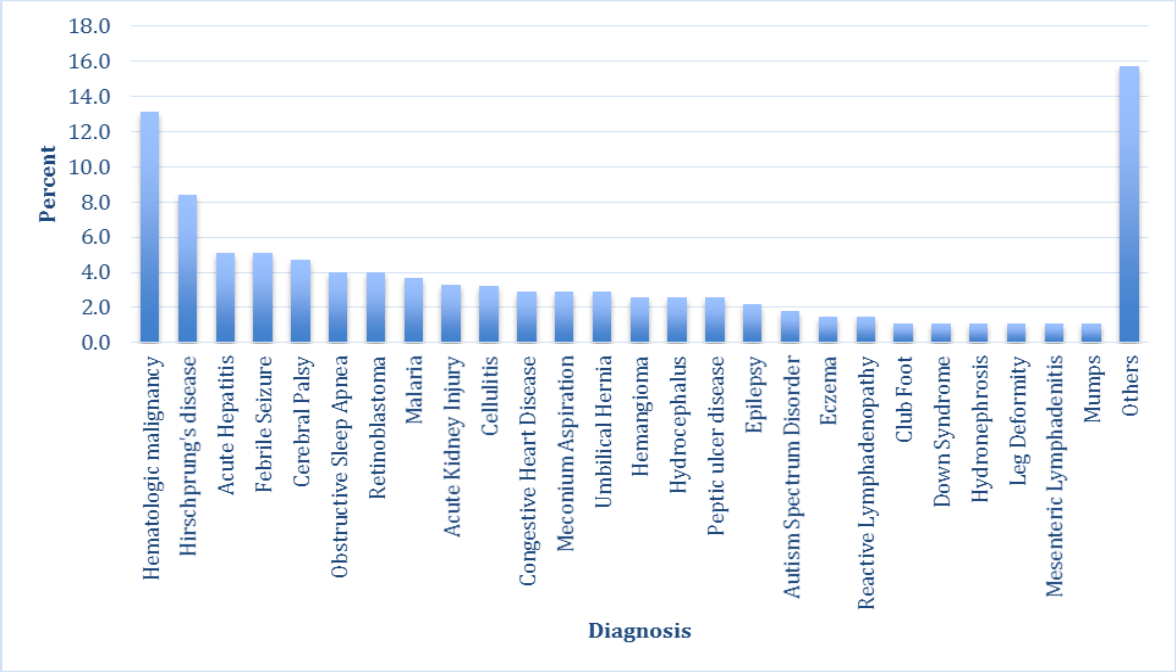

Figure 2. Diagnosis of excluded pediatric patients due to absence of diagnoses in the MedBrain database (n=274).

N.B. Others *= includes conditions with two or fewer cases (n=43), such as kwashiorkor, tonsillitis, mucocele, subgaleal hemorrhage, lymphadenopathy, congenital hydrocephalus, dog bite, hypoglycemia, neuroblastoma, and other rare conditions.*

Baseline Demographic and Clinical Characteristicsof Participants

Of the 930 included participants, 761 (81.8%) were from SPHMMC and 169 (18.2%) from ACSH. Most participants were infants (33.8%) or children under 5 years (31.9%). Neonates accounted for 14.4%, older children for 16.2%, and adolescents for 3.7%. The sex distribution showed a slight male predominance (56.5%)**(Table 1).**

Table 1: Baseline demographic and clinical characteristics of pediatric patients who presented to emergency departments of the two hospitals in Ethiopia(n=930).

Variable	Frequency	Percentage
Hospital		
ACSH	169	18.2
SPHMMC	761	81.8
Age group		
Neonate	134	14.4
Infant	314	33.8
Under 5	297	31.9
Child	151	16.2
Adolescent	34	3.7
Sex		
Male	525	56.5
Female	405	43.5

Pediatricians' top presumptive diagnoses identified pneumonia as the most common condition diagnosed in 190 cases (20.4%), followed by acute bronchitis in 120 (12.9%), late-onset neonatal sepsis (LONS) in 89 (9.6%), acute gastroenteritis in 74 (8.0%), bronchiolitis in 70 (7.5%), and meningitis in 48 (5.2%). **(Figure 3)**

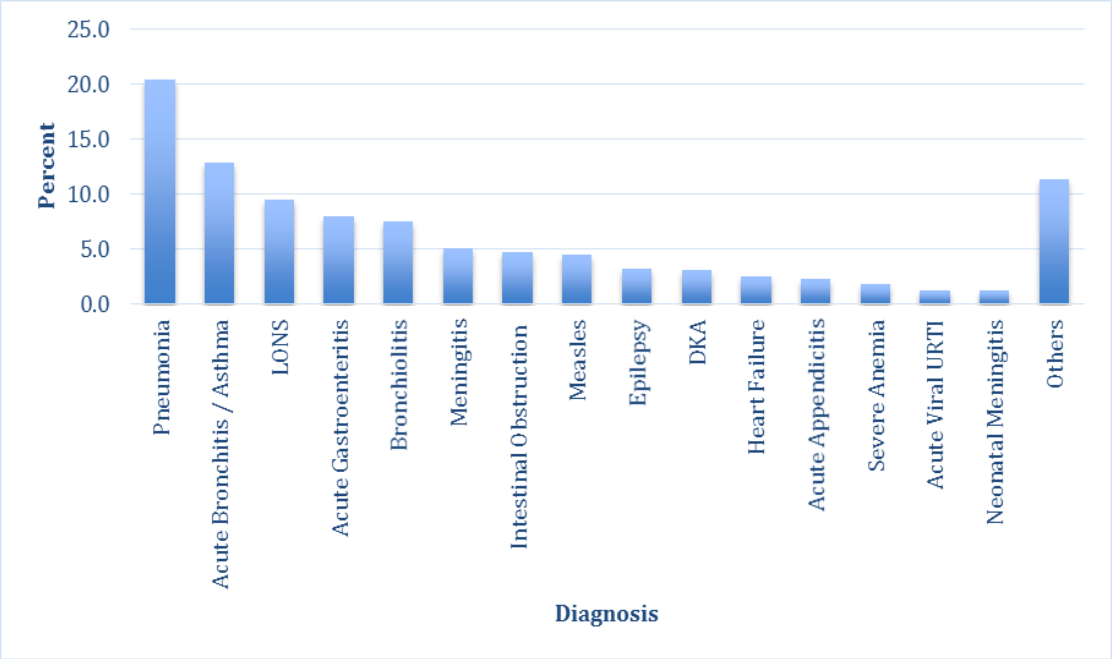

Figure 3: Distribution of diagnoses of the studied pediatric patients from emergency departments of the two hospitals in Ethiopia (n=930).

N.B.*LONS = Late Onset Neonatal Sepsis, DKA = Diabetic Ketoacidosis, Others = includes conditions with a frequency of < 1% (n=106), including candidiasis, pulmonary tuberculosis, tinea, early onset neonatal sepsis, pertussis, cystitis/lower urinary tract infection, malaria, rickets, and other conditions.*

MedBrain's Triage Accuracy

The triage results showed that 113 (12.2%) were triaged as high urgency, 286 (30.8%) as priority, and 531 (57.1%) as non-urgent cases. Agreement between MedBrain and healthcare professionals' triage was observed in 671 cases (72.2%).For the remaining cases, MedBrain over-triaged 34 cases (3.7%) and under-triaged 225 cases (24.2%).

MedBrain's Total Diagnostic Accuracy

Of the 930 cases, MedBrain's first diagnosis matched the pediatrician's presumptive diagnosis in 782 (84.1%) cases. MedBrain provided the correct diagnosis as a second and third diagnosis in an additional 69 (7.4%) and 17 (1.8%) cases, respectively. The remaining 62 (6.7%) cases were classified as failures. Thus, considering all disease conditions, the overall diagnostic accuracyof MedBrain was 84.1% (95% CI = 81.8, 86.3) forTop 1, 91.5% (95% CI: 89.8–93.4) for Top 2, and 93.3% (95% CI: 91.7–94.9) for Top 3 diagnosis.

Given that all 930 cases in this study had a confirmed pediatrician diagnosis, the dataset contained no true negatives (healthy cases). Hence, there were no false positives cases. Accordingly, MedBrain's overall Sn was 93.3% and PPV was 100%. However, overall specificity, NPV, LR+, and LR- could not be meaningfully calculated or interpreted because there were no true-negative or false-positive cases **(Table 2).**

Table 2: MedBrain's total diagnostic accuracy for pediatric patients from emergency departments of the two hospitals in Ethiopia (n=930).

Variable	Frequency	Percentage	95% CI
MedBrain's correct diagnosis rank			
1st position	782	84.1	81.9, 86.5
2nd position	69	7.4	5.8, 9.0
3rd position	17	1.8	1.0, 2.7
Failure (> 3 or no diagnosis)	62	6.7	5.2, 8.2
Overall diagnostic accuracy			
Top 1	782	84.1	81.8, 86.3
Top 2	851	91.5	89.8, 93.4
Top 3	868	93.3	91.7, 94.9
Sensitivity (Sn)	868	93.3	91.7, 94.9
Specificity (Sp)	-	-	-
Positive Predictive Value (PPV)	868	100	-
Negative Predictive Value (NPV)	-	-	-
Positive Likelihood Ratio (LR+)	-	-	-
Negative Likelihood Ratio (LR-)	-	-	-

MedBrain's Disease-Specific Diagnostic Accuracy and Reliability

The diagnostic performance of MedBrain was assessed for the six most prevalent conditions in the study population with prevalence exceeding 5%, including pneumonia, acute bronchitis, LONS, acute gastroenteritis, bronchiolitis, and meningitis. Across all conditions, diagnostic accuracy was high, with overall accuracy exceeding 97.4% and approaching 100% at the Top 3 rank.

A notable finding across all disease conditions was that Sp and PPV remained perfect at 100%, reflecting complete agreement between MedBrain and pediatricians in identifying negative cases.Across all disease conditions, a clear pattern of improving diagnostic performance was observed as the number of diagnoses considered increased. Sn at the first rank ranged from moderate to high (73.0–98.6%) but improved substantially with additional diagnostic ranks, reaching at least 90% for all conditions by the second rank and approaching or achieving 100% by the third. NPV was similarly strong, ranging from 97.1–99.9% at the first rank, increasing to 98.9 –100% at the second, and reaching near-perfect or perfect levels (99.2–100%) at the third. LR– values showed progressive improvement across ranks, starting higher at the first rank (0.014–0.270), decreasing at the second (0–0.100), and approaching zero at the third (0–0.079), showing MedBrain's strong capability in ruling out conditions when not listed among its top predictions.

Furthermore, the degree of agreement, as indicated by Cohen's Kappa values, similarly demonstrated almost perfect agreement, increasing from 0.830–0.993 at the first rank to 0.943–1.000 at the second rank, and remaining almost perfect, 0.955–1.000,at the third.
While the diagnosis of most of the prevalent conditions showed near-perfect performance by the second or third diagnostic rank, condition-specific variations were also observed: acute gastroenteritis achieved perfect diagnostic performance from the second rank onwards, whereas meningitis showed the most pronounced improvement between the first and second ranks.

Despite these high-performance rates, diagnostic failures were also recorded. The failure rate was high for LONS (0.8%), followed by bronchiolitis (0.5%), and pneumonia (0.4%). In contrast, MedBrain had no recorded failures for acute gastroenteritis. **(Table 3)**

Table 3: MedBrain's disease-specific diagnostic accuracy and reliability for pediatric patients from emergency departments of the two hospitals in Ethiopia (n=930).

Diagnosis	Diagnosis rank	Pediatrician and MedBrain Diagnostic agreement						Reliability
		Overall accuracy (%)	Sn	Sp	LR-	PPV	NPV	Cohen's Kappa (95% CI)
Pneumonia	Top 1	908 (97.6)	88.4	100.0	0.116	100.0	97.1	0.924 (0.893, 0.955)
	Top 2	922 (99.1)	95.8	100.0	0.042	100.0	98.9	0.973 (0.955, 0.991)
	Top 3	926 (99.6)	97.9	100.0	0.021	100.0	99.5	0.987 (0.973, 1.001)
	Failure	4 (0.4)						
Acute bronchitis	Top 1	915 (98.4)	87.5	100.0	0.125	100.0	98.2	0.924 (0.887, 0.961)
	Top 2	923 (99.2)	94.2	100.0	0.058	100.0	99.1	0.966 (0.941, 0.991)
	Top 3	927 (99.7)	97.5	100.0	0.025	100.0	99.6	0.985 (0.969, 1.001)
	Failure	3 (0.3)						
LONS	Top 1	906 (97.4)	73.0	100.0	0.270	100.0	97.2	0.830 (0.763, 0.897)
	Top 2	922 (99.1)	91.0	100.0	0.090	100.0	99.1	0.948 (0.913, 0.983)
	Top 3	923 (99.2)	92.1	100.0	0.079	100.0	99.2	0.955 (0.922, 0.988)
	Failure	7 (0.8)						
Acute Gastroenteritis	Top 1	929 (99.9)	98.6	100.0	0.014	100.0	99.9	0.993 (0.979, 1.007)
	Top 2	930 (100.0)	100.0	100.0	0	100.0	100.0	1.000
	Top 3	930 (100.0)	100.0	100.0	0	100.0	100.0	1.000
	Failure	0						
Bronchiolitis	Top 1	917 (98.6)	81.4	100.0	0.186	100.0	98.5	0.890 (0.831, 0.949)
	Top 2	923 (99.2)	90.0	100.0	0.100	100.0	99.2	0.943 (0.902, 0.984)
	Top 3	925 (99.5)	92.9	100.0	0.071	100.0	99.4	0.960 (0.925, 0.995)
	Failure	5 (0.5)						
Meningitis	Top 1	921 (99.0)	81.3	100.0	0.187	100.0	99.0	0.892 (0.821, 0.963)
	Top 2	927 (99.7)	93.8	100.0	0.062	100.0	99.7	0.966 (0.927,1.005)
	Top 3	927 (99.7)	93.8	100.0	0.062	100.0	99.7	0.966 (0.927, 1.005)
	Failure	3 (0.3)						

N.B.: *Positive likelihood ratio (LR+) is not calculated because specificity was 100% (1.0)for all disease conditions.*

Discussion

This study evaluated the diagnostic and triage performance of MedBrain, a digital decision support system (DDSS), for common pediatric conditions in the emergency departments of two large hospitals in Ethiopia. A total of 930 participants were included, nearly two-thirds (65.6%) under 5 years of age, with a slight male predominance (56.5%). The most prevalent conditions were pneumonia, acute bronchitis, late-onset neonatal sepsis (LONS), acute gastroenteritis, bronchiolitis, and meningitis. These findings align with epidemiological data showing that pneumonia and diarrheal diseases remain leading causes of morbidity and mortality among children under 5in Sub-Saharan Africa (1–3, 5). This indicates that the study population was representative of the pediatric disease burden in Ethiopia, strengthening the external validity of the findings.

MedBrain's triage performance showed 72.2% overall agreement with healthcare professionals, with 3.7% of cases over-triaged and 24.2% under-triaged. The relatively low over-triage rate suggests the tool is unlikely to overwhelm already resource-limited systems. However, the high under-triage rate raises safety concerns, as it could delay care for critically ill children, which is particularly dangerous in high-mortality settings like Ethiopia, where delays in treatment can rapidly worsen outcomes. This limitation likely reflects MedBrain's stronger focus on diagnostic reasoning than on evaluating the full range of clinical severity indicators considered by pediatricians. Compared to previous studies of DDSS and symptom checkers, which reported varied triage accuracy and frequent misclassification of urgent cases (6, 9, 18), MedBrain performed within the range of other tools but still requires improvement in its triage algorithms by incorporating broader clinical severity indicators.

In terms of diagnostic performance, MedBrain demonstrated strong accuracy and reliability. Its first-position diagnosis matched the pediatrician's in 84.1% of cases, rising to 91.5% for the top two diagnoses and 93.3% for the top three. The high sensitivity (93.3%) and perfect positive predictive value (100%) show that when MedBrain provided a diagnosis, it was highly likely to be correct. These findings are consistent with prior studies showing that DDSSs can improve clinician performance and reduce diagnostic error rates(13–16). This also aligns with results from internal validation of MedBrain, which reported high diagnostic accuracy in non-clinical case testing (34). Compared to online symptom checkers, which have shown much lower diagnostic accuracy (Top-1 accuracy ranging from19% to 36%) (6, 23), MedBrain demonstrated superior clinical performance, attributed to its specialized, pattern-based approach.

The disease-specific analysis further confirmed MedBrain's capacity to identify the most prevalent pediatric conditions with high accuracy. Across pneumonia, acute bronchitis, LONS, gastroenteritis, bronchiolitis, and meningitis, diagnostic agreement exceeded 97.4%, with nearly perfect performance (≥99%) by the second and third diagnostic ranks,indicating that MedBrain was very effective at identifying these diseases when they were present and demonstrating its strong capability to identify and rule out these diseases. In particular, MedBrain achieved perfect accuracy for acute gastroenteritis in the top two ranks, demonstrating its ability to correctly identify one of the most common and high-burden childhood diseases in Ethiopia. In contrast, symptom checkers in high -income settings have shown variable diagnostic accuracy for gastroenteritis and other infectious conditions, often underperforming compared to general practitioners(23, 24, 30). This suggests that MedBrain's targeted design for pediatric emergencies in low-resource settings is a key factor in its superior performance.

Moreover, MedBrain consistently showed perfect specificity and positive predictive value across all prevalent conditions, meaning it never falsely diagnosed a disease when it was not present. In clinical practice, this minimizes unnecessary treatments, thereby reducing overtreatment and resource wastage.The high negative predictive value (97%–100%) and low negative likelihood ratios further demonstrate its effectiveness at ruling out conditions when not suggested. Importantly, inter-rater reliability analysis showed almost perfect agreement between MedBrain and pediatricians, with Cohen's Kappa values ranging from 0.830 for LONS to 1.000 for acute gastroenteritis. This indicates that MedBrain's diagnostic outputs are both accurate and highly consistent with expert clinical judgment. Compared with prior DDSS studies, which often reported lower specificity, PPV, and clinician agreement due to false positives or inconsistent reasoning (6, 12, 18), MedBrain demonstrates a superior performance. Furthermore, the low failure rate, especially the absence of failures in diagnosing acute gastroenteritis, further shows its clinical utility.

Despite these strengths, the study also has important limitations. MedBrain excluded 274 patients with conditions not yet represented in its database, including malaria, which is endemic in Ethiopia. This indicates that the current version of the system is not comprehensive enough to cover all major or prevalent diseases. Similar concerns about limited disease coverage and generalizability have been raised in previous evaluations of DDSSs (10, 16, 17, 19). Furthermore, although the study population reflects the country's epidemiological profile, the study was conducted exclusively in urban hospitals. This ensured consistent access to pediatricians, who were necessary to establish a reliable gold standard, but it limits the generalizability of findings to rural areas, where patient profiles and disease prevalence may differ. MedBrain's performance, therefore, still requires validation in rural settings, where healthcare infrastructure is limited and mid-level professionals, the tool's intended users, provide frontline care.

Conclusion

The study demonstrated that MedBrain achieved high diagnostic accuracy and reliability in Ethiopian emergency settings for common acute pediatric conditions. Its diagnostic outputs were highly consistent with pediatricians' clinical judgment, with powerful performance for prevalent conditions such as pneu-

monia, acute gastroenteritis, and bronchiolitis. These findings suggest that MedBrain has the potential to strengthen clinical decision-making and improve diagnostic precision in emergency settings where timely and accurate diagnosis is critical and where mid-level professionals provide frontline care with limited resources.

However, the system showed limitations in triage performance, with a high under-triagerate, and its diagnostic library did not yet include several important conditions, such as malaria. Moreover, while the findings provide strong evidence of its utility in urban hospitals, further validation in rural health facilities is needed to confirm its effectiveness in the settings for which it was designed.

Decaration

Authors' Contribution: TWL, TH, and FL conceived and designed the study. SBD, TGG, and KTH contributed to the conception and design of the study. TWL performed statistical analysis and drafted the initial manuscript. TH, BWW, TS, TGH, FMA, and BTG contributed to the interpretation of the findings and drafting of the manuscript. SBD, TGG, and KTH revised the manuscript. SWB provided oversight and critical revisions. All authors approved the final version of the manuscript.

Competing Interests: The authors declare no competing interests. MedBrain, the Digital Decision Support System evaluated in this study, was developed and financially supported by MedBrain Global SL. The funder covered study-related costs, including payments for data collectors and facilitators, but had no role in the study design, data collection, statistical analysis, data interpretation, or manuscript preparation. Data was collected by independent nurses using the DDSS in clinical settings. For analysis, the principal investigator was granted direct access to the DDSS database to ensure data integrity. All datasets were anonymized before analysis.

Consent for publication: Not applicable

Availability of data and materials: All relevant data are available upon reasonable request from the corresponding author.

Funding: This study was financially supported by MEDBRAIN GLOBAL SL, the developer of MedBrain Digital Decision Support System.

Acknowledgment: The authors would like to thank the data collectors, supervisors, and facilitators at St. Paul's Hospital Millennium Medical College and Alert Comprehensive Specialized Hospital for their invaluable contributions to this study.

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10258):1204–22. doi:10.1016/S0140-6736(20)30925-9
2. 2.United Nations. The Sustainable Development Goals Report: Special Edition. New York: UN; 2023. Available from: https://unstats.un.org/sdgs/report/2023/
3. 3.Ethiopian Federal Ministry of Health. *Annual Performance Report 2015 EFY*. Addis Ababa: FMoH; 2023.
4. World Bank. *World Development Indicators*. Washington, DC: World Bank; 2023. Available from: https://datacatalog.worldbank.org
5. Ethiopian Public Health Institute (EPHI) [Ethiopia] and ICF. Ethiopia Mini Demographic and Health Survey 2019: Final Report. Rockville, Maryland, USA: EPHI and ICF, 2021.
6. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med. 2022;5:118. doi:10.1038/s41746-022-00667-w
7. orld Health Organization. Global strategy on human resources for health: Workforce 2030. Geneva: WHO; 2016.
8. World Health Organization. The state of the world's health workforce 2022: evidence to attain universal health coverage. Geneva: WHO; 2022.
9. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res. 2022;24(5):e31810. doi:10.2196/31810
10. Ben-Shabat N, Sloma A, Weizman T, Kiderman D, Amital H. Assessing the performance of a new artificial intelligence-driven diagnostic support tool using medical board exam simulations: clinical vignette study. JMIR Med Inform. 2021;9(11):e32507. doi:10.2196/32507
11. Moreno Barriga E, Irene P. Experience of Mediktor®, a new symptom checker based on artificial intelligence, in patients treated in an emergency department. Emergencias. 2017;29:391–6.
12. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics—results from an initial performance evaluation. Arch Dis Child. 2003;88(5):408–

13. doi:10.1136/adc.88.5.408
13. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA. 1999;282 (19):1851–6. doi:10.1001/jama.282.19.1851
14. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293(10):1223–38. doi:10.1001/jama.293.10.1223
15. Faqar-Uz-Zaman S, Anantharajah L, Baumartz P, Sobotta P, Filmann N, Zmuc D, et al. The diagnostic efficacy of an app-based healthcare application in the emergency room: eRadaR-trial. Ann Surg. 2022;276(5):935–42. doi:10.1097/SLA.0000000000005614
16. Elkin PL, Liebow M, Bauer BA, Chaliki S, Wahner-Roedler D, Bundrick J, et al. The introduction of a diagnostic decision support system (DXplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging DRGs. Int J Med Inform. 2010;79(11):772–7. doi:10.1016/j.ijmedinf.2010.09.004
17. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnosis (DDX) generators: a systematic review and meta-analysis. PLoS One. 2016;11 (3):e0148991. doi:10.1371/journal.pone.0148991
18. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar M, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res. 2023;25:e43803. doi:10.2196/43803
19. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One. 2021;16(7):e0254088. doi:10.1371/journal.pone.0254088
20. Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, et al. A comparison of artificial intelligence and human doctors for triage and diagnosis. Front Artif Intell. 2020;3:543405. doi:10.3389/frai.2020.543405
21. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. doi:10.1038/s41586-023-06291-2
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2 (2):e0000198. doi:10.1371/journal.pdig.0000198
23. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self-diagnosis and triage: audit study. BMJ. 2015;351:h3480. doi:10.1136/bmj.h3480
24. Shen C, Nguyen M, Gregor A, Isaza G, Beattie A. Accuracy of a popular online symptom checker for ophthalmic diagnoses. JAMA Ophthalmol. 2019;137(6):690.
25. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open. 2020;10:e040269.
26. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Med J Aust. 2020;212(11):514–8.
27. 27Berry AC, Cash BD, Mulekar MS, Charlton RC, Greenwald BD, Daskalakis C, et al. Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. Epidemiol Infect. 2019;147:e104.
28. Nazario Arancibia JC, Olivares F, Soto A, Astudillo P, Pino R. Evaluation of a diagnostic decision support system for the triage of patients in a hospital emergency department. Int J Interact Multimed Artif Intell. 2019;5:60–7.
29. Yoshida Y, Clark GT. Accuracy of online symptom checkers for diagnosis of orofacial pain and oral medicine disease. J Prosthodont Res. 2021;65(2):168–90.
30. Yu SWY, Ma MH, Tsai JC, Chiu TF, Chen SC, et al. Triage accuracy of online symptom checkers for accident and emergency department patients. Hong Kong J Emerg Med. 2020;27(4):217–22.
31. Cotte F, Mueller T, Gilbert S, Blümke B, Multmeier J, Hirsch M, et al. Safety of triage self-assessment using a symptom assessment app for walk-in patients in the emergency care setting: observational prospective cross-sectional study. JMIR Mhealth Uhealth. 2022;10(3):e32340. doi:10.2196/32340
32. Hageman MGJS, Anderson J, Blok R, Bossen JKJ, Ring D. Internet self-diagnosis in hand surgery. Hand (N Y). 2015;10(4):565–9.
33. Powley L, McIlroy G, Simons G, Raza K. Are online symptom checkers useful for patients with inflammatory arthritis? BMC Musculoskelet Disord. 2016;17:362.
34. MedBrain. Available from: https://medbrain.io/
35. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. doi:10.1136/bmj.h5527
36. World Health Organization. Interagency Integrated Triage Tool (IITT). Geneva: WHO; [cited 2025 Jan 5]. Available from: https://www.who.int/tools/triage