PIDSP JOURNAL
Pediatric Infectious Disease Society of the Philippines
Vol. 27 No. 1 · 2026
ISSN 2782-9510

ORIGINAL ARTICLE

External Validation of the Febrile Infants Risk Score at Triage (FIRST) for Early Detection of Serious Bacterial Infections in Infants Less than Three Months Old: A Retrospective Cohort Study in a Philippine Tertiary Emergency Department

Marione Thea B. Rodriguez, MD;[1] Soraya B. Sarapuddin, MD, DPPS[2]
[1]Institute of Pediatrics, The Medical City, Ortigas, Pasig City
[2]Neonatal Intensive Care Unit, The Medical City, Ortigas Avenue, Pasig City

Received: 06 MAR 2026 | **Accepted:** 12 MAR 2026 | **Published:** 26 JUN 2026 | **DOI:**.10.56964/pidspj20262701007

ABSTRACT

Objectives: Early identification of serious bacterial infection (SBI) in febrile infants younger than three months remains challenging. This study externally validated the diagnostic performance of the Febrile Infants Risk Score at Triage (FIRST) and the modified FIRST+ score for detecting SBI in a Filipino pediatric population.

Methodology: A single-center retrospective cohort study was conducted among febrile infants less than 3 months old presenting to a tertiary emergency department in the Philippines between January 1, 2020 and December 31, 2024. Medical records were reviewed to identify SBI, defined as urinary tract infection (UTI), bacteremia, or meningitis based on the final clinical diagnosis. The discrimination and calibration of FIRST and FIRST+ were evaluated using the original pre-specified cut-offs described in the derivation study. Sensitivity, specificity, likelihood ratios, calibration metrics, and decision curve analysis were also determined.

Results: A total of 248 infants were included, of whom 114 (46%) had SBI. UTI was the most common diagnosis (31%), followed by bacteremia (13%) and meningitis (1%). Several clinical and laboratory characteristics associated with SBI included male sex, higher admission temperature, maternal group B *Streptococcus* positivity, elevated absolute neutrophil count, and increased C-reactive protein levels. FIRST demonstrated poor discrimination (AUC 0.59) despite high sensitivity, while FIRST+ demonstrated good discrimination (AUC 0.84) and balanced sensitivity and specificity at their respective original cut-offs. Exploratory analysis suggested that a higher threshold (37.5) improved specificity and positive likelihood ratio, although this finding requires further validation.

Conclusion: FIRST+ demonstrated better discrimination and calibration than the triage-only FIRST score in this cohort of Filipino febrile infants. While FIRST+ may serve as a useful adjunctive risk stratification tool when laboratory testing is available, further prospective multicenter validation is necessary before broader clinical implementation.

KEYWORDS: Febrile Infants, Serious Bacterial Infection, Urinary Tract Infection, Bacteremia, Risk Assessment

INTRODUCTION

Fever is a common reason for emergency visits in infants younger than three months, with up to 10% presenting with serious bacterial infections (SBIs), including urinary tract infections (UTIs), bacteremia, and meningitis.[1,2] Early recognition is important because delayed diagnosis may lead to serious complications, while overly aggressive management may expose infants to unnecessary procedures and antibiotics. Physicians often adopt low thresholds for invasive procedures and antibiotic use, which may result in unnecessary costs, prolonged hospitalization, and antimicrobial resistance.[3]

CORRESPONDENCE:
Dr. Marione Thea B. Rodriguez
rodriguez.thea12@gmail.com

The authors declare that the data presented are original material and have not been previously published, accepted or considered for publication elsewhere; that the manuscript has been approved by all authors, and that the authors have met the requirements for authorship.

Cite this article as:
Rodriguez, MTB, Sarapuddin, SB. External Validation of the Febrile Infants Risk Score at Triage (FIRST) for Early Detection of Serious Bacterial Infections in Infants Less than Three Months Old: A Retrospective Cohort Study in a Philippine Tertiary Emergency Department. PIDSPJ. 2026;27(1):43-54. Available from: https://doi.org/10.56964/pidspj20262701007

In low- and middle-income countries such as the Philippines, the challenge of balancing early SBI detection with judicious resource utilization is particularly important. Limited access to rapid diagnostics, variability in laboratory turnaround time, and restricted healthcare resources may contribute to prolonged admissions and unnecessary antimicrobial exposure. Despite these challenges, local data on prediction tools for SBI among young febrile infants remain limited.

Over the years, several clinical prediction tools have been developed to aid decision-making in febrile infants, including the Pediatric Emergency Care Applied Research Network (PECARN) rule, the Invasive Bacterial Infection (IBI) score, and the Step-by-Step approach.[4–6]

The American Academy of Pediatrics also emphasizes risk stratification and selective laboratory evaluation in the management of well-appearing febrile infants.[7] Although useful, these models have important limitations, including age restrictions, reliance on laboratory tests that are not always accessible in resource-limited settings, and limited external validation outside the populations in which they were developed.

To address these gaps, Chong et al. developed the Febrile Infants Risk Score at Triage (FIRST) from a prospective Singaporean cohort of infants under three months old. It was designed as an early triage decision tool to identify febrile infants at risk for SBI using two components: FIRST, based on demographics, vital signs, and clinical history, and FIRST+, which incorporates urine leukocyte esterase and procalcitonin.[8] In the derivation cohort, FIRST demonstrated moderate discrimination (AUC 0.71), while FIRST+ demonstrated stronger performance (AUC 0.87).[8] However, prediction models may perform differently across populations because of variations in disease prevalence, healthcare systems, and clinical practices. External validation in Filipino infants has not yet been performed. This study therefore aimed to externally validate the FIRST and FIRST+ scoring systems among febrile infants younger than three months presenting to a tertiary emergency department in the Philippines. The primary objective was to evaluate model discrimination and calibration using the original pre-specified thresholds reported in the derivation study. Secondary objectives included assessment of diagnostic performance through sensitivity, specificity, predictive values, likelihood ratios, and decision curve analysis. Exploratory analyses were performed to examine alternative score thresholds and their potential diagnostic performance in this cohort.

METHODS

Study Design

The study is a retrospective cohort study that involved the review of electronic medical records of infants younger than three months who presented with fever (≥ 38.0°C or history within 24 hours) to The Medical City Emergency Department between January 1, 2020 and December 31, 2024.

Population and Sample Size

Included in the study are infants younger than three months who presented with fever (≥ 38.0°C or history within 24 hours), with complete clinical and laboratory records necessary to compute FIRST and FIRST+ scores, such as demographic characteristics (age, sex), clinical parameters (temperature, duration of fever, presence of respiratory distress), laboratory findings (absolute neutrophil count or ANC, urine leukocyte esterase, procalcitonin), and culture results (blood, urine, cerebrospinal fluid or CSF). Exclusion criteria were congenital anomalies, severe prematurity (< 32 weeks age of gestation), immunodeficiency, prior antibiotic therapy within 48 hours, confirmed viral infections without a secondary SBI, or incomplete data.

Prior to full data extraction, pilot testing of the data abstraction form and variable definitions was conducted using a subset of records to ensure consistency and completeness of data collection procedures.

A complete-case analysis approach was used because retrospective retrieval of missing laboratory variables from the electronic medical record was not feasible. Multiple imputation was considered; however, several important variables, particularly procalcitonin and C-reactive protein, had substantial non-random missingness related to clinical ordering practices, which may violate assumptions required for reliable imputation. Consequently, missing variables were neither replaced nor estimated. Excluding incomplete records may have introduced selection bias and potentially overestimated model performance.

Because this study primarily aimed to externally validate an existing prediction model rather than develop a new model, sample adequacy was interpreted in relation to the number of outcome events rather than conventional diagnostic test sample size computation. A total of 114 SBI events were available for validation analyses, allowing estimation of discrimination and calibration performance measures. The originally

planned sample size calculation based on diagnostic test methodology was retained for reference but interpreted cautiously given the differing methodological requirements for prediction model validation studies. The original sample size calculation estimated a minimum sample size of 267 patients, assuming a significance level of 5%, an anticipated AUC of 0.87, and a precision of 0.056 based on diagnostic test methodology.[8-10] However, interpretation of study adequacy was guided primarily by the number of SBI events available for validation.

The study protocol was reviewed and approved by The Medical City Institutional Review Board. Given the retrospective nature of the study involving review of existing medical records and minimal anticipated risk to participants, a waiver of informed consent was granted by the Institutional Review Board. Patient confidentiality was strictly maintained throughout data collection and analysis. All identifiable information was assigned a unique code, and data were stored in a password-protected database accessible only to the investigators.

Operational Definition of Terms

A febrile infant is defined as an infant younger than 90 days with a documented axillary or rectal temperature ≥38.0°C in the emergency department, or a history of fever within the preceding 24 hours. Serious Bacterial Infection (SBI) is a collective term for urinary tract infection (UTI), bacteremia, or bacterial meningitis, based on the final diagnosis documented in the medical record after review of all available clinical, laboratory, microbiologic, and imaging findings. UTI was defined as a positive urine culture with ≥ 100,000 colony-forming units per milliliter (CFU/mL) from a clean-catch urine specimen, and/or urinalysis findings suggestive of infection, such as pyuria, bacteriuria, positive leukocyte esterase, and/or positive nitrites. Bacteremia was defined as the isolation of a pathogenic organism in blood culture that was considered clinically relevant and not a contaminant. Bacterial meningitis was defined as the isolation of a pathogenic organism in CSF culture and/or a clinical diagnosis supported by CSF findings suggestive of bacterial infection (such as pleocytosis, elevated protein, or low glucose) or neuroimaging findings. Final SBI classification was determined by the investigator through review of microbiologic, laboratory, and clinical data documented in the medical record. Conversely, isolated positive cultures that the medical team assessed as contaminants and did not consider representative of true infection were not classified as SBI.

Data Collection

Data extracted from electronic records included variables required for computation of FIRST and FIRST+ scores and other clinical information such as demographic characteristics (age, sex), maternal factors (maternal Group B Streptococcus status), clinical parameters (temperature, duration of fever, heart rate, respiratory rate, presence of respiratory distress), laboratory findings (white blood cell count, absolute neutrophil count, hemoglobin, platelet count, C-reactive protein, procalcitonin, urine leukocyte esterase), microbiologic results (blood, urine, and cerebrospinal fluid cultures), and hospital outcomes (admission, ICU admission, antibiotic administration, and length of stay). Both FIRST (triage-only model) and FIRST+ (extended model including laboratory results) scores were calculated according to the original published scoring system.

Statistical Analysis

Descriptive statistics were used to summarize the general and clinical characteristics of the participants. Frequency and proportion were used for categorical variables (nominal/ordinal), while mean and standard deviation were reported for normally distributed interval/ratio variables, and median and interquartile range (IQR) for non-normally distributed interval/ratio variables.

Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC), while calibration was assessed through calibration plots, calibration slope, calibration intercept, expected-to-observed ratio, Brier score, and Hosmer–Lemeshow goodness-of-fit testing.

Sensitivity, specificity, predictive values, and likelihood ratios were calculated for the original published thresholds. Secondary exploratory analyses examined alternative thresholds using receiver operating characteristic analysis and Youden's J index. Comparative analyses between SBI and non-SBI groups were performed primarily for descriptive characterization of the external validation cohort. Decision curve analysis was additionally performed to estimate potential clinical net benefit across varying threshold probabilities. Statistical analyses were conducted using R version 4.1.3, with significance set at $\alpha = 0.05$.

RESULTS

Among 281 retrieved records of infants younger than three months who presented to the emergency department with fever (≥ 38.0°C or history within 24 hours), 32 were excluded, leaving 248 infants who met the eligibility criteria. **Table 1** presents the clinico-demographic and laboratory profiles of 248 febrile infants, of whom 134 (54.0%) had no SBI and 114 (46.0%) had SBI. Median age did not differ significantly between groups [60 days (IQR 30–87) vs. 54.5 days (IQR 22–84.5), p = 0.164], nor did age distribution (p = 0.147). Overall, the baseline characteristics of the two groups were not significantly different, except for a higher proportion of male sex in the SBI group (66.7% vs. 50.0%, p = 0.012). Infants with SBI presented with higher median temperatures [38.5°C vs. 38.2°C, p = 0.001], and maternal group B *Streptococcus* positivity was strongly associated with SBI (12.3% vs. 0.8%, p < 0.001).

Laboratory markers showed that ANC [5698 vs. 4443, p = 0.003] and C-reactive protein [10.9 mg/L vs. 0.6 mg/L, p = 0.017] were significantly higher in the SBI group, while total white blood cell counts (12.23 x10^9/L among infants with SBI vs 11.10 x10^9/L among infants without SBI, p = 0.077) and procalcitonin (0.38 ng/ml vs 0.14 ng/ml, p = 0.393) did not differ. FIRST scores were higher among SBI patients both at triage [47 vs. 43, p = 0.014] and after incorporation of laboratory results [47 vs. 27, p < 0.001]. Urine culture (68.4% vs. 0.8%, p < 0.001) and blood culture (17.5% vs. 1.5%, p < 0.001) positivity were strongly associated with SBI, while CSF culture yielded only one positive case (p = 0.409).

Hospital admission (93.9% vs. 76.1%, p < 0.001), intensive care unit (ICU) admission (7.0% vs. 0.8%, p = 0.013), antibiotic use (98.3% vs. 65.7%, p < 0.001), and length of hospital stay (4 vs. 3 days, p < 0.001) were all higher in the SBI group. Other parameters, including duration of fever, heart rate, respiratory rate, hemoglobin, and platelet count, showed no significant differences (all p > 0.05).

Table 1. Demographic, clinical and laboratory profile of febrile infants with and without SBI

	All (n = 248)	**Without SBI** (n = 134, 54.03%)	**With SBI** (n = 114, 45.97%)	**p-value**
		Median (IQR); Frequency (%)		
Age, days	60 (29-86)	60 (30-87)	54.50 (22-84.50)	.164§
<21 days	41 (16.53)	17 (12.69)	24 (21.05)	.147ψ
21-28 days	19 (7.66)	9 (6.72)	10 (8.77)	
>28 days	188 (75.81)	108 (80.60)	80 (70.18)	
Sex				**.012**ψ
Male	143 (57.66)	67 (50.00)	76 (66.67)	
Female	105 (42.34)	67 (50.00)	38 (33.33)	
Temperature, °C	38.30 (38-38.80)	38.20 (38-38.60)	38.50 (38.10-39)	**.001**§
Duration of Fever, days	1 (1-2)	1 (1-1)	1 (1-2)	.104§
Heart Rate, bpm	150 (140-165)	149.50 (140-165)	151 (140-164.80)	.730§
Respiratory Rate, bpm	42 (35.75-49)	42 (36-50)	42 (35-47.75)	.385§
Maternal GBS Status				**<.001**ψ
Negative	233 (93.95)	133 (99.25)	100 (87.72)	
Positive	15 (6.05)	1 (0.75)	14 (12.28)	
Total white blood cells, x10^9/L	11.46 (8.62-15.27)	11.10 (8.50-13.86)	12.23 (8.64-16.07)	.077§
Urine Leukocyte Esterase				**<.001**ψ
Negative	51 (20.56)	37 (27.61)	14 (12.28)	
Mild	12 (4.84)	5 (3.73)	7 (6.14)	
Moderate	18 (7.26)	3 (2.24)	15 (13.16)	
Strong	59 (23.79)	1 (0.75)	58 (50.88)	
Not Done	108 (43.55)	88 (65.67)	20 (17.54)	
ANC	4905.75 (2885.40-7191.27)	4442.90 (2682.43-6533)	5698.20 (3603.20-8084)	**.003**§
Hemoglobin, g/dL	112 (102-129)	112 (104-127)	112 (100-136)	.719§
Platelet Count	397.50 (320-510.20)	407 (322-508.20)	380 (314.80-518.50)	.631§
Procalcitonin, ng/mL (n = 11)	0.36 (0.10-4.26)	0.14 (0.09-1.94)	0.38 (0.23-4.89)	.393§
C-Reactive Protein, mg/L (n = 48)	5.59 (0.52-37.37)	0.60 (0.34-10.22)	10.91 (1.18-59.19)	**.017**§

FIRST Score (Triage)	43 (34-51)	43 (30-47)	47 (34-56)	**.014**§
FIRST+ Score (After Labs)	32 (23-46)	27 (18-33)	47 (34.25-63)	**<.001**§
Blood Culture				**<.001**ψ
Negative	122 (49.19)	60 (44.78)	62 (54.39)	
Positive	22 (8.87)	2 (1.49)	20 (17.54)	
Not Done	104 (41.94)	72 (53.73)	32 (28.07)	
Urine Culture				**<.001**†
Negative	4 (1.61)	3 (2.24)	1 (0.88)	
Positive	79 (31.85)	1 (0.75)	78 (68.42)	
Not Done	165 (66.53)	130 (97.01)	35 (30.70)	
CSF Culture				.409†
Negative	3 (1.21)	1 (0.75)	2 (1.75)	
Positive	1 (0.40)	0	1 (0.88)	
Not Done	244 (98.39)	133 (99.25)	111 (97.37)	
Hospital Admission	209 (84.27)	102 (76.12)	107 (93.86)	**<.001**ψ
ICU Admission	9 (3.63)	1 (0.75)	8 (7.02)	**.013** †
Antibiotic Admission	200 (80.65)	88 (65.67)	112 (98.25)	**<.001**ψ
Length of Hospital Stay, days	3 (2-7)	3 (1-5)	4 (3-7)	**<.001**§

Statistical Analysis Used: *–Independent t-test; §–Mann-Whitney; ψ–Chi-square test; †–Fisher's Exact test. SBI – Serious Bacterial Infection; GBS – Group B *Streptococcus*; ANC – Absolute Neutrophil Count; CRP – C-Reactive Protein; ICU – Intensive Care Unit; FIRST – Febrile Infants Risk Score at Triage.

Table A1 (see appendix) shows the distribution of SBI and their etiologies among the 248 infants. UTI was the most common SBI, diagnosed in 78 patients (31.5%), followed by bacteremia in 33 patients (13.3%) and meningitis in 3 patients (1.2%). The number of SBI diagnoses was not fully concordant with the number of positive microbiologic cultures because SBI classification was based on the predefined study definition using the final clinical diagnosis and all available supporting data, rather than culture results alone.

Among bacteremia cases, *Staphylococcus hominis* was the most frequently isolated pathogen (2.4%), followed by *Staphylococcus epidermidis* (1.2%), *Escherichia coli* and *Streptococcus agalactiae* (0.8% each). Less common isolates included *Staphylococcus aureus*, *Serratia marcescens*, *Pseudomonas aeruginosa*, *Streptococcus pyogenes*, *Enterobacter cloacae* complex, *Pasteurella multocida*, and *Sphingomonas paucimobilis* (0.4% each). Nearly half of blood cultures had no growth (48.8%). For meningitis, *Klebsiella pneumoniae* was the only identified organism (0.4%), while most samples yielded no growth (1.2%).

UTIs were predominantly caused by *E. coli* (18.9%) and *Klebsiella pneumoniae* (5.7%). Less frequent causes included *Enterococcus faecalis* (1.6%), *Enterobacter cloacae* complex (1.2%), *Citrobacter koseri* (0.8%), and *Staphylococcus haemolyticus* (0.8%). A wide variety of pathogens were detected in single cases (0.4% each), including *Proteus mirabilis*, *Staphylococcus epidermidis*, *Enterobacter aerogenes*, *Citrobacter amalonaticus*, Acine*tobacter junii*, *Klebsiella aerogenes*, *Klebsiella oxytoca*, and *Morganella morganii*. Four UTI cases (1.6%) showed no growth. Overall, UTI was the leading cause of SBI in this cohort, with *E. coli* as the predominant pathogen.

Table 2. Model calibration and discrimination

	FIRST Score	FIRST+ Score
E/O Ratio	1.000 (0.832-1.212)	1.000 (0.832-1.212)
Hosmer-Lemeshow test, p-value	.109	.645
Calibration slope (95% CI)	1.000 (0.21, 1.83)	1.00 (0.76, 1.28)
Calibration intercept (95% CI)	0.00 (-0.28, 0.29)	0.00 (-0.33, 0.34)
Brier score	0.242	0.150
C-statistic (Discrimination)	0.590	0.843

E/O = Expected-to-Observed ratio.

Table 2 presents the calibration and discrimination of the FIRST and FIRST+ scores for predicting SBI. Both models demonstrated acceptable calibration based on the Hosmer–Lemeshow test (FIRST: p = 0.109; FIRST+: p = 0.645). For FIRST+, the calibration intercept was 0.00 (95% CI: –0.33,0.34), suggesting no systematic over- or underprediction of risk. Its calibration slope was 1.00 (95% CI: 0.76,1.28), indicating that predicted risks closely matched observed outcomes without evidence of overfit or underfit.

The FIRST score also had values near the ideal, with an intercept of 0.00 (95% CI: –0.28,0.29) and a slope of 1.00 (95% CI: 0.21,1.83). However, the wider confidence intervals suggested less stable calibration compared with FIRST+. Consistency between expected and observed outcomes was further reflected in the E/O ratio, which was 1.00 (95% CI: 0.83,1.21) for both models.

Predictive accuracy, measured using the Brier score, was better for FIRST+ (0.150 vs. 0.242, where lower values indicate better accuracy). Discrimination was limited for FIRST (C-statistic 0.590) but substantially improved for FIRST+ (C-statistic 0.843), demonstrating good ability to distinguish between infants with and without SBI. These findings suggest that FIRST+ demonstrated substantially better calibration and discrimination than the triage-only FIRST score in this cohort.

Figure 1 shows the calibration curves of the FIRST (Panel a) and FIRST+ (Panel b) scores. The FIRST score demonstrated only reasonable agreement between predicted and observed probabilities, with noticeable deviation from the ideal line at higher predicted risk levels. In contrast, the FIRST+ score closely followed the ideal line across thresholds, indicating better agreement between predicted and observed risk. The low mean absolute error (0.022) further supports the good overall calibration of FIRST+ compared with the less stable calibration of the triage-only FIRST score.

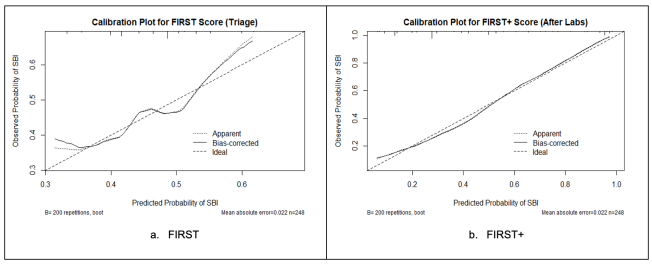

Figure 1. Calibration curves of FIRST (triage) and FIRST+ (after laboratory results)

Figure 2 shows the decision curve analysis for the FIRST and FIRST+ scores. Across a wide range of threshold probabilities, the FIRST+ score (after laboratory results) consistently provided higher net clinical benefit than the triage-only FIRST score. Both models outperformed the default "treat all" or "treat none" strategies, indicating that applying either score offers value over unselective management. Importantly, the superior performance of FIRST+ suggests that incorporating laboratory parameters improves risk stratification and helps clinicians target treatment to infants most likely to have SBI, thereby reducing unnecessary interventions while minimizing missed cases. The net benefit of FIRST+ remained higher than FIRST across clinically relevant thresholds between 10–50%, which corresponds to the range where most real-world admission and treatment decisions are made.

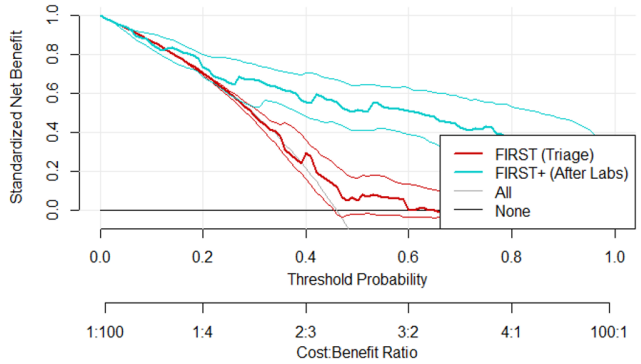

Figure 2. Decision curve in predicting Serious Bacterial Infection

Table A2 (see appendix) summarizes the diagnostic performance of the FIRST score at the pre-specified cut-off of ≥30. Sensitivity was 88.6% (95% CI: 81.3–93.8), meaning that approximately 9 in 10 infants with SBI were correctly identified. However, specificity was only 14.2% (95% CI: 8.8–21.3), indicating that nearly 9 in 10 infants without SBI were misclassified as having SBI. Positive predictive value (PPV), negative predictive value (NPV), and overall accuracy were limited (PPV 46.8%, NPV 59.4%, accuracy 48.4%). Both positive (LR+) and negative (LR–) likelihood ratios were uninformative (LR+ 1.03; LR– 0.80), suggesting minimal impact on post-test probability. Thus, although this cut-off ensures that most true cases are detected, the poor specificity and low discriminative ability substantially limit its clinical utility.

Table A3 (see appendix) presents the diagnostic performance of the FIRST+ score at the pre-specified cut-off of ≥36. Sensitivity was 71.9% (95% CI: 62.7,79.9), correctly identifying about 7 in 10 infants with SBI. Specificity improved substantially to 81.3% (95% CI: 73.7,87.6), meaning that about 8 in 10 infants without SBI were correctly classified. Predictive values were balanced, with a PPV of 76.6% and NPV of 77.3%, resulting in an overall accuracy of 77.0%. The LR+ was 3.86 (95% CI: 2.66,5.59), which increases the probability of SBI nearly fourfold when positive, while the LR– was 0.35 (95% CI: 0.25–0.47), which moderately reduces the probability of SBI when negative. These findings indicate that the FIRST+ cut-off of ≥36 provides a more clinically useful balance of sensitivity and specificity and is particularly helpful in confirming SBI once laboratory results are available.

Figure 3 shows the receiver operating characteristic (ROC) curves for the FIRST and FIRST+ scores. The ROC curve for FIRST closely followed the diagonal, with an AUC of 0.590, indicating discrimination only marginally better than chance (AUC

0.5). At its optimal cut-off of 32, the score prioritized sensitivity (84.2%) at the expense of specificity (30.6%), confirming its limited ability to accurately classify infants with and without SBI. In contrast, the ROC curve for FIRST+ bowed prominently toward the upper-left corner, yielding an AUC of 0.843. This represents good discrimination (AUC 0.8–0.9). At its optimal cut-off of 37.5, FIRST+ achieved a balanced profile, with moderate sensitivity (63.2%) and high specificity (94.0%). This suggests better discrimination and fewer false-positive classifications. Overall, the ROC analysis shows the limited discriminatory ability of the triage-only FIRST score, whereas FIRST+ showed good discrimination and improved ability to distinguish infants with and without SBI.

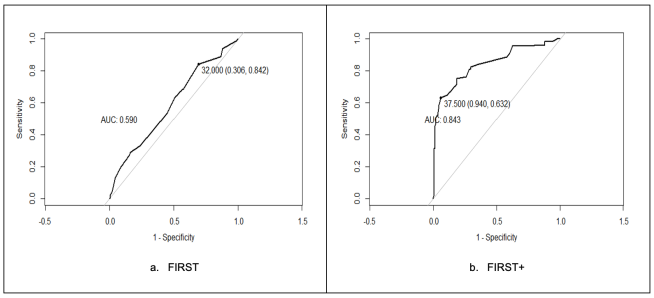

Figure 3. ROC curve of FIRST (triage) and FIRST+ (after laboratory results)

Table A4 (see appendix) shows the performance of the FIRST score at its data-driven optimal cut-off of ≥32. Sensitivity remained high at 84.2% (95% CI: 76.2,90.4), correctly identifying more than 8 in 10 infants with SBI. However, specificity improved only modestly to 30.6% (95% CI: 22.9,39.1), meaning that nearly 7 in 10 infants without SBI remained misclassified. Accuracy reached 55.2%, with a PPV of 50.8% and an NPV of 69.5%.

The likelihood ratios highlighted its limitations: LR+ was only 1.21 (95% CI: 1.06,1.39), which provides a negligible increase in post-test probability when positive, while LR– was 0.52 (95% CI: 0.31,0.85), which moderately decreases but does not rule out SBI when negative. The ROC area was 0.590 (95% CI: 0.519,0.657), confirming poor overall discrimination. Thus, even at its optimized cut-off, the FIRST score remains primarily sensitive but not specific, limiting its clinical utility due to weak likelihood ratios and high false-positive rates. Compared with the pre-specified cut-off of ≥30 (**Table A2**, see appendix), the optimized threshold of ≥32 (**Table A4**, see appendix) modestly improved specificity (14.2% to 30.6%) but did not achieve clinically meaningful discrimination, as likelihood ratios remained weak and the ROC area confirmed poor overall performance.

Table A5 (see appendix) shows the diagnostic performance of the FIRST+ score at its optimized cut-off of ≥37.5. Sensitivity declined to 63.2% (95% CI: 53.6,72.0), correctly identifying about 6 in 10 infants with SBI, but specificity increased markedly to 94.0% (95% CI: 88.6,97.4), meaning that nearly all infants without SBI were correctly classified. The PPV was very high at 90.0%, while the NPV was 75.0%, yielding an overall accuracy of 79.8%.

The likelihood ratios showed that LR+ was 10.58 (95% CI: 5.33,21.02), which provides strong confirmatory evidence of SBI when positive, whereas LR– was 0.39 (95% CI: 0.31,0.50), which only moderately reduces the probability when negative, limiting its utility as a rule-out test. The ROC area was 0.843 (95% CI: 0.792,0.890), reflecting good discrimination. Taken together, the optimized FIRST+ cut-off functions primarily as a confirmatory tool for SBI rather than as a screening instrument. Compared with the pre-specified cut-off of ≥36 (**Table A3**, see appendix), the optimized threshold of ≥37.5 further improved specificity (81.3% to 94.0%) and markedly strengthened LR+, making the score more clinically valuable for confirming SBI, though at the cost of reduced sensitivity.

The diagnostic performance of the FIRST score at triage was evaluated across varying cut-off points using Youden's J Index, which summarizes the trade-off between sensitivity and specificity (range –1 to 1, with higher values indicating better overall discrimination). The index was generally low at all thresholds, reflecting the poor discriminatory ability of the triage-only score. At very low cut-offs (≤23.5), sensitivity remained extremely high (≥88.6%) but specificity was near zero (<14.2%), producing J Index values close to zero. As the threshold increased, sensitivity declined while specificity rose modestly, with the best balance occurring at a cut-off of 32, where sensitivity was 84.2% and specificity 30.6%, yielding the highest J Index of only 0.1481. Beyond this point, further increases in specificity were offset by sharp losses in sensitivity, and J Index values remained consistently low (<0.13). Overall, these findings indicate that the FIRST score at triage has limited diagnostic value regardless of threshold, as no cut-off achieved a clinically meaningful balance of sensitivity and specificity.

In contrast, the FIRST+ score, which incorporates laboratory data, demonstrated substantially stronger discriminatory ability, with J Index values rising steadily as specificity improved. The optimal balance was achieved at a cut-off of 37.5, where sensitivity was 63.2% and specificity 94.0%, yielding the highest J Index of

0.5719. This indicates that FIRST+ achieved a meaningful balance between identifying true SBI cases and minimizing false positives. Other thresholds near this range also performed well, such as 33.5 (sensitivity 75.4%, specificity 81.3%, J = 0.5678) and 39.0 (sensitivity 60.5%, specificity 94.8%, J = 0.5530), indicative of the robustness of the score around the optimal zone. At very low cut-offs, sensitivity was preserved but specificity was poor, resulting in low J values, while very high cut-offs (>45) sharply reduced sensitivity with only marginal gains in specificity.

Overall, FIRST+ consistently outperformed the triage-only FIRST score. Whereas FIRST never achieved a J Index above 0.15, FIRST+ reached values above 0.50, highlighting its ability to meaningfully balance sensitivity and specificity. These results reinforce the clinical utility of the optimized FIRST+ cut-off around 37.5 as the most efficient threshold for confirming SBI.

DISCUSSION

In this retrospective external validation study involving 248 febrile infants younger than three months, FIRST+ demonstrated substantially better discrimination, calibration, and diagnostic utility than the triage-only FIRST score for identifying SBI. Infants with SBI more frequently required hospitalization, ICU admission, antibiotic therapy, and longer hospital stay, highlighting the clinical importance of early risk stratification in this population.

Male sex, higher admission temperature, elevated ANC, and increased C-reactive protein levels were commonly observed among infants with SBI. These findings should be interpreted as descriptive rather than independent predictive associations.

UTI was the leading SBI (31%), predominantly due to *E. coli*, but also involving *Klebsiella* and *Enterococcus* species. Bacteremia accounted for 13%, often with coagulase-negative staphylococci (*S. hominis, S. epidermidis*), while *Klebsiella pneumoniae* was the sole meningitis pathogen. The better performance of FIRST+ likely reflects the added value of laboratory parameters, particularly inflammatory biomarkers and urine leukocyte esterase, in refining SBI risk stratification. Prior meta-analyses have shown that procalcitonin-based approaches may improve specificity and overall discrimination in febrile infant risk stratification, supporting the improved performance observed with FIRST+ in this cohort.[11]

The relatively high prevalence of SBI observed in this cohort compared with prior derivation and validation studies likely reflects referral and spectrum bias associated with the tertiary referral nature of the institution and selective admission and microbiologic testing practices. As a result, these findings may not fully reflect febrile infants seen in lower-acuity settings.

The triage-only FIRST score performed poorly (AUC 0.590), with very high sensitivity but low specificity at prespecified cut-offs. By contrast, FIRST+ showed markedly better performance, with strong discrimination (AUC 0.843), superior accuracy (Brier score 0.150 vs. 0.242), and excellent calibration (slope 1.00, intercept 0.00). At an optimized cut-off of 37.5, FIRST+ achieved 94% specificity and a positive likelihood ratio >10, making SBI considerably more likely when the score was positive. Although FIRST+ demonstrated favorable calibration metrics, near-ideal calibration estimates are uncommon in external validation studies and may partly reflect optimistic performance associated with retrospective single-center datasets, complete-case analysis, and exploratory threshold optimization performed within the same cohort.

Decision curve analysis suggested that FIRST+ may provide greater potential net benefit across clinically relevant threshold probabilities compared with the triage-only FIRST score. At prespecified cut-offs, FIRST (≥30) remained limited by excessive false positives, whereas FIRST+ (≥36) demonstrated a more balanced sensitivity and specificity profile. Optimized analysis further distinguished the models: FIRST (≥32) retained poor specificity, while FIRST+ (≥37.5) had improved specificity, although this finding should be interpreted cautiously because the threshold was derived and evaluated within the same dataset, which may have resulted in optimistic performance estimates.

In this cohort, FIRST+ demonstrated better discrimination, calibration, and decision-analytic performance than the triage-only FIRST score. However, because this was a retrospective observational study, the findings do not establish that use of FIRST+ directly improves admission decisions, antibiotic stewardship, or patient outcomes. Accordingly, FIRST+ should be considered a supplementary risk stratification tool that may assist clinical assessment when laboratory testing is available, rather than a replacement for clinical judgment.

The findings of this study parallel those of Chong et al., who developed FIRST in Singapore.[8] Their cohort reported a lower SBI prevalence (22.4% vs. 46% in our study) and higher discrimination for the triage-only FIRST (AUC 0.71 vs. 0.590), while FIRST+ performed similarly (AUC 0.87 vs. 0.843). Our cohort demonstrated

substantially higher SBI prevalence and poorer performance of the triage-only FIRST score compared with the derivation cohort. These differences underscore the importance of external validation across healthcare systems with varying disease prevalence, referral patterns, microbiologic testing practices, and resource availability.

Finally, the results in this study echo broader patterns in pediatric SBI prediction tools such as PECARN, Step-by-Step, and the IBI score, which typically achieve high sensitivity but limited specificity in external cohorts. The improvement seen with FIRST+ supports prior evidence that incorporating laboratory data improves specificity and overall clinical utility, though it requires resource availability and timely access to testing.

This study has several important limitations. First, its retrospective single-center design may limit generalizability and introduce information bias related to medical record documentation. Second, the use of complete-case analysis and incomplete availability of laboratory variables, particularly procalcitonin and C-reactive protein, may have introduced selection bias and affected estimates of model performance. Third, the relatively high prevalence of SBI can be attributed to the spectrum and admission bias associated with a tertiary referral-center population. Fourth, exploratory optimization of diagnostic thresholds within the same dataset may have resulted in overfitting and optimistic performance estimates. Finally, incomplete microbiologic testing in some infants may have affected SBI classification.

CONCLUSION

FIRST+ demonstrated substantially better discrimination, calibration, and overall diagnostic performance than the triage-only FIRST score in identifying SBI among Filipino febrile infants younger than three months. In contrast, the triage-only FIRST score showed limited discriminatory ability and may not be sufficiently reliable as a standalone screening tool in similar settings.

FIRST+ may be a useful adjunctive tool when laboratory testing is available, although prospective multicenter studies are still needed before broader clinical use. The optimized thresholds identified in this study were exploratory and require prospective multicenter validation before broader clinical implementation. These findings support further evaluation of FIRST+ in larger and more diverse Filipino pediatric populations.

Future studies should focus on prospective multicenter external validation of FIRST and FIRST+ across diverse healthcare settings in the Philippines. Additional research may evaluate standardized laboratory testing strategies, methods for handling missing data, and potential recalibration of thresholds using independent validation cohorts. Implementation and cost-effectiveness studies may also help determine the feasibility of incorporating these tools into routine clinical workflows in resource-limited settings.

ACKNOWLEDGMENT

The authors thank The Medical City Health Informatics and Technology Group for access to patient records, and the Institute of Pediatrics for institutional support.

Conflicts of Interest

None declared.

REFERENCES

1. Mahajan P, VanBuren J, Tzimenatos L, Cruz AT, Vitale M, Powell EC, et al. Serious bacterial infections in young febrile infants with positive urinalysis results. Pediatrics. 2022;150(4):e2021055633. doi:10.1542/peds.2021-055633.
2. Pantell RH. Fever. In: Kliegman RM, St Geme JW III, Blum NJ, Shah SS, Tasker RC, Wilson KM, editors. Nelson Textbook of Pediatrics. 21st ed. Philadelphia: Elsevier; 2020. Chapter 201, p. 1386–1388.
3. Sutiman N, Khoo ZX, Ong GY, Piragasam R, Chong SL. Validation and comparison of the PECARN rule, Step-by-Step approach and Lab-score for predicting serious and invasive bacterial infections in young febrile infants. Ann Acad Med Singap. 2022;51(10):595–604. doi:10.47102/annals-acadmedsg.2022260.
4. Kuppermann N, Dayan PS, Levine DA, et al. A clinical prediction rule to identify febrile infants 60 days and younger at low risk for serious bacterial infections. JAMA Pediatr. 2019;173(4):342–351. doi:10.1001/jamapediatrics.2018.5501.
5. Aronson PL, Shabanova V, Shapiro ED, et al. A prediction model to identify febrile infants ≤60 days at low risk of invasive bacterial infection. Pediatrics. 2019;144(1):e20183604. doi:10.1542/peds.2018-3604.
6. Mintegi S, Bressan S, Gomez B, et al. Accuracy of a sequential approach to identify young febrile infants at low risk for invasive bacterial infection. Emerg Med J. 2014;31(1):e19–e24. doi:10.1136/emermed-2013-202449.
7. Pantell RH, Roberts KB, Adams WG, Dreyer BP, Kuppermann N, O'Leary ST, et al. Clinical practice guideline: evaluation and management of well-appearing febrile infants 8 to 60 days old. Pediatrics. 2021;148(2):e2021052228. doi:10.1542/peds.2021-052228.
8. Chong SL, Niu C, Ong GY, Piragasam R, Khoo ZX, Koh ZX, et al. Febrile infants risk score at triage (FIRST) for the

early identification of serious bacterial infections. Sci Rep. 2023;13(1):15845. doi:10.1038/s41598-023-42854-z.

9. Peacock JL, Peacock PJ. Research design. In: Oxford Handbook of Medical Statistics. 1st ed. Oxford: Oxford University Press; 2011. p. 60–61.
10. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014;48:193–204. doi:10.1016/j.jbi.2014.02.013.
11. Gomez B, Mintegi S, Bressan S, Da Dalt L, Gervaix A, Lacroix L, et al. Diagnostic value of procalcitonin in well-appearing young febrile infants. Pediatrics. 2012;130(5):815–822. doi:10.1542/peds.2012-0329.

APPENDIX

Table A1. Etiology of serious bacterial infections among pediatric patients

	Frequency (%)
Type of SBI	
Bacteremia	33 (13.31)
Meningitis	3 (1.21)
UTI	78 (31.45)
Etiology of Bacteremia	
Staphylococcus hominis	6 (2.42)
Staphylococcus epidermidis	3 (1.21)
Group B *Streptococcus* (GBS)	2 (0.81)
Escherichia coli (E. coli)	2 (0.81)
Streptococcus agalactiae	2 (0.81)
Staphylococcus aureus	1 (0.40)
Serratia marscescens	1 (0.40)
Pseudomonas aeruginosa	1 (0.40)
Streptococcus pyogenes	1 (0.40)
Enterobacter cloacae complex	1 (0.40)
Pasteurella multocida	1 (0.40)
Sphingomonas paucimobilis	1 (0.40)
No growth	121 (48.79)
Etiology of Meningitis	
Klebsiella pneumoniae	1 (0.40)
No growth	3 (1.21)
Etiology of UTI	
Escherichia coli (E. coli)	47 (18.95)
Klebsiella pneumoniae	14 (5.65)
Enterococcus faecalis	4 (1.61)
Enterobacter cloacae complex	3 (1.21)
Citrobacter koseri	2 (0.81)
Staphylococcus haemolyticus	2 (0.81)
Proteus mirabilis	1 (0.40)
Staphylococcus epidermidis	1 (0.40)
Enterobacter aerogenes	1 (0.40)
Citrobacter amalonaticus	1 (0.40)
Acinetobacter junii	1 (0.40)
Klebsiella aerogenes	1 (0.40)
Klebsiella oxytoca	1 (0.40)
Morganella morganii	1 (0.40)
No growth	4 (1.61)

UTI – urinary tract infection

Table A2. Diagnostic performance of the FIRST score (triage) in predicting Serious Bacterial Infection (pre-specified cut-off)

	With SBI	Without SBI	Total
	Frequency		
FIRST ≥30	101	115	216
FIRST <30	13	19	32
Total	114	134	248

Sensitivity (%, 95% CI)	88.60 (81.29, 93.79)	Positive LR (LR, 95% CI)	1.03 (0.94, 1.15)
Specificity (%, 95% CI)	14.18 (8.76, 21.25)	Negative LR (LR, 95% CI)	0.80 (0.42, 1.56)
PPV (%, 95% CI)	46.76 (39.96, 53.65)	Accuracy (%, 95% CI)	48.39 (42.02, 54.80)
NPV (%, 95% CI)	59.38 (40.64, 76.30)		

LR–Likelihood ratio; NPV–Negative predictive value; PPV–Positive predictive value

Table A3. Diagnostic performance of the FIRST+ score (after laboratory results) in predicting Serious Bacterial Infection (pre-specified cut-off)

	With SBI	Without SBI	Total
	Frequency		
FIRST+ ≥36	82	25	107
FIRST+ <36	32	109	141
Total	114	134	248

Sensitivity (%, 95% CI)	71.93 (62.74, 79.94)	Positive LR (LR, 95% CI)	3.86 (2.66, 5.59)
Specificity (%, 95% CI)	81.34 (73.70, 87.55)	Negative LR (LR, 95% CI)	0.35 (0.25, 0.47)
PPV (%, 95% CI)	76.64 (67.47, 84.27)	Accuracy (%, 95% CI)	77.02 (71.27, 82.10)
NPV (%, 95% CI)	77.30 (69.50, 83.93)		

LR–Likelihood ratio; NPV–Negative predictive value; PPV–Positive predictive value

Table A4. Diagnostic performance of the FIRST Score (triage) at the optimal cut-off for SBI

	With SBI	Without SBI	Total
	Frequency		
FIRST ≥32	96	93	189
FIRST <32	18	41	59
Total	114	134	248

Sensitivity (%, 95% CI)	84.21 (76.20, 90.37)	Positive LR (LR, 95% CI)	1.21 (1.06, 1.39)
Specificity (%, 95% CI)	30.60 (22.93, 39.14)	Negative LR (LR, 95% CI)	0.52 (0.31, 0.85)
PPV (%, 95% CI)	50.79 (43.44, 58.12)	Accuracy (%, 95% CI)	55.24 (48.82, 61.54)
NPV (%, 95% CI)	69.49 (56.13, 80.81)	ROC area (area, 95% CI)	0.590 (0.519, 0.657)

LR–Likelihood ratio; NPV–Negative predictive value; PPV–Positive predictive value

Table A5. Diagnostic performance of the FIRST+ score (after laboratory results) at the optimal cut-off for SBI

	With SBI	Without SBI	Total
	Frequency		
FIRST+ ≥37.50	72	8	80
FIRST+ <37.50	42	126	168
Total	114	134	248

Sensitivity (%, 95% CI)	63.16 (53.61, 72.00)	Positive LR (LR, 95% CI)	10.58 (5.33, 21.02)
Specificity (%, 95% CI)	94.03 (88.58, 97.39)	Negative LR (LR, 95% CI)	0.39 (0.31, 0.50)
PPV (%, 95% CI)	90.00 (81.24, 95.58)	Accuracy (%, 95% CI)	79.84 (74.30, 84.65)
NPV (%, 95% CI)	75.00 (67.75, 81.35)	ROC area (area, 95% CI)	0.843 (0.792, 0.890)

LR–Likelihood ratio; NPV–Negative predictive value; PPV–Positive predictive value