Table of Contents

Online Resource 1. Supplementary Information

Undertriage and Overtriage in Prehospital Mass Casualty Triage: A Systematic Review and Meta-Analysis

Supplementary Appendix S1
Full Search Strategies

Databases searched for the systematic review of undertriage and overtriage in prehospital mass-casualty-incident triage systems.

Searches were run on 2 March 2026 across bibliographic databases and trial-registry sources. The final combined strategy conceptually intersected three blocks: (1) prehospital MCI triage algorithms, (2) mass-casualty/disaster context, and (3) triage error/performance terms.

Web of Science Core Collection

Date run: 2 March 2026

Final combined set: #4

Results: 113

Set #1

TS=("mass casualty triage" OR "START triage" OR "SALT triage" OR "triage sieve" OR "MPTT" OR "modified physiological triage tool" OR "CareFlight triage" OR "Sacco triage" OR "JumpSTART" OR "prehospital triage" OR "field triage" OR "disaster triage")

Set #2

TS=("mass casualty" OR "mass casualty incident" OR "multiple casualty" OR "disaster" OR "major incident" OR "MCI")

Set #3

TS=("undertriage" OR "over-triage" OR "overtriage" OR "under-triage" OR "triage accuracy" OR "sensitivity" OR "specificity" OR "misclassification" OR "triage error" OR "triage performance" OR "triage validation")

Final query (#4)

#1 AND #2 AND #3

Cochrane CENTRAL

Date run: 2 March 2026

Final combined set: #10

Results: 24

Set #1

("mass casualty triage" OR "START triage" OR "SALT triage" OR "triage sieve" OR "MPTT" OR "modified physiological triage tool" OR "CareFlight triage" OR "Sacco triage" OR "JumpSTART" OR "prehospital triage" OR "field triage" OR "disaster triage"):ti,ab,kw

Set #2

[mh "Triage"]

Set #3

#1 OR #2

Set #4

("mass casualty" OR "mass casualty incident" OR "multiple casualty" OR "disaster" OR "major incident" OR "MCI"):ti,ab,kw

Set #5

[mh "Mass Casualty Incidents"] OR [mh "Disasters"]

Set #6

#4 OR #5

Set #7

("undertriage" OR "over-triage" OR "overtriage" OR "under-triage" OR "triage accuracy" OR "sensitivity" OR "specificity" OR "misclassification" OR "triage error" OR "triage performance" OR "triage validation"):ti,ab,kw

Set #8

[mh "Sensitivity and Specificity"]

Set #9

#7 OR #8

Final query (#10)

#3 AND #6 AND #9

CINAHL Ultimate (EBSCOhost)

Date run: 2 March 2026

Final combined set: S10

Results: 104

Set S1

TI ("mass casualty triage" OR "START triage" OR "SALT triage" OR "triage sieve" OR "MPTT" OR "modified physiological triage tool" OR "CareFlight triage" OR "Sacco triage" OR "JumpSTART" OR "prehospital triage" OR "field triage" OR "disaster triage") OR AB ("mass casualty triage" OR "START triage" OR "SALT triage" OR "triage sieve" OR "MPTT" OR "modified physiological triage tool" OR "CareFlight triage" OR "Sacco triage" OR "JumpSTART" OR "prehospital triage" OR "field triage" OR "disaster triage")

Set S2

(MH "Triage+")

Set S3

S1 OR S2

Set S4

TI ("mass casualty" OR "mass casualty incident" OR "multiple casualty" OR "disaster" OR "major incident" OR "MCI") OR AB ("mass casualty" OR "mass casualty incident" OR "multiple casualty" OR "disaster" OR "major incident" OR "MCI")

Set S5

(MH "Mass Casualty Incidents") OR (MH "Disasters+")

Set S6

S4 OR S5

Set S7

TI ("undertriage" OR "over-triage" OR "overtriage" OR "under-triage" OR "triage accuracy" OR "sensitivity" OR "specificity" OR "misclassification" OR "triage error" OR "triage performance" OR "triage validation") OR AB ("undertriage" OR "over-triage" OR "overtriage" OR "under-triage" OR "triage accuracy" OR "sensitivity" OR "specificity" OR "misclassification" OR "triage error" OR "triage performance" OR "triage validation")

Set S8

(MH "Sensitivity and Specificity+")

Set S9

S7 OR S8

Final query (S10)

S3 AND S6 AND S9

Scopus

Date run: 2 March 2026

Results: 103

Final query

TITLE-ABS-KEY("mass casualty triage" OR "START triage" OR "SALT triage" OR "triage sieve" OR "MPTT" OR "modified physiological triage tool" OR "CareFlight triage" OR "Sacco triage" OR "JumpSTART" OR "prehospital triage" OR "field triage" OR "disaster triage") AND TITLE-ABS-KEY("mass casualty" OR "mass casualty incident" OR "multiple casualty" OR "disaster" OR "major incident" OR "MCI") AND TITLE-ABS-KEY("undertriage" OR "over-triage" OR "overtriage" OR "under-triage" OR "triage accuracy" OR "sensitivity" OR "specificity" OR "misclassification" OR "triage error" OR "triage performance" OR "triage validation")

PubMed/MEDLINE

Date run: 2 March 2026

Final combined set: #10

Results: 273

Set #1

"mass casualty triage"[tiab] OR "START triage"[tiab] OR "SALT triage"[tiab] OR "triage sieve"[tiab] OR "MPTT"[tiab] OR "modified physiological triage tool"[tiab] OR "CareFlight triage"[tiab] OR "Sacco triage"[tiab] OR "JumpSTART"[tiab] OR "prehospital triage"[tiab] OR "field triage"[tiab] OR "disaster triage"[tiab]

Set #2

"Triage"[MeSH Terms]

Set #3

#1 OR #2

Set #4

"mass casualty"[tiab] OR "mass casualty incident"[tiab] OR "multiple casualty"[tiab] OR "disaster"[tiab] OR "major incident"[tiab] OR "MCI"[tiab]

Set #5

"Mass Casualty Incidents"[MeSH Terms] OR "Disasters"[MeSH Terms]

Set #6

#4 OR #5

Set #7

"undertriage"[tiab] OR "over-triage"[tiab] OR "overtriage"[tiab] OR "under-triage"[tiab] OR "triage accuracy"[tiab] OR "sensitivity"[tiab] OR "specificity"[tiab] OR "misclassification"[tiab] OR "triage error"[tiab] OR "triage performance"[tiab] OR "triage validation"[tiab]

Set #8

"Sensitivity and Specificity"[MeSH Terms]

Set #9

#7 OR #8

Final query (#10)

("mass casualty triage"[Title/Abstract] OR "START triage"[Title/Abstract] OR "SALT triage"[Title/Abstract] OR "triage sieve"[Title/Abstract] OR "MPTT"[Title/Abstract] OR "modified physiological triage tool"[Title/Abstract]

OR "CareFlight triage"[Title/Abstract] OR "Sacco triage"[Title/Abstract] OR "JumpSTART"[Title/Abstract] OR "prehospital triage"[Title/Abstract] OR "field triage"[Title/Abstract] OR "disaster triage"[Title/Abstract] OR "Triage"[MeSH Terms]) AND ("mass casualty"[Title/Abstract] OR "mass casualty incident"[Title/Abstract] OR "multiple casualty"[Title/Abstract] OR "disaster"[Title/Abstract] OR "major incident"[Title/Abstract] OR "MCI"[Title/Abstract] OR ("Mass Casualty Incidents"[MeSH Terms] OR "Disasters"[MeSH Terms])) AND ("undertriage"[Title/Abstract] OR "over-triage"[Title/Abstract] OR "overtriage"[Title/Abstract] OR "under-triage"[Title/Abstract] OR "triage accuracy"[Title/Abstract] OR "sensitivity"[Title/Abstract] OR "specificity"[Title/Abstract] OR "misclassification"[Title/Abstract] OR "triage error"[Title/Abstract] OR "triage performance"[Title/Abstract] OR "triage validation"[Title/Abstract] OR "Sensitivity and Specificity"[MeSH Terms])

ClinicalTrials.gov

Date run: 2 March 2026

Results: 6

Search structure: Trial-registry search

Condition/Disease

"mass casualty" OR "disaster triage"

Other terms

"triage" AND ("accuracy" OR "undertriage" OR "overtriage")

Supplementary Table 14
Full-text articles excluded after eligibility assessment, with primary reasons for exclusion

Each article was assigned a single primary reason for exclusion for transparent PRISMA reporting. Where more than one limitation was present, the reason judged most decisive for eligibility was retained.

No.	First author	Year	Title	Primary reason for exclusion
1	Sunyoto	2014	Providing emergency care and assessing a patient triage system in a referral hospital in Somaliland	Not prehospital/field/MCI triage context
2	Cheng	2022	Comparison of prehospital professional accuracy, speed, and interrater reliability of six pediatric triage strategies	No valid reference standard against patient outcomes
3	Hart	2018	Intuitive versus Algorithmic Triage	No valid reference standard against patient outcomes
4	de Visser	2026	Field Triage Errors: A Cross-Sectional Study of Emergency Responders in a Virtual Reality Mass-Casualty Simulation	No valid reference standard against patient outcomes
5	Cuttance	2017	Paramedic Application of a Triage Sieve: A Paper-Based Exercise	No valid reference standard against patient outcomes
6	Cicero	2017	Pediatric Disaster Triage: Multiple Simulation Curriculum Improves Prehospital Care Providers' Assessment Skills	No valid reference standard against patient outcomes
7	Cone	2011	Comparison of the SALT and Smart triage systems using a virtual reality simulator with paramedic students	No valid reference standard against patient outcomes
8	Claudius	2015	Comparison of Computerized Patients versus Live Moulaged Actors for a Mass-casualty Drill	No valid reference standard against patient outcomes
9	Cicero	2016	Comparing the Accuracy of Three Pediatric Disaster Triage Strategies: A Simulation-Based Investigation	No valid reference standard against patient outcomes
10	Behzadi	2024	Digitalisation of information and management optimisation in Multiple Victim Incidents: Analytical study	No valid reference standard against patient outcomes
11	Cone	2009	Pilot Test of the SALT Mass Casualty Triage System	No valid reference standard against patient outcomes
12	Dittmar	2018	Primary mass casualty incident triage: evidence for the benefit of yearly brief re-training from a simulation study	No valid reference standard against patient outcomes
13	Franc	2024	Repeatability, Reproducibility, and Diagnostic Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage Using the START Protocol	No valid reference standard against patient outcomes
14	Chumvanichaya	2025	A comparison of SIEVE, SORT, and START triage training effectiveness between immersive interactive 3D and conventional training environments	No valid reference standard against patient outcomes
15	Nordberg	2016	Primary Trauma Triage Performed by Bystanders: An Observation Study	Not prehospital/field/MCI triage context
16	Ingrassia	2015	Virtual reality and live simulation: a comparison between two simulation tools for assessing mass casualty triage skills	No valid reference standard against patient outcomes
17	Lee	2016	First Responder Accuracy Using SALT during Mass-casualty Incident Simulation	No valid reference standard against patient outcomes
18	Kman	2024	Virtual Reality Simulation for Assessment of Hemorrhage Control and SALT Triage Accuracy in First Responders	No valid reference standard against patient outcomes
19	Lee	2015	First Responder Accuracy Using SALT after Brief Initial Training	No valid reference standard against patient outcomes
20	Cicero	2012	Simulation Training with Structured Debriefing Improves Residents' Pediatric Disaster Triage Performance	No valid reference standard against patient outcomes
21	Cross	2012	Independent Application of the Sacco Disaster Triage Method to Pediatric Trauma Patients	Insufficient quantitative data for meta-analysis
22	Carenzo	2024	Factors affecting the accuracy of prehospital triage application and prehospital scene time in simulated mass casualty incidents	No valid reference standard against patient outcomes
23	Badiali	2017	Testing the START Triage Protocol: Can It Improve the Ability of Non-medical Personnel to Better Triage Patients During Disasters and Mass Casualty Incidents?	No valid reference standard against patient outcomes

No.	First author	Year	Title	Primary reason for exclusion
24	Cicero	2018	Correlation Between Paramedic Disaster Triage Accuracy in Screen-Based Simulations and Immersive Simulations	No valid reference standard against patient outcomes
25	Arshad	2015	A Modified Simple Triage and Rapid Treatment Algorithm from the New York City (USA) Fire Department	Insufficient quantitative data for meta-analysis
26	Alenyo	2018	A Comparison Between Differently Skilled Prehospital Emergency Care Providers in Major-Incident Triage in South Africa	No valid reference standard against patient outcomes
27	Mohamadi	2022	Comparison of Medical Students and Paramedics Using START and Sacco Triage Method in MCI: A Simulated Study	No valid reference standard against patient outcomes
28	Kman	2025	Virtual Reality Simulation for Assessment of Hemorrhage Control and SALT Triage Performance: A Comparison of Prehospital to In-Hospital Emergency Responders	No valid reference standard against patient outcomes
29	Sapp	2010	Triage Performance of First-Year Medical Students Using a Multiple-Casualty Scenario, Paper Exercise	No valid reference standard against patient outcomes
30	Vassallo	2019	Investigating the effects of under-triage by existing major incident triage tools	Insufficient quantitative data for meta-analysis
31	Vassallo	2014	UK Triage: the validation of a new tool to counter an evolving threat	Overlapping population with preferred study
32	Cheung	2012	The accuracy of existing prehospital triage tools for injured children in England: an analysis using trauma registry data	Overlapping population with preferred study
33	Vassallo	2018	Major incident triage and the implementation of a new triage tool, the MPTT-24	Overlapping population with preferred study
34	Cross	2015	A Better START for Low-acuity Victims: Data-driven Refinement of Mass Casualty Triage	Overlapping population with preferred study
35	Horne	2013	UK triage – An improved tool for an evolving threat	Overlapping population with preferred study
36	Wallis	2006	Validation of the Pediatric Triage Tape	Overlapping population with preferred study
37	Cross	2013	Head-to-Head Comparison of Disaster Triage Methods in Pediatric, Adult, and Geriatric Patients	Insufficient quantitative data for meta-analysis
38	Gebhart	2007	START Triage: Does It Work?	Insufficient quantitative data for meta-analysis
39	Jones	2014	Randomized Trial Comparing Two Mass Casualty Triage Systems (JumpSTART versus SALT) in a Pediatric Simulated Mass Casualty Event	No valid reference standard against patient outcomes
40	Khorram-Manesh	2023	The implication of a translational triage tool in mass casualty incidents: part three: a multinational study, using validated patient cards	No valid reference standard against patient outcomes
41	Lin	2020	Comparison between simple triage and rapid treatment and Taiwan Triage and Acuity Scale for the emergency department triage of victims following an earthquake-related mass casualty incident: a Retrospective cohort study	Not prehospital/field/MCI triage context
42	Mehralian	2023	Development and validation of SALT Triage method to facilitate the identification and classification of patients in Mass Casualty Incidents	No valid reference standard against patient outcomes
43	Paul	2009	Validierung der Vorsichtung nach dem mSTaRT-Algorithmus: Pilotstudie zur Entwicklung einer multizentrischen Evaluation	Non-English full text
44	Rehn	2010	A concept for major incident triage: full-scaled simulation feasibility study	No valid reference standard against patient outcomes
45	Schenker	2006	Triage Accuracy at a Multiple Casualty Incident Disaster Drill: The Emergency Medical Service, Fire Department of New York City Experience	No valid reference standard against patient outcomes
46	Vassallo	2017	The Prospective Validation of the Modified Physiological Triage Tool (MPTT): An Evidence-Based Approach to Major Incident Triage	Overlapping population with preferred study
47	Vassallo	2017	The civilian validation of the Modified Physiological Triage Tool (MPTT): an evidence-based approach to primary major incident triage	Overlapping population with preferred study
48	Wolf	2014	Evaluation of a novel algorithm for primary mass casualty triage by paramedics in a physician manned EMS system: a dummy based trial	No valid reference standard against patient outcomes
49	Xu	2024	Triage in major incidents: development and external validation of novel machine learning-derived primary and secondary triage tools	Overlapping population with preferred study

Supplementary Appendix S2
Study-arm level handling of overlap and eligibility-related arm selection

Overlap management was prespecified to operate at the level of the algorithm within a partially shared underlying study population, rather than automatically excluding all arms from a multi-arm publication. This was necessary because some studies contributed both unique and non-unique estimates. In the pediatric TARN literature, **Malik et al. (2021)**, **Vassallo et al. (2022)**, and **Price et al. (2016)** all evaluated multiple triage tools in registry-derived pediatric trauma populations, creating the potential for repeated inclusion of closely related patient populations and duplicate algorithm coverage.

For Vassallo et al. (2022), only the SPTT arm was included. Although that study reported several pediatric and adult triage tools, the remaining comparator arms were not included in the pooled analysis because they duplicated algorithm–registry combinations already represented in the overlapping pediatric TARN dataset analyzed by Malik et al. (2021). Therefore, retaining all Vassallo arms would have caused the same underlying pediatric TARN population to contribute multiple times to between-algorithm comparisons. The SPTT arm was preserved because it constituted the study's unique algorithmic contribution within the overlap set.

Price et al. (2016) retained only the Triage Sort. Although Price reported five pediatric tools from the TARN population, the other comparator arms were not retained because substantially overlapping pediatric TARN registry-derived algorithm estimates were already represented by Malik et al. (2021). Therefore, these overlapping arms were excluded from the pooled analyses to avoid duplicate registry–algorithm representation. Triage Sort was retained because it was the only algorithm reported by Price et al. (2016) that was not independently represented in Malik et al. (2021) pediatric TARN analysis. All other algorithms evaluated by Price—JumpSTART, START (>8 years), CareFlight, and PTT/Sieve—were also evaluated by Malik et al. using a substantially overlapping pediatric TARN population, and their inclusion would have resulted in double-counting.

A different issue arose in **Jerome et al. (2023)**. In that Alberta Trauma Registry analysis, both **BCD** and **MITT** satisfied the review's eligibility criteria and were therefore retained as distinct prespecified arms. However, the published results showed identical classification performance for the two tools. This is consistent with the source article's explanation that **BCD** and **MITT** use the same criteria except that **MITT classifies patients aged <2 years as Immediate**; in the analysed cohort, that distinction did not materially affect the resulting 2×2 table. Accordingly, both arms were retained in extraction but treated as non-independent estimates within the same study, and were therefore handled within the study-level clustering structure used in synthesis.

In Peng et al. (2021), arm selection was driven not by cohort overlap but by a prespecified eligibility. The study compared START, CareFlight, REMS, T-RTS, and TEWS using a retrospective earthquake casualty database. However, only START and CareFlight were retained because REMS, T-RTS, and TEWS are physiological or early warning scoring systems intended primarily for mortality/severity prediction and not structured field mass-casualty triage algorithms. The Peng et al. paper itself describes REMS, T-RTS, and TEWS as score-based methods using multiple physiological indicators, and its primary analyses focus on ROC/AUC performance for hospital death, ICU admission, and ISS >15 rather than algorithmic field triage categorization. Therefore, these three non-triage scores were excluded from the pooled algorithm-level analyses as they fell outside the prespecified scope of eligible MCI triage tools.

This arm-level selection strategy was preferred to whole-study exclusion because it preserved eligible information on algorithms not otherwise represented, while reducing artificial precision that would arise from repeated inclusion of substantially overlapping registry-derived cohorts or from mixing fundamentally different tool types within the same synthesis. The included analyzable arms are presented in **Supplementary Table 1.**

These study-arm retention decisions complement the study-level clustering approach used in the primary statistical synthesis and were intended to minimize duplicate representation without discarding unique, eligible algorithm-specific information. For clarity, explicit study-arm retention decisions related to overlap or eligibility-based restrictions are summarized in Supplementary Table 18.

Supplementary Table 18
Crosswalk of partially overlapping data sources and arm-retention decisions

Study	Underlying issue	Eligible or reported arms in source study	Arms retained for synthesis	Arms not retained	Explicit rationale for retention / exclusion
Price et al. (2016)	Partial overlap with the pediatric TARN literature	Five pediatric tools reported from the TARN population: Triage Sort, JumpSTART, START (>8 years), CareFlight, and PTT/Sieve	Triage Sort only	JumpSTART; START (>8 years); CareFlight; PTT/Sieve	The non-retained arms were excluded because substantially overlapping pediatric TARN registry-derived estimates were already represented by Malik et al. (2021). Triage Sort was retained because it was the only algorithm in Price et al. that was not independently represented in the overlapping Malik pediatric TARN analysis.
Vassallo et al. (2022)	Partial overlap with the pediatric TARN literature	Several pediatric and adult triage tools were reported	SPTT only	Remaining comparator arms	Comparator arms were not entered into the pooled analysis because they duplicated algorithm–registry combinations already represented in the overlapping pediatric TARN dataset analysed by Malik et al. (2021). SPTT was retained as the study's unique algorithmic contribution within that overlap set.
Jerome et al. (2023)	Within-study non-independence rather than cross-publication exclusion	Both BCD and MITT satisfied the review's eligibility criteria	BCD and MITT both retained as distinct prespecified arms	None	Both arms were retained because both were eligible, even though the published results showed identical classification performances. Identical 2×2 tables were handled as non-independent estimates within the same study-level clustering structure.
Peng et al. (2021)	Eligibility-based arm restriction	START, CareFlight, REMS, T-RTS, and TEWS were compared	START; CareFlight	REMS; T-RTS; TEWS	REMS, T-RTS, and TEWS were excluded because they were considered physiological or early warning scoring systems for mortality/severity prediction rather than structured field mass-casualty triage algorithms, and therefore, they fell outside the prespecified scope of eligible MCI triage tools.

Abbreviations: *MCI, mass-casualty incident; TARN, Trauma Audit and Research Network.*
Note. *This table summarizes the explicit study-arm retention decisions described in Supplementary Appendix S2 for source studies affected by partial population overlap, within-study non-independence, or eligibility-based arm restriction. Overlap management was prespecified to operate at the algorithm level within a partially shared underlying study population rather than by excluding entire studies. In the pediatric TARN literature, retention was used to avoid duplicate algorithm–registry representation while preserving uniquely eligible information. In Jerome et al. (2023), BCD and MITT were retained despite identical reconstructed performance because both satisfied the eligibility criteria and were handled as non-independent estimates within the same study. In Peng et al. (2021), arm restriction was driven by eligibility rather than by overlap. This table complements the narrative description provided in Supplementary Appendix S2.*

Supplementary Table 1
Full study-arm extraction fields and 2×2-derived data

Each row represents an analyzable study arm included in the multilevel meta-analysis. Undertriage (UT) and overtriage (OT) rates are shown as calculated proportions from the extracted 2×2 data.

First author	Year	Country	Algorithm family	Algorithm	Design	Data context	Population	Reference standard type	N total	N critical	N non-critical	TP	FP	FN	TN	Eligible reference standards reported in source study	Reference standard used for extracted 2×2	Collapsed reference-standard family (2-category)
Bhalla MC	2015	USA	START family	START	Retrospective	Registry-based	Adult	Hospital outcome/intervention timing	100	29	71	4	5	25	66	Hospital outcome / intervention timing	Hospital outcome / intervention timing	Non-intervention-based
Bhalla MC	2015	USA	SALT family	SALT	Retrospective	Registry-based	Adult	Hospital outcome/intervention timing	100	29	71	6	5	23	66	Hospital outcome / intervention timing	Hospital outcome / intervention timing	Non-intervention-based
Heller AR	2019	Germany	PRIOR family	PRIOR	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	45	203	5	239	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Heller AR	2019	Germany	START family	MSTART	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	39	88	11	354	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Heller AR	2019	Germany	START family	START	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	39	75	11	367	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Heller AR	2019	Germany	ASAV family	ASAV	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	39	84	11	358	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Heller AR	2019	Germany	FTS family	FTS	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	15	13	35	429	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Heller AR	2019	Germany	Sieve family	Triage Sieve	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	17	18	33	424	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Heller AR	2019	Germany	CareFlight family	CareFlight	Retrospective	Prehospital HEMS	Mixed	Weighted expert physician reference triage	492	50	442	35	57	15	385	Weighted expert physician reference triage	Weighted expert physician reference triage	Non-intervention-based
Jerome D	2023	Canada	START family	START	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	1112	980	1416	5144	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based
Jerome D	2023	Canada	START family	JumpSTART	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	784	674	1744	5450	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based
Jerome D	2023	Canada	SALT family	SALT	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	1036	857	1492	5267	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based
Jerome D	2023	Canada	RAMP family	RAMP	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	935	735	1593	5389	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based
Jerome D	2023	Canada	MPTT family	MPTT	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	1921	2817	607	3307	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based
Jerome D	2023	Canada	Sieve family	BCD	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	1770	2388	758	3736	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based

First author	Year	Country	Algorithm family	Algorithm	Design	Data context	Population	Reference standard type	N total	N critical	N non-critical	TP	FP	FN	TN	Eligible reference standards reported in source study	Reference standard used for extracted 2×2	Collapsed reference-standard family (2-category)
Jerome D	2023	Canada	Sieve family	MITT	Retrospective	Registry-based	Mixed	Urgent life-saving interventions (LSI)	8652	2528	6124	1770	2388	758	3736	Urgent life-saving interventions (LSI)	Urgent life-saving interventions (LSI)	Intervention/acuity-based
Heffernan RW	2019	USA	SALT family	SALT	Prospective	Hospital-based	Pediatric	Expert-panel-derived criterion standard categories	115	4	111	3	6	1	105	Expert-panel-derived criterion standard categories	Expert-panel-derived criterion standard categories	Non-intervention-based
Heffernan RW	2019	USA	START family	JumpSTART	Prospective	Hospital-based	Pediatric	Expert-panel-derived criterion standard categories	115	4	111	2	3	2	108	Expert-panel-derived criterion standard categories	Expert-panel-derived criterion standard categories	Non-intervention-based
Heffernan RW	2019	USA	Sieve family	Triage Sieve	Prospective	Hospital-based	Pediatric	Expert-panel-derived criterion standard categories	115	4	111	2	4	2	107	Expert-panel-derived criterion standard categories	Expert-panel-derived criterion standard categories	Non-intervention-based
Heffernan RW	2019	USA	CareFlight family	CareFlight	Prospective	Hospital-based	Pediatric	Expert-panel-derived criterion standard categories	115	4	111	2	4	2	107	Expert-panel-derived criterion standard categories	Expert-panel-derived criterion standard categories	Non-intervention-based
Challen K	2013	UK	START family	START	Retrospective	Real incident	Mixed	Modified Baxt/Upeniek criticality criteria	124	4	120	3	17	1	103	Modified Baxt/Upeniek criticality criteria	Modified Baxt/Upeniek criticality criteria	Intervention/acuity-based
Challen K	2013	UK	Sieve family	Triage Sieve	Retrospective	Real incident	Mixed	Modified Baxt/Upeniek criticality criteria	127	4	123	3	20	1	103	Modified Baxt/Upeniek criticality criteria	Modified Baxt/Upeniek criticality criteria	Intervention/acuity-based
Challen K	2013	UK	CareFlight family	CareFlight	Retrospective	Real incident	Mixed	Modified Baxt/Upeniek criticality criteria	128	4	124	3	21	1	103	Modified Baxt/Upeniek criticality criteria	Modified Baxt/Upeniek criticality criteria	Intervention/acuity-based
Price CL	2016	UK	Sort family	Triage Sort	Retrospective	Registry-based	Pediatric	ISS >15	31292	6842	24450	4830	5355	2012	19095	ISS >15	ISS >15	Non-intervention-based
Malik NS	2021	UK	Sieve family	BCD	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	1017	2099	326	1520	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	CareFlight family	CareFlight	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	543	188	800	3431	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	START family	JumpSTART	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	477	235	866	3384	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	Sieve family	MIMMS	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	504	442	839	3177	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	MPTT family	MPTT	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	795	2360	548	1259	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	MPTT family	MPTT-24	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	761	2200	582	1419	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	START family	MSTART	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	684	434	659	3185	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	Sieve family	NARU	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	649	767	694	2852	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	RAMP family	RAMP	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	533	185	810	3434	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	START family	START	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	1343	3619	661	413	682	3206	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Malik NS	2021	UK	Pediatric tape-based family	PTT_LT12Y	Retrospective	Registry-based	Pediatric	Pediatric LSI	2516	693	1823	310	233	383	1590	Pediatric LSI	Pediatric LSI	Intervention/acuity-based
Vassallo J	2022	UK	MPTT family	SPTT	Retrospective	Registry-based	Pediatric	Pediatric LSI	4962	875	4087	807	3592	68	495	Pediatric LSI	Pediatric LSI	Intervention/acuity-based

First author	Year	Country	Algorithm family	Algorithm	Design	Data context	Population	Reference standard type	N total	N critical	N non-critical	TP	FP	FN	TN	Eligible reference standards reported in source study	Reference standard used for extracted 2×2	Collapsed reference-standard family (2-category)
Peng Y	2021	China	START family	START	Retrospective	Registry-based	Mixed	Death/ICU/ISS>15 (composite)	29523	217	29306	119	1362	98	27944	Death/ICU/ISS>15 composite	Death/ICU/ISS>15 composite	Non-intervention-based
Peng Y	2021	China	CareFlight family	CareFlight	Retrospective	Registry-based	Mixed	Death/ICU/ISS>15 (composite)	29523	217	29306	113	1287	104	28019	Death/ICU/ISS>15 composite	Death/ICU/ISS>15 composite	Non-intervention-based
Tiyawat G	2024	USA	META family	META	Retrospective	Hospital-based	Adult	Lerner consensus criteria	3097	458	2639	353	1027	105	1612	RTS/Sort triage; ISS; Lerner consensus criteria	Lerner consensus criteria	Intervention/acuity-based
Tiyawat G	2024	USA	SALT family	SALT	Retrospective	Hospital-based	Adult	Lerner consensus criteria	3097	458	2639	314	744	144	1895	RTS/Sort triage; ISS; Lerner consensus criteria	Lerner consensus criteria	Intervention/acuity-based
McKee CH	2020	USA	SALT family	SALT	Prospective	Hospital-based	Adult	LSI	125	5	120	2	23	3	97	LSI / expert-derived life-saving intervention standard	LSI	Intervention/acuity-based
McKee CH	2020	USA	START family	START	Prospective	Hospital-based	Adult	LSI	125	5	120	4	11	1	109	LSI / expert-derived life-saving intervention standard	LSI	Intervention/acuity-based
McKee CH	2020	USA	Sieve family	Triage Sieve	Prospective	Hospital-based	Adult	LSI	125	5	120	3	7	2	113	LSI / expert-derived life-saving intervention standard	LSI	Intervention/acuity-based
McKee CH	2020	USA	CareFlight family	CareFlight	Prospective	Hospital-based	Adult	LSI	125	5	120	3	10	2	110	LSI / expert-derived life-saving intervention standard	LSI	Intervention/acuity-based
Lin YK	2022	Taiwan	START family	START	Retrospective	Hospital-based	Mixed	Expert outcome review	47	2	45	2	1	0	44	Expert outcome review	Expert outcome review	Non-intervention-based
Vassallo J	2019	UK	Sort family	Triage Sort	Retrospective	Registry-based	Adult	LSI	127233	24791	102442	3951	1398	20840	101044	LSI	LSI	Intervention/acuity-based
Vassallo J	2017	UK/Afghanistan	MPTT family	MPTT	Retrospective	Hospital-based	Adult	LSI	3654	1738	1916	1215	665	523	1251	LSI	LSI	Intervention/acuity-based
Vassallo J	2017	UK/Afghanistan	Sieve family	Military Sieve	Retrospective	Hospital-based	Adult	LSI	3654	1738	1916	761	123	977	1793	LSI	LSI	Intervention/acuity-based
Vassallo J	2017	UK/Afghanistan	Sieve family	Modified Military Sieve	Retrospective	Hospital-based	Adult	LSI	3654	1738	1916	885	240	853	1676	LSI	LSI	Intervention/acuity-based
Vassallo J	2017	UK/Afghanistan	Sieve family	Triage Sieve	Retrospective	Hospital-based	Adult	LSI	3654	1738	1916	431	102	1307	1814	LSI	LSI	Intervention/acuity-based
Vassallo J	2017	UK/Afghanistan	START family	START	Retrospective	Hospital-based	Adult	LSI	3654	1738	1916	673	59	1065	1857	LSI	LSI	Intervention/acuity-based
Vassallo J	2017	UK/Afghanistan	CareFlight family	CareFlight	Retrospective	Hospital-based	Adult	LSI	3654	1738	1916	582	31	1156	1885	LSI	LSI	Intervention/acuity-based
Kahn CA	2009	USA	START family	START	Retrospective	Real incident	Adult	LSI / Baxt criteria	148	2	146	2	20	0	126	LSI / Baxt criteria	LSI / Baxt criteria	Intervention/acuity-based
Malik NS	2021	UK	Sieve family	BCD	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	11194	27316	4706	52090	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	CareFlight family	CareFlight	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	6885	5717	9015	73689	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	START family	JumpSTART	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	7441	8496	8459	70910	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	Sieve family	MIMMS	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	6646	5241	9254	74165	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	MPTT family	MPTT	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	7934	32477	7966	46929	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	MPTT family	MPTT-24	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	7616	29460	8284	49946	LSI	LSI	Intervention/acuity-based

First author	Year	Country	Algorithm family	Algorithm	Design	Data context	Population	Reference standard type	N total	N critical	N non-critical	TP	FP	FN	TN	Eligible reference standards reported in source study	Reference standard used for extracted 2×2	Collapsed reference-standard family (2-category)
Malik NS	2021	UK	START family	MSTART	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	9095	8735	6805	70671	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	Sieve family	NARU	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	7139	9211	8761	70195	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	RAMP family	RAMP	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	6265	5320	9635	74086	LSI	LSI	Intervention/acuity-based
Malik NS	2021	UK	START family	START	Retrospective	Registry-based	Adult (16-64 years)	LSI	95306	15900	79406	8538	7226	7362	72180	LSI	LSI	Intervention/acuity-based
Wallis LA (a)	2006	South Africa	Pediatric tape-based family	PTT	Prospective	Hospital-based	Pediatric	Garner criteria	3461	200	3261	83	35	117	3226	ISS; NISS; Garner criteria	Garner criteria	Intervention/acuity-based
Wallis LA (b)	2006	South Africa	CareFlight family	CareFlight	Prospective	Hospital-based	Pediatric	Garner criteria	3461	200	3261	92	36	108	3225	ISS; NISS; Garner criteria	Garner criteria	Intervention/acuity-based
Wallis LA (b)	2006	South Africa	START family	JumpSTART	Prospective	Hospital-based	Pediatric	Garner criteria	2441	125	2316	1	53	124	2263	ISS; NISS; Garner criteria	Garner criteria	Intervention/acuity-based
Wallis LA (b)	2006	South Africa	START family	START	Prospective	Hospital-based	Pediatric	Garner criteria	1020	75	945	29	201	46	744	ISS; NISS; Garner criteria	Garner criteria	Intervention/acuity-based

Abbreviations: TP, true positive; FP, false positive; FN, false negative; TN, true negative.

Note. *In partially overlapping pediatric TARN studies, retention was performed at the registry–algorithm level; thus, Vassallo et al. (2022) contributed only SPTT, whereas Price et al. (2016) contributed only Triage Sort because the remaining comparator arms were already represented in substantially overlapping pediatric TARN analyses, principally Malik et al. (2021). In Jerome et al. (2023), both BCD and MITT were retained as non-independent estimates within the same study because they produced identical 2×2 tables in the Alberta Trauma Registry analysis. Peng et al. (2021) retained only START and CareFlight because REMS, T-RTS, and TEWS were considered ineligible physiological scoring systems rather than structured MCI triage algorithms. For studies reporting more than one eligible reference standard, the "Reference standard used for extracted 2×2" column identifies the source-level standard against which the analyzable contingency table was reconstructed. The final collapsed 2-category family used in the moderator analyses is shown separately to distinguish source-level operationalization from analytic grouping.*

Supplementary Table 2
Study-level narrative justification for revised QUADAS-2 judgments

A high risk was reserved for clearly biased design features. Incompletely reported but not clearly problematic methods were classified as unclear.

Study	Patient selection	Index test	Reference standard	Flow & timing	Overall
Wallis 2006a	Low: Prospective pediatric cohort in a relevant emergency setting with no clear spectrum restriction.	Low: Real-time index testing with explicit triage rule.	Low: The comparator acuity standard was explicit and clinically aligned.	Low: Flow and outcome ascertainment were complete.	Low: No material bias signals across domains.
Wallis 2006b	Low: Prospective pediatric cohort in a relevant emergency setting with no clear spectrum restriction.	Low: Real-time index testing with explicit triage rule.	Low: The comparator acuity standard was explicit and clinically aligned.	Low: Flow and outcome ascertainment were complete.	Low: No material bias signals across domains.
Kahn 2009	Unclear: Relevant incident sample, but inclusion and exclusion were incompletely described.	Low: The START-family test was clearly defined and not obviously outcome-driven.	Unclear: The Hybrid LSI/Baxt standard was plausible but insufficiently detailed.	Unclear: Case flow and verification completeness were not fully reported.	Unclear: Plausible methods but key reporting gaps remained.
Challen 2013	Unclear: Retrospective real-world cohort with incomplete sampling and exclusion details.	Unclear: Algorithms were recreated from recorded data with limited mapping details.	Unclear: The Modified Baxt/Upeniek criteria were plausible but not fully described.	Unclear: The included and excluded case flows were insufficiently documented.	Unclear: Under-reporting dominated more than a clear methodological failure.
Bhalla 2015	Unclear: Registry-based adult cohort, but sampling and exclusions were not fully described.	Unclear: Retrospective algorithm reconstruction was plausible but was incompletely operationalized.	High: Hospital outcome/intervention timing was an indirect proxy for field triage acuity.	Unclear: Flow details were insufficient to exclude attrition or verification concerns.	High: The overall judgment was driven by an indirect reference standard.
Price 2016	Low: The large pediatric registry cohort appeared to be broadly representative.	Unclear: Sort-family triage was reconstructed retrospectively from registry data.	Unclear: ISS>15 was explicit but only an indirect intervention proxy.	Low: Registry flow appeared stable without any major attrition.	Unclear: No high-risk domain, but uncertainty remained in the index test and reference standard.
Vassallo 2017	Low: The adult hospital cohort was clearly defined and clinically relevant.	High: Triage categories were computer-derived retrospectively rather than assessed in real time.	Low: The LSI-based standard was explicit and clinically appropriate.	Low: Flow and timing were acceptable.	High: The overall judgment was driven by index test reconstruction.
Heffernan 2019	Unclear: Prospective pediatric cohort, but recruitment details were limited.	Unclear: Multiple algorithms were applied, but the operational details were incomplete.	Unclear: Expert panel categories were plausible but insufficiently specified.	Low: The case flow appeared reasonably complete once enrolled.	Unclear: Several domains were uncertain rather than biased.
Heller 2019	Low: Mixed HEMS/prehospital cohort was relevant and appropriate.	Unclear: Computational reconstruction used a limited threshold and mapping details.	Unclear: Weighted expert-physician triage was relevant but insufficiently described.	Low: No clear major attrition signals.	Unclear: Uncertainty mainly stemmed from the reconstructed testing and expert reference standards.
Vassallo 2019	Low: A large adult registry cohort was relevant and likely representative.	Unclear: Sort-family assignments were derived retrospectively from the registry variables.	Low: The LSI-based standard was explicit and clinically appropriate.	Low: Registry flow supported low-risk judgment.	Unclear: The main residual concern was retrospective index-test reconstruction.
McKee 2020	Unclear: The prospective adult cohort was relevant, but recruitment and exclusions were incompletely reported.	Unclear: Algorithm operationalization from the collected data was not fully specified.	Unclear: LSI was appropriate, but the implementation details were limited.	Low: Patient flow was reasonably complete.	Unclear: Selection, index test, and reference standard reporting remained incomplete.
Malik 2021 ped	Low: A large pediatric registry cohort was relevant and broadly appropriate.	Unclear: Algorithms were generated from registry variables with incomplete coding details.	Low: The Pediatric LSI was explicit and well aligned with the target condition.	Low: Registry flow and timing appeared stable.	Unclear: Overall concern centered on retrospective index-test reconstruction.
Malik 2021 adult	Low: A large adult registry cohort was relevant and broadly appropriate.	Unclear: Algorithms were generated from registry variables with incomplete coding details.	Low: The Adult LSI was explicit and well aligned with the target condition.	Low: Registry flow and timing appeared stable.	Unclear: Overall concern centered on retrospective index-test reconstruction.

Study	Patient selection	Index test	Reference standard	Flow & timing	Overall
Peng 2021	Low: A large mixed registry cohort was relevant and reasonably representative.	Unclear: Index test reconstruction from registry variables was plausible but incomplete.	High: Mortality alone is an inadequate proxy for immediate triage acuity.	Low: Flow/timing appeared acceptable at the registry level.	High: The overall judgment was driven by the mortality-based reference standard.
Lin 2022	Unclear: Small mixed hospital cohort with limited details on inclusion and exclusion.	Low: START-family index was clearly identifiable and not obviously data-driven.	Unclear: Expert outcome review was plausible but insufficiently detailed.	Unclear: The flow and completeness of the analyzed cases were not fully reported.	Unclear: The main issue was under-reporting rather than a clear systematic bias.
Vassallo 2022	Low: The pediatric registry cohort was relevant and likely representative.	Unclear: MPTT-family allocation was retrospectively derived from the registry variables.	Low: The Pediatric LSI standard was explicit and clinically appropriate.	Low: Flow and timing were acceptable.	Unclear: Only the reconstructed index test remained unclear.
Jerome 2023	Low: The large mixed registry-based trauma cohort was relevant and broadly representative.	Unclear: Algorithms were assigned retrospectively based on the available variables.	Low: The Urgent LSI was explicit and well aligned with the review question.	Low: Registry flow was acceptable without any verification problems.	Unclear: Limited reporting of retrospective algorithm mapping remains the main concern.
Tiyawat 2024	Unclear: Adult hospital cohort with incomplete recruitment and exclusion details.	Unclear: META and SALT-family applications from recorded data lacked operational details.	Low: The Lerner consensus-based acuity criteria were explicit and clinically relevant, although the operationalization of recorded variables remained indirect.	Low: Flow/timing appeared acceptable once the cases were analyzed.	Unclear: Selection and index test reporting were insufficient for a low-risk judgment.

Abbreviations: LSI, life-saving intervention; HEMS, helicopter emergency medical service; ISS, Injury Severity Score.

Supplementary Table 13
QUADAS-2 applicability concerns for included studies

Study	Patient Selection	Index Test	Reference Standard	Overall
Bhalla MC 2015	High	High	High	High
Challen K 2013	Some concerns	High	Some concerns	High
Heffernan RW 2019	High	High	High	High
Heller AR 2019	High	High	High	High
Jerome D 2023	High	High	Low	High
Kahn CA 2009	Low	Low	High	High
Lin YK 2022	Low	Low	High	High
Malik NS 2021 (pediatric)	Low	High	Low	High
Malik NS 2021 (adult)	Low	High	Low	High
McKee CH 2020	High	High	High	High
Peng Y 2021	High	High	High	High
Price CL 2016	Low	High	High	High
Tiyawat G 2024	Low	High	High	High
Vassallo J 2017	High	High	Low	High
Vassallo J 2019	Low	High	Low	High
Vassallo J 2022	Low	High	Low	High
Wallis LA 2006a	Low	Low	Low	Low
Wallis LA 2006b	Low	Low	Low	Low

Note: *Applicability was assessed across three QUADAS-2 domains: patient selection, index test, and reference standard. Each domain was judged as having low, some, or high concern regarding the applicability of the study to the review question.*

Patient selection*: High concern was assigned when the study population did not closely match a prehospital MCI setting (e.g., single-center emergency department cohorts and highly selected registry subsets).*

Index test: *High concern was assigned when triage algorithms were reconstructed retrospectively from registry or chart data rather than applied prospectively in a real or simulated pre-hospital environment.*

Reference standard: *High concern was assigned when the reference standard did not directly reflect the clinical construct of immediate triage priority in the MCI context (e.g., ISS-only thresholds without intervention-based criteria).*

Overall: *Overall applicability was judged to be of high concern if any individual domain was rated as high concern.*

Summary*: Low concern was assigned to 10/18 (patient selection), 4/18 (index test), and 8/18 (reference standard) studies. The index test domain showed the highest proportion of applicability concerns, reflecting the predominance of retrospective algorithm applications in the included evidence base.*

Supplementary Table 3
Sensitivity analysis summary

Sensitivity analyses for pooled undertriage (UT) and overtriage (OT) proportions from multilevel models.

Sensitivity analysis	Clusters	Undertriage			Overtriage			Interpretation
		Pooled (95% CI)	95% PI	σ^2	Pooled (95% CI)	95% PI	σ^2	
Main analysis	18	0.453 (0.333–0.579)	0.095–0.868	0.894	0.131 (0.067–0.241)	0.006–0.790	2.204	Reference values for all subsequent sensitivity checks.
Overall RoB low + unclear only	15	0.415 (0.285–0.558)	0.078–0.857	0.9	0.143 (0.064–0.288)	0.005–0.851	2.543	Most balanced quality-restricted analysis; direction preserved for both outcomes.
Flow & timing low only	14	0.449 (0.322–0.583)	0.095–0.863	0.835	0.142 (0.060–0.300)	0.004–0.867	2.694	Robustness was maintained; heterogeneity persisted despite the lower-risk flow/timing subset.
Patient selection low only	11	0.468 (0.309–0.633)	0.079–0.899	0.993	0.146 (0.049–0.363)	0.003–0.917	3.205	The pooled direction was retained, but the OT remained highly unstable.
Reference standard low only	9	0.477 (0.278–0.683)	0.057–0.932	1.238	0.167 (0.042–0.476)	0.002–0.960	3.864	Very selective subset; wide uncertainty limits the interpretability.
Index test low only	4	0.579 (0.480–0.672)	0.480–0.672	0.0	0.049 (0.007–0.285)	0.000593–0.815	1.502	Sparse subset; treat cautiously because only four clusters remain.
Strict LSI-only	8	0.444 (0.229–0.682)	0.041–0.938	1.37	0.230 (0.065–0.560)	0.004–0.958	3.004	UT was stable, whereas OT rose under the strictest intervention-based standard.
Expanded LSI-based	12	0.449 (0.290–0.619)	0.071–0.897	1.046	0.160 (0.061–0.357)	0.004–0.900	2.83	UT remained unchanged, whereas OT was attenuated when Garner/Baxt-related standards were added.
Adult-only	7	0.564 (0.315–0.785)	0.077–0.953	1.033	0.201 (0.00006–0.999)	0.000–1.000	198.700	Undertriage increased relative to the main analysis, overtriage remained highly unstable with extremely wide uncertainty, and substantial residual heterogeneity persisted.

Abbreviations: CI, confidence interval; LSI, life-saving intervention; PI, prediction interval; RoB, risk of bias.

Note. *The index test low-only subset was retained for completeness but should be interpreted cautiously because only four clusters remained. Across all restrictions, the direction of the pooled estimates was generally preserved, whereas residual heterogeneity remained substantial. The "Strict LSI-only" and "Expanded LSI-based" subsets were defined according to the original source-level intervention-oriented reference-standard classification and should not be interpreted as equivalent to the revised 2-category collapsed reference-standard moderator used in the Supplementary Table 6. The clusters correspond to studies in the multilevel models.*

Supplementary Table 17
Sensitivity analysis excluding tools more commonly positioned outside immediate primary scene triage

Analysis	Arms (k)	Clusters (n)	Undertriage pooled (95% CI)	Undertriage 95% PI	Overtriage pooled (95% CI)	Overtriage 95% PI	Interpretation
Main analysis	67	18	0.453 (0.333–0.579)	0.095–0.868	0.131 (0.067–0.241)	0.006–0.790	Reference analysis
Excluding Triage Sort and SPTT	64	15	0.485 (0.401–0.569)	0.224–0.754	0.134 (0.089–0.197)	0.026–0.471	Modest shift in pooled estimates; overall interpretation unchanged

Abbreviations: *CI, confidence interval; PI, prediction interval.*
Note. *Triage Sort arms from Price et al. (2016) and Vassallo et al. (2019), and the SPTT arm from Vassallo et al. (2022), were excluded because these tools are more commonly positioned outside immediate primary scene triage in the literature. This post hoc sensitivity analysis was performed to assess whether inclusion of these arms materially influenced pooled estimates or residual heterogeneity. After exclusion, pooled undertriage increased modestly, whereas pooled overtriage remained essentially unchanged. Wide prediction intervals and substantial residual heterogeneity persisted in both models. Main-analysis values are taken from the primary pooled models reported in the manuscript, and excluded-analysis values are derived from the multilevel pooled models after removal of Triage Sort and SPTT.*

Supplementary Table 15
Additional pediatric-only and overlap-focused sensitivity analyses

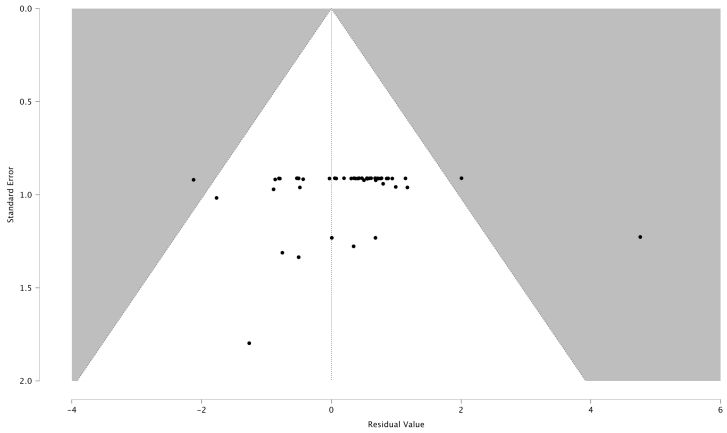
Analysis	Outcome	Estimates (k)	Clusters (n)	Pooled proportion	95% CI	95% PI	Q_e (df)	σ^2	Note
Pediatric-only pooled analysis	Undertriage	21	6	0.383	0.162–0.667	0.030–0.927	1921.48 (20)	1.165	Exploratory age-restricted pooled analysis; pooled effect non-significant [t(4.98) = −1.04, p = .345]; substantial residual heterogeneity remained.
Pediatric-only pooled analysis	Overtriage	21	6	0.161	0.018–0.665	3.998×10^{-4}–0.989	14247.41 (20)	4.938	Exploratory age-restricted pooled analysis; pooled effect non-significant [t(5.00) = −1.82, p = .129]; overtriage estimates were highly imprecise.
Conservative overlap-focused exclusion sensitivity analysis	Undertriage	43	13	0.477	0.374–0.583	0.187–0.784	3288.60 (42)	0.349	Potentially overlapping registry-derived publication sets were excluded; pooled estimate directionally consistent with the main analysis [t(10.13) = −0.47, p = .648].
Conservative overlap-focused exclusion sensitivity analysis	Overtriage	43	13	0.102	0.056–0.178	0.010–0.550	15764.78 (42)	1.101	Potentially overlapping registry-derived publication sets were excluded; pooled estimate directionally consistent with main analysis and statistically significant [t(11.86) = −7.33, p < .001].

Abbreviations: *CI, confidence interval; PI, prediction interval; Q_e, residual heterogeneity statistic; σ^2, between-cluster variance component. Pediatric-only analyses: 6 clusters (Heffernan 2019, Malik 2021 pediatric, Price 2016, Vassallo 2022, Wallis 2006a, Wallis 2006b), 21 estimates. Conservative overlap-focused exclusion: 13 clusters, 43 estimates; potentially overlapping registry-derived publication sets excluded, while preserving the multilevel modelling structure. Cluster-robust variance estimation with small-sample correction (clubSandwich CR2) was applied to all models.*

Supplementary Table 4
Leave-one-cluster-out analyses for pooled undertriage and overtriage

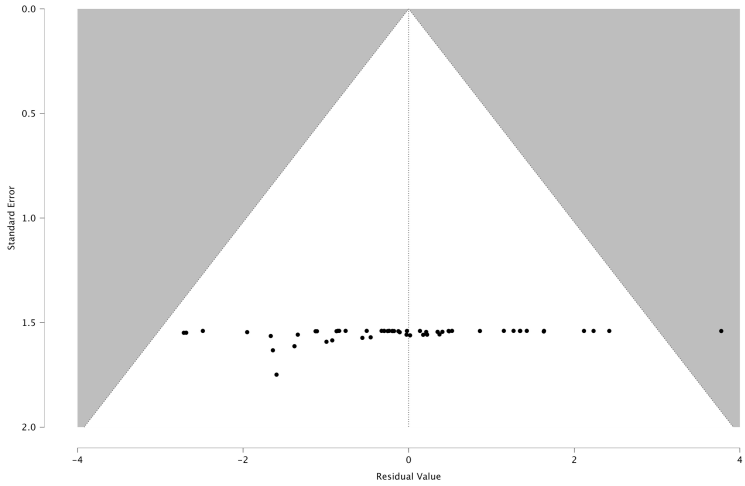
Each row shows the pooled estimate and heterogeneity metrics after the omission of one study from the multilevel model. PI = prediction interval; σ^2 = estimated between-cluster variance component.

Removed study	UT pooled (95% CI)	UT PI	UT σ^2	OT pooled (95% CI)	OT PI	OT σ^2	Interpretation
Bhalla	0.429 (0.314–0.553)	0.095–0.843	0.793	0.135 (0.066–0.255)	0.006–0.812	2.316	minimal change
Challen	0.461 (0.335–0.592)	0.093–0.876	0.923	0.129 (0.063–0.247)	0.005–0.807	2.341	minimal change
Heffernan	0.452 (0.325–0.586)	0.088–0.876	0.948	0.139 (0.069–0.260)	0.006–0.810	2.247	minimal change
Heller	0.458 (0.330–0.592)	0.090–0.878	0.947	0.127 (0.062–0.242)	0.005–0.802	2.323	minimal change
Jerome	0.450 (0.322–0.585)	0.086–0.876	0.960	0.124 (0.061–0.237)	0.005–0.794	2.286	minimal change
Kahn	0.459 (0.337–0.587)	0.096–0.872	0.900	0.130 (0.063–0.249)	0.005–0.809	2.342	minimal change
Lin	0.459 (0.337–0.587)	0.096–0.872	0.900	0.139 (0.069–0.258)	0.006–0.806	2.224	minimal change
Malik_2021	0.448 (0.320–0.583)	0.086–0.875	0.957	0.123 (0.060–0.235)	0.005–0.789	2.262	minimal change
Malik_2021b	0.449 (0.321–0.584)	0.086–0.875	0.958	0.127 (0.062–0.242)	0.005–0.802	2.324	minimal change
Mckee	0.454 (0.327–0.588)	0.089–0.877	0.948	0.131 (0.064–0.251)	0.005–0.810	2.344	minimal change
Peng	0.451 (0.323–0.586)	0.087–0.877	0.960	0.139 (0.069–0.260)	0.006–0.810	2.252	minimal change
Price	0.464 (0.337–0.596)	0.095–0.878	0.919	0.126 (0.062–0.242)	0.005–0.801	2.321	minimal change
Tiyawat	0.466 (0.339–0.597)	0.097–0.877	0.907	0.123 (0.060–0.233)	0.005–0.787	2.247	minimal change
Vassallo 2017a	0.445 (0.318–0.579)	0.086–0.873	0.947	0.129 (0.063–0.247)	0.005–0.808	2.344	minimal change
Vassallo 2019	0.424 (0.315–0.540)	0.105–0.822	0.693	0.148 (0.078–0.265)	0.008–0.786	1.948	variance reduction
Vassallo 2022	0.496 (0.394–0.599)	0.163–0.832	0.534	0.107 (0.062–0.180)	0.010–0.599	1.330	larger shift; variance reduction
Wallis 2006a	0.444 (0.317–0.578)	0.086–0.871	0.940	0.150 (0.080–0.265)	0.009–0.779	1.881	variance reduction
Wallis 2006b	0.444 (0.318–0.578)	0.086–0.871	0.942	0.133 (0.065–0.254)	0.005–0.812	2.334	minimal change



Supplementary Fig. 1 Residual funnel plot for undertriage

Note. Residual funnel plot of study-level residuals from the multilevel meta-analysis model of undertriage. The x-axis represents standardized residual values, and the y-axis represents the standard error. The shaded triangular region indicates the expected distribution of residuals in the absence of small-study effects. The vertical dashed line represents the model-centered residual value (zero). Visual inspection shows a broadly symmetrical distribution of residuals around zero, without clear evidence of asymmetry or small-study effects. Residual funnel plots were constructed using studentized residuals derived from the multilevel random-effects model fitted with restricted maximum likelihood (REML) and cluster-robust variance estimations. This approach accounts for within-study dependency across multiple study arms and provides a more appropriate assessment of small-study effects in multilevel meta-analytic settings.



Supplementary Fig. 2 Residual funnel plot for overtriage

Note. Residual funnel plot of study-level residuals from the multilevel meta-analysis model of overtriage. The x-axis represents the standardized residual values, and the y-axis represents the standard error. The shaded triangular region denotes the expected dispersion of the residuals under symmetry assumptions. The vertical dashed line corresponds to a zero residual value. The observed distribution appeared relatively symmetrical, with no consistent pattern of asymmetry suggestive of small-study effects or publication bias. Residual-based funnel plots were preferred over conventional effect-size funnel plots because of the multilevel structure of the data, in which multiple correlated estimates were nested within studies. The use of cluster-robust variance estimation (RVE) mitigates bias due to dependency and provides a more reliable visual diagnosis of potential small-study effects.

Supplementary Table 5
Summary of univariable moderator analyses for undertriage and overtriage

*Each row reports the omnibus moderator test from a separate univariable multilevel meta-regression model. Fm = cluster-robust F-statistic with Satterthwaite-adjusted degrees of freedom. σ^2 = between-cluster variance component. UT = undertriage; OT = overtriage. * indicates p < .05.*

Moderator	Outcome	Fm	df1	df2	p	σ^2	Interpretation
Algorithm family	UT	0.05	1	1.34	.858	0.900	Not significant
Algorithm family	OT	560.07	1	1.28	.012*	2.419	Significant
Population age	UT	1.18	2	8.58	.353	0.864	Not significant
Population age	OT	0.07	2	9.57	.933	2.464	Not significant
Study design	UT	0.84	1	4.90	.403	0.937	Not significant
Study design	OT	4.24	1	4.89	.096	1.955	Borderline, not significant
Data context	UT	0.70	3	1.30	.661	1.033	Not significant
Data context	OT	0.34	3	1.36	.808	2.415	Not significant
Setting	UT	9.11	3	2.55	.068	1.005	Borderline, not significant
Setting	OT	0.28	3	2.57	.836	2.499	Not significant
Provider type	UT	0.02	2	5.19	.983	1.008	Not significant
Provider type	OT	1.55	2	7.09	.276	2.075	Not significant
Civilian vs Military	UT	0.004	1	1.15	.958	—	Not significant; sparse
Civilian vs Military	OT	2.38×10^{-5}	1	1.12	.997	—	Not significant; sparse

Note. *The algorithm family was modeled as a numeric predictor in the univariable moderator analyses above. The categorical algorithm-family estimated marginal means for overtriage are reported separately in Supplementary Table 7. The collapsed reference-standard moderator is reported separately in Supplementary Table 6 using the revised 2-category classification (intervention/acuity-based vs. non-intervention-based). Multivariable models incorporating the same revised reference standard variables are presented in Supplementary Table 12. Civilian vs. Military results are reproduced from Supplementary Table 10 for completeness. Provider type was examined as a moderator, but as noted in Supplementary Appendix S3, the reporting was insufficient to support this analysis reliably.*

Supplementary Table 6
Revised univariable moderator analysis of collapsed reference-standard family

Meta-regression coefficients (logit scale) and category-level estimates from univariable multilevel meta-regression models with REML estimation and cluster-robust variance (CR2, Satterthwaite-adjusted df). Reference standards were collapsed into two prespecified categories: Intervention/acuity-based and Non-intervention-based.

A. Omnibus moderator tests

Outcome	Reference-standard comparison	Fm	df1	df2	p	σ^2	Interpretation
Undertriage	Intervention/acuity-based vs Non-intervention-based	0.05	1	11.34	.830	0.952	Not significant
Overtriage	Intervention/acuity-based vs Non-intervention-based	0.98	1	12.64	.342	2.252	Not significant

B. Meta-regression coefficients

Outcome	Variable	Estimate	SE	95% CI	t	df	p
Undertriage	Intercept (Intervention/acuity-based)	−0.153	0.332	−0.899 to 0.592	−0.462	9.478	.655
Undertriage	Reference-standard contrast: Non-intervention-based vs Intervention/acuity-based	−0.105	0.477	−1.151 to 0.941	−0.220	11.344	.830
Overtriage	Intercept (Intervention/acuity-based)	−1.659	0.533	−2.846 to −0.472	−3.115	9.999	.011
Overtriage	Reference-standard contrast: Non-intervention-based vs Intervention/acuity-based	−0.621	0.629	−1.983 to 0.741	−0.988	12.641	.342

C. Category-specific back-transformed estimates

Outcome	Reference-standard family	Pooled proportion	95% CI	95% PI
Undertriage	Intervention/acuity-based	0.454	0.325–0.589	0.087–0.878
Undertriage	Non-intervention-based	0.428*	—	—
Overtriage	Intervention/acuity-based	0.160	0.055–0.384	0.005–0.869
Overtriage	Non-intervention-based	0.093	0.043–0.188	0.002–0.816

Abbreviations*: CI, confidence interval; PI, prediction interval.*
Note. *In these two-level models, Intervention/acuity-based was the reference category; therefore, the coefficient for Non-intervention-based represents the single contrast against this reference. For overtriage, category-specific estimated marginal means were directly provided by the software. For undertriage, the intervention/acuity-based value corresponds to the pooled effect reported by the moderator model, whereas the non-intervention-based value is shown descriptively as the back-transformed estimate implied by the model coefficient; because the contrast was non-significant, category-specific confidence intervals for that derived value were not emphasized in the interpretation. These results indicate that the dichotomized reference-standard classification did not explain the statistically detectable proportion of between-study variability.*

Supplementary Table 7
Overtriage algorithm-family marginal means and model coefficients

Family specific estimated marginal means were obtained from the overtriage multilevel moderator model. Model coefficients are shown where provided by the software output; omitted blank cells indicate levels for which a stable coefficient display was not available in the extracted output.

Algorithm family	Marginal mean (95% CI)	95% PI	Model coefficient (95% CI)	p
MPTT family	0.412 (0.283–0.554)	0.055–0.893	1.495 (0.021–2.968)	0.049
PRIOR family	0.361 (0.227–0.522)	0.044–0.874	1.282 (0.927–1.637)	0.014
Sieve family	0.227 (0.123–0.380)	0.023–0.786	0.625 (-1.688–2.938)	0.259
META family	0.140 (0.082–0.228)	0.013–0.662	0.034 (-1.563–1.631)	0.909
ASAV family	0.136 (0.073–0.238)	0.013–0.659		
START family	0.106 (0.060–0.181)	0.010–0.590	-0.277 (-1.821–1.266)	0.302
SALT family	0.091 (0.052–0.154)	0.008–0.546	-0.452 (-2.067–1.164)	0.264
Pediatric tape-based family	0.083 (0.017–0.324)	0.005–0.635	-0.556 (-8.348–7.235)	0.597
CareFlight family	0.077 (0.044–0.131)	0.007–0.500		
RAMP family	0.071 (0.041–0.120)	0.006–0.477	-0.723 (-1.375–-0.072)	0.041
Sort family	0.059 (0.000–1.000)	0.000–1.000	-0.924 (-12.848–11.000)	0.637
FTS family	0.021 (0.006–0.064)	0.001–0.237	-2.015 (-5.980–1.950)	0.098

Note. *The overall overtriage algorithm family moderator effect was significant (p=0.012). The estimated marginal means were highest for the MPTT and PRIOR families and lowest for the FTS and RAMP families. The contrasts of the estimated marginal means in the source output were Bonferroni-adjusted.*

Supplementary Table 8
Univariable meta-regression coefficients for undertriage

*Coefficients are reported on the logit scale from separate univariable multilevel meta-regression models with REML estimation and cluster-robust variance (CR2 with Satterthwaite-adjusted df). Each section header shows the omnibus moderator test and residual between-cluster variance (σ^2). The intercept corresponds to the reference category, as shown in parentheses. * indicates $p < .05$.*

Variable	Estimate	SE	95% CI	t	df	p
Population age (Fm(2, 8.58) = 1.18, p = .353, σ^2 = 0.864)						
Intercept (Adult)	0.268	0.414	−0.771 to 1.307	0.647	5.46	.544
Mixed	−0.723	0.453	−1.782 to 0.335	−1.596	7.46	.152
Pediatric	−0.745	0.617	−2.113 to 0.623	−1.207	10.38	.254
Study design (Fm(1, 4.90) = 0.84, p = .403, σ^2 = 0.937)						
Intercept (Prospective)	0.067	0.200	−0.572 to 0.706	0.335	2.98	.760
Retrospective	−0.335	0.366	−1.280 to 0.611	−0.915	4.90	.403
Data context (Fm(3, 1.30) = 0.70, p = .661, σ^2 = 1.033)						
Intercept (Hospital-based)	−0.162	0.232	−0.742 to 0.419	−0.696	5.51	.514
Prehospital HEMS	−0.393	5.227	−31.031 to 30.245	−0.075	1.53	.949
Real incident	−0.901	0.414	−3.712 to 1.910	−2.175	1.38	.216
Registry-based	0.105	0.510	−1.009 to 1.219	0.206	11.71	.841
Setting (Fm(3, 2.55) = 9.11, p = .068, σ^2 = 1.005)						
Intercept (Hospital-based)	−0.040	0.240	−0.588 to 0.507	−0.167	8.46	.871
Other real-world	−1.569	0.240	−2.117 to −1.022	−6.548	8.46	<.001*
Prehospital HEMS	−0.514	5.140	−36.182 to 35.153	−0.100	1.36	.933
Registry	−0.244	0.599	−1.561 to 1.072	−0.408	11.18	.691
Provider type (Fm(2, 5.19) = 0.02, p = .983, σ^2 = 1.008)						
Intercept (Clinical)	−0.115	0.476	−2.037 to 1.807	−0.242	2.14	.830
Computer	−0.116	0.598	−1.883 to 1.650	−0.194	3.47	.857
Research abstraction	−0.047	0.662	−1.798 to 1.704	−0.071	4.56	.947

Supplementary Table 9
Univariable meta-regression coefficients for overtriage

*Coefficients are reported on the logit scale from separate univariable multilevel meta-regression models with REML estimation and cluster-robust variance (CR2 with Satterthwaite-adjusted df). Each section header shows the omnibus moderator test and residual between-cluster variance (σ^2). The intercept corresponds to the reference category, as shown in parentheses. The algorithm family coefficients for overtriage are reported in Supplementary Table 7 (estimated marginal means) and the corresponding meta-regression output. * indicates $p < .05$.*

Variable	Estimate	SE	95% CI	t	df	p
Population age (Fm(2, 9.57) = 0.07, p = .933, σ^2 = 2.464)						
Intercept (Adult)	−2.036	0.432	−3.094 to −0.978	−4.710	6.00	.003*
Mixed	0.031	0.631	−1.407 to 1.468	0.048	8.60	.963
Pediatric	0.389	1.006	−1.834 to 2.611	0.386	10.70	.707
Study design (Fm(1, 4.89) = 4.24, p = .096, σ^2 = 1.955)						
Intercept (Prospective)	−2.970	0.548	−4.713 to −1.226	−5.420	3.00	.012*
Retrospective	1.389	0.675	−0.357 to 3.135	2.059	4.89	.096
Data context (Fm(3, 1.36) = 0.34, p = .808, σ^2 = 2.415)						
Intercept (Hospital-based)	−2.487	0.479	−3.662 to −1.313	−5.190	5.96	.002*
Prehospital HEMS	1.188	22.984	−167.612 to 169.988	0.052	1.32	.965
Real incident	0.753	0.487	−1.835 to 3.341	1.548	1.65	.287
Registry-based	0.976	0.814	−0.788 to 2.739	1.199	12.62	.253
Setting (Fm(3, 2.57) = 0.28, p = .836, σ^2 = 2.499)						
Intercept (Hospital-based)	−2.243	0.367	−3.073 to −1.412	−6.115	8.96	<.001*
Other real-world	0.423	0.367	−0.407 to 1.253	1.153	8.96	.279
Prehospital HEMS	0.944	23.360	−195.901 to 197.788	0.040	1.22	.973
Registry	0.793	0.945	−1.292 to 2.878	0.839	10.80	.419
Provider type (Fm(2, 7.09) = 1.55, p = .276, σ^2 = 2.075)						
Intercept (Clinical)	−2.979	0.626	−4.987 to −0.971	−4.761	2.96	.018*
Computer	1.576	0.844	−0.531 to 3.683	1.867	5.55	.115
Research abstraction	1.007	0.748	−0.796 to 2.809	1.346	6.41	.224

Supplementary Table 10
Additional univariable moderator analyses not emphasized in the primary interpretation

Outcome	Moderator	F	df 1	df2	p	Brief interpretation
Undertriage	Civilian_Military_Simple	0.004	1	1.147	.958	No significant moderation; estimates were highly imprecise because military-context data were sparse.
Overtriage	Civilian_Military_Simple	2.381×10^{-5}	1	1.124	.997	No significant moderation; estimates were highly imprecise because military-context data were sparse.

Supplementary Table 11
Conceptual synthesis of multilevel meta-regression findings for undertriage and overtriage in MCI triage systems

Domain	Undertriage (UT)	Overtriage (OT)	Interpretation
Pooled direction	High pooled UT persisted across models	Lower pooled OT compared to UT	MCI triage systems appear less reliable in identifying truly critical patients than in limiting unnecessary high-priority classification
Reference standard	Not a statistically significant moderator (p = .830)	Not a statistically significant moderator (p = .342)	Outcome definition heterogeneity remains an interpretive challenge, but the dichotomized reference-standard classification did not explain a statistically detectable proportion of variability
Algorithm family	No robust independent contribution after multivariable adjustment; estimation unstable (negative residual df)	Some signals were observed, but they were inconsistent overall and unstable in multivariable modeling	Differences between triage algorithms are not consistently reproducible across studies
Data context / setting	No significant independent effect	No significant independent effect	Operational setting alone does not adequately explain heterogeneity
Residual heterogeneity	Remained substantial after adjustment	Remained substantial after adjustment	Indicates presence of unmeasured structural or contextual factors
Overall inference	Variability reflects complex interplay of methodological, definitional, and contextual factors	No single moderator consistently explains OT variability	Standardization of reference standards remains important on conceptual grounds, even though formal moderator tests were non-significant

Abbreviations: *UT, undertriage; OT, overtriage; ISS, Injury Severity Score; MCI, mass casualty incident.*

This table summarizes the findings from the multilevel multivariate meta-regression models incorporating the reference standard type, data context, and algorithm family as moderators. Effect estimates were derived from logit-transformed proportions and back-transformed for interpretation. Models were fitted using restricted maximum likelihood (REML) with cluster-robust variance estimation to account for within-study dependencies. No single moderator, including the reference standard type, emerged as a consistently dominant explanatory factor. Substantial residual heterogeneity persisted despite the adjustments, indicating the presence of additional unmeasured sources of variability.

Supplementary Table 12
Multivariable meta-regression coefficients for undertriage and overtriage

This table reports the coefficient-level outputs from the multivariable multilevel meta-regression models. Both the undertriage and overtriage models included a collapsed reference-standard family (2-category), data context, and algorithm family. The reference-standard type did not reach statistical significance in either model, and the algorithm-family estimation was unstable for several contrasts because of sparse subgroup representation and problematic residual degrees of freedom. The coefficients are reported on the logit scale.

Undertriage model

Variable	Estimate	SE	95% CI	p
Intercept	−0.933	0.318	−1.733 to −0.133	.030
Reference-standard contrast: Non-intervention-based vs Intervention/acuity-based	−0.058	0.674	−1.683 to 1.568	.934
Data context (Prehospital HEMS)	−0.243	2.772	−8.384 to 7.898	.935
Data context (Real incident)	−0.736	0.548	−3.786 to 2.314	.339
Data context (Registry-based)	0.047	0.555	−1.193 to 1.286	.935
Algorithm family — CareFlight	0.890	0.081	0.550 to 1.230	.008
Algorithm family — FTS	2.063	3.796×10^{-6}	2.063 to 2.063	<.001
Algorithm family — META	0.724	0.116	−0.067 to 1.514	.056
Algorithm family — MPTT	0.530	0.348	−0.915 to 1.974	.263
Algorithm family — PRIOR	−0.879	4.519×10^{-6}	−0.879 to −0.879	<.001
Algorithm family — Pediatric tape-based	0.806	0.100	−0.278 to 1.891	.069
Algorithm family — RAMP	1.072	0.066	0.811 to 1.333	.002
Algorithm family — SALT	1.164	0.116	0.236 to 2.093	.038
Algorithm family — START	0.596	0.131	−0.266 to 1.458	.083
Algorithm family — Sieve	0.505	0.144	−0.271 to 1.280	.097
Algorithm family — Sort	1.309	1.453	−5.839 to 8.458	.474

Overtriage model

Variable	Estimate	SE	95% CI	p
Intercept	−2.261	0.523	−3.551 to −0.971	.005
Reference-standard contrast: Non-intervention-based vs Intervention/acuity-based	−0.362	0.639	−1.862 to 1.139	.589
Data context (Prehospital HEMS)	1.178	11.784	−58.111 to 60.466	.931
Data context (Real incident)	0.894	0.470	−1.499 to 3.287	.219
Data context (Registry-based)	0.904	0.592	−0.411 to 2.219	.157

Variable	Estimate	SE	95% CI	p
Algorithm family — CareFlight	−0.633	0.051	−0.926 to −0.340	.015
Algorithm family — FTS	−2.015	1.158×10^{-6}	−2.015 to −2.015	<.001
Algorithm family — META	0.037	0.214	−2.180 to 2.253	.890
Algorithm family — MPTT	1.496	0.163	0.669 to 2.324	.019
Algorithm family — PRIOR	1.282	1.244×10^{-6}	1.282 to 1.282	<.001
Algorithm family — Pediatric tape-based	−0.552	0.761	−9.870 to 8.766	.599
Algorithm family — RAMP	−0.722	0.053	−1.124 to −0.320	.024
Algorithm family — SALT	−0.449	0.214	−2.694 to 1.796	.266
Algorithm family — START	−0.276	0.055	−0.495 to −0.056	.032
Algorithm family — Sieve	0.627	0.304	−0.857 to 2.110	.192
Algorithm family — Sort	−1.238	1.863	−10.408 to 7.933	.583

Abbreviations: *HEMS, helicopter emergency medical services; LSI, life-saving intervention.*
Note. *The models were fitted using multilevel mixed-effects meta-regression with REML estimation and cluster-level dependence handling. The reference standard was modeled as a 2-category variable, with Intervention/acuity-based used as the reference level; accordingly, only the non-intervention-based vs. Intervention/acuity-based contrast coefficient is displayed. Algorithm family coefficients should be interpreted cautiously because several contrasts were associated with sparse subgroup representation and unstable residual degrees of freedom.*

Supplementary Note 1. Exploratory injury-mechanism meta-regression

Exploratory injury mechanism meta-regression was attempted for undertriage and overtriage; however, the estimation was unstable because of sparse and uneven subgroup representation, software termination/errors, and extremely imprecise coefficient estimates. Therefore, these findings were not included in the primary interpretation.

Supplementary Table 16

Study-specific reference-standard operationalization and binary harmonization used in the meta-analysis

Because source studies used different acuity constructs and triage-category structures, study-specific reference-standard operationalization and binary harmonization rules were pre-specified during data extraction.

Study	Original reference-standard label	Exact operational definition used for extraction	Collapsed group	Index-test positivity threshold used	Category collapse rule used for reconstruction
Bhalla 2015	Hospital outcome/intervention timing	Outcome-based 4-category mapping: minor/green = discharge without major intervention; delayed/yellow = intervention >12 h; immediate/red = intervention <12 h; dead/expectant/black = death within 48 h or CPC 4–5. The study reports a 4-level agreement: binary DTA collapse reconstructed from the extracted counts.	Non-intervention-based	Study reports 4-level agreement; binary DTA collapse reconstructed from extracted counts	Not directly binary in source article
Heller 2019	Weighted expert physician reference triage	Nineteen emergency physicians judged HEMS missions in triplicate using a weighted consensus reference category.	Non-intervention-based	T1 considered positive	T1 vs non-T1
Jerome 2023	Urgent lifesaving interventions (LSI)	The immediate category was defined as the need for specific urgent lifesaving interventions translated from Lerner's consensus definitions using Alberta Trauma Registry data.	Intervention/acuity-based	Immediate / Priority 1 positive	Immediate vs delayed/expectant/dead-other
Heffernan 2019	Expert-panel-derived criterion standard categories	Hospital record review mapped to criterion-standard disaster triage categories.	Non-intervention-based	Immediate / critical positive against all other categories	Immediate or higher acuity vs delayed/minimal
Challen 2013	Modified Baxt/Upeniek criticality criteria	Criticality was defined from ED/hospital notes using the modified Baxt and Upeniek criteria.	Intervention/acuity-based	Working extraction used top-two-category positive cutoff	P1 and P2 vs P3/P4
Price 2016	ISS >15	Major trauma/priority one target condition defined by Injury Severity Score threshold >15 in the pediatric TARN cohort.	Non-intervention-based	Priority One / major trauma positive test	ISS >15 vs ISS ≤15
Malik 2021 (pediatric)	Paediatric LSI	Pediatric TARN cohort; Priority One defined as the receipt of at least one life-saving intervention.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One
Vassallo 2022	Paediatric LSI	Pediatric TARN cohort; Priority One, defined as at least one life-saving intervention.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One
Peng 2021	Death/ICU/ISS>15 composite	The outcome reference standard was based on in-hospital deaths among earthquake casualties.	Non-intervention-based	Death / non-survival positive test	Died vs survived
Tiyawat 2024	Lerner consensus criteria	Adult trauma cohort; critical patients defined by life-saving intervention requirements.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One

Study	Original reference-standard label	Exact operational definition used for extraction	Collapsed group	Index-test positivity threshold used	Category collapse rule used for reconstruction
McKee 2020	LSI	Expert-derived life-saving intervention reference standard applied to the ED-based MCI simulation cohort.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One
Lin 2022	Expert outcome review	Expert outcome review of derailment casualties: immediate/critical vs. non-critical.	Non-intervention-based	Immediate / critical positive	Immediate (critical) vs non-critical
Vassallo 2019	LSI	Adult TARN cohort; Priority One, defined by life-saving intervention criteria.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One
Vassallo 2017	LSI	Military JTTR cohort; Priority One defined by life-saving intervention criteria used for MPTT derivation/validation.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One
Kahn 2009	LSI / Baxt criteria	Critical patient defined by Baxt-based outcome criteria/need for life-saving intervention after a real train crash.	Intervention/acuity-based	Immediate / critical positive	Immediate (critical) vs non-critical
Malik 2021 (adult, 16–64 y)	LSI	Adult TARN cohort; Priority One defined by the receipt of at least one life-saving intervention.	Intervention/acuity-based	Priority One / immediate positive test	Immediate / Priority One vs not Immediate / Priority One
Wallis 2006a / 2006b	Garner criteria	Pediatric ED cohort evaluated against the Garner-based urgent intervention criteria.	Intervention/acuity-based	Priority One / urgent intervention positive test	Binary reconstruction based on extracted 2×2 mapping used in the review

Abbreviations: *CPC, Cerebral Performance Category; ED, emergency department; HEMS, helicopter emergency medical services; ISS, Injury Severity Score; JTTR, Joint Theatre Trauma Registry; LSI, life-saving intervention; TARN, Trauma Audit and Research Network. Note. This table summarizes the source-level reference-standard wording and prespecified binary harmonization rules recorded during data extraction. For moderator analyses, source reference standards were collapsed into two categories: intervention/acuity-based (including LSI, pediatric LSI, Lerner consensus, and Garner criteria) and non-intervention-based (including ISS>15, expert-panel, outcome-based, mortality, and hospital-outcome-based frameworks). In studies with different category labels across triage systems, positivity was harmonized to a binary high-priority threshold using the recorded positivity threshold and category collapse rule fields from the master extraction sheet. * Wallis 2006a used Garner criteria operationalized analogously to urgent intervention criteria for reconstruction; Wallis 2006b required a modified collapse rule. Collapsed group colour coding: Intervention/acuity-based | Non-intervention-base*

Supplementary Table 19

Source-level reference-standard frameworks represented in the included evidence base

Source-level reference standard	Collapsed family used in moderator analyses	Studies contributing	Arms contributing	Conceptual basis	Note
LSI	Intervention/acuity-based	4	12	Urgent life-saving intervention requirement	It is used in adult and mixed-population studies as an intervention-based indicator of critical priority.
Urgent LSI	Intervention/acuity-based	1	7	Urgent life-saving intervention requirement	Reported by Jerome et al. (2023) and treated as intervention/acuity-based.
Pediatric LSI	Intervention/acuity-based	2	12	Pediatric life-saving intervention requirement	It is used in pediatric registry-based studies and is treated as intervention/acuity-based.
Lerner consensus criteria	Intervention/acuity-based	1	2	Consensus-based acuity criteria	Used in Tiyawat et al. (2024); classified as intervention/acuity-based in the revised collapsed scheme.
Modified Baxt/Upeniek criticality criteria	Intervention/acuity-based	1	3	Criticality criteria based on urgent priority constructs	Used in Challen et al. (2013); grouped as intervention/acuity-based in the final collapsed classification.
LSI / Baxt criteria	Intervention/acuity-based	1	1	Hybrid intervention/criticality framework	Used in Kahn et al. (2009) and retained within the intervention/acuity-based family.
Garner criteria	Intervention/acuity-based	2 analytically distinct reports	4	Acuity criteria used in pediatric trauma triage evaluation	Used in Wallis 2006a/2006b and treated as intervention/acuity-based in the revised collapsed classification.
ISS >15	Non-intervention-based	1	1	Anatomical injury severity threshold	Used in Price et al. (2016) and treated as non-intervention-based.
Death/ICU/ISS>15 composite	Non-intervention-based	1	2	Composite outcome/severity framework	Used in Peng et al. (2021) and treated as non-intervention-based.
Hospital outcome / intervention timing	Non-intervention-based	1	2	Outcome/timing-based proxy for triage acuity	Used in Bhalla et al. (2015) and treated as non-intervention-based.
Expert-panel-derived criterion standard categories	Non-intervention-based	1	4	Expert-derived categorical acuity assignment	Used in Heffernan et al. (2019) and treated as non-intervention-based.
Weighted expert physician reference triage	Non-intervention-based	1	7	Expert-derived weighted triage reference	Used in Heller et al. (2019) and treated as non-intervention-based.
Expert outcome review	Non-intervention-based	Non-intervention-based	1	1	Expert review of outcomes

Abbreviations: ICU, intensive care unit; ISS, Injury Severity Score; LSI, life-saving intervention.
Note. This table provides a descriptive summary of the source-level reference standard frameworks represented in the included evidence base. It is intended to complement, not replace, the revised collapsed 2-category reference-standard classification used in the moderator analyses. As described in the manuscript, formal moderator models used the broader categories of intervention/acuity-based versus non-intervention-based reference standards, whereas this table displays the underlying source-level operational diversity that contributed to these broader groupings. The exact study-arm-level source labels and extracted 2×2 reference standards are reported in Supplementary Table 1, and positivity thresholds/category-collapse rules are summarized in Supplementary Table 16.

Supplementary Appendix S3
Deviations from the Registered PROSPERO Protocol

Minor deviations from the registered PROSPERO protocol occurred during the completion of the review, which are detailed below.

1. Expansion of information sources

The Web of Science Core Collection was added as an additional bibliographic database to improve the search coverage beyond the originally prespecified sources. ClinicalTrials.gov was also searched as a trial registry source to identify potentially relevant registered studies beyond the originally prespecified bibliographic databases.

2. Handling of multi-arm dependence

In the final analyses, dependency arising from multiple algorithm-specific estimates contributed within the same study was handled using multilevel modeling with study-level clustering rather than by systematic exclusion of multi-arm studies.

3. Assessment of small-study effects

Residual funnel plots were preferred over conventional funnel plot asymmetry testing because the primary synthesis incorporated dependent effect sizes within a multilevel framework, limiting the interpretability of standard small-study-effect methods.

4. Provider-type analysis not performed

The PROSPERO protocol prespecified an analysis of undertriage and overtriage according to the provider type (paramedics, physicians, nurses, or computer-applied algorithms). However, the included studies reported provider information inconsistently or not at all, preventing reliable categorization of the data. Accordingly, this prespecified analysis was not conducted.

5. Expansion of the author team after registration

The PROSPERO protocol listed only two authors (ÖA and DK), reflecting the composition of the review team at the time of registration of the protocol. As the review progressed and the analytical and reporting workflows became more complex, three additional authors (HI, RKE, and OG) were invited to contribute to the statistical analysis, data verification, and senior methodological supervision. Their specific roles are detailed in the Author' Contributions section, and all authors have reviewed and approved the final manuscript.

6. Geriatric subgroup not included in the primary pooled analysis

The PROSPERO protocol listed geriatric populations as eligible. However, the elderly subgroup (≥65 years, n = 100,403) reported separately within Malik et al. (2021) was not included in the primary pooled analysis. That study analysed adults aged 16–64 years and elderly patients aged ≥65 years as separate age-stratified groups, whereas comparable geriatric-specific strata were not available across most of the remaining included studies. Inclusion of the elderly subgroup would therefore have introduced a disproportionately large age-specific contribution from a single study into the main synthesis. The primary pooled analysis was therefore based on the more consistently represented age groups across the evidence base, and this decision is acknowledged as a post hoc deviation.

7. Age-group moderator analysis modified

Although age group was prespecified as a moderator variable, a formal pediatric-versus-adult-versus-mixed subgroup analysis was not performed as a standalone moderator model because the adult-only sensitivity analysis already addressed the most clinically relevant age-related comparison, and the remaining subgroups (pediatric-only and mixed) had limited and unbalanced cluster representations.

8. Language restriction during review conduct

The PROSPERO registration stated that no language restrictions would apply. However, during the review, eligibility was restricted to English-language publications for feasibility reasons, as the review team did not have the capacity for reliable screening, extraction, and quality assessment in other languages. This deviation may have introduced a language bias by excluding potentially relevant non-English studies.

9. Data extraction workflow differed from the working protocol

Minor discrepancies existed across the PROSPERO record, the standalone working protocol, and the final review. In particular, while the working protocol initially described independent dual data extraction, the review used a single-extractor-plus-verification workflow, whereby one reviewer (OA) performed the primary extraction and a second reviewer (RKE) checked the extracted dataset in full for accuracy, completeness, and consistency. Accordingly, the manuscript reports the methods as actually conducted and should be regarded as the definitive account of the review.

These deviations were transparently documented and were considered unlikely to materially change the overall interpretation of the review's findings.

Supplementary Appendix S4
Expanded Statistical Rationale

Primary statistical analyses were conducted using a multilevel random-effects proportion meta-analysis on the logit scale. This framework was selected to address two structural features of the evidence base: substantial methodological and clinical heterogeneity across studies, and within-study dependence arising from multiple algorithm-specific estimates contributed by the same study. Under these conditions, multilevel modeling was considered preferable to simpler conventional approaches, because it allowed eligible multi-arm information to be retained without treating correlated estimates as fully independent and without requiring arbitrary exclusion of study arms. The resulting framework therefore provided a transparent and structurally appropriate primary synthesis strategy for pooled undertriage and overtriage proportions.

Random-effects modeling was preferred a priori because substantial between-study variability was expected in relation to the study design, reference standard framework, population mix, and operational context. The between-study variance was estimated using the restricted maximum likelihood (REML). For each pooled estimate, the review reports the pooled proportion, 95% confidence interval, 95% prediction interval, residual heterogeneity statistics, and variance component (σ^2), as these metrics were considered more informative than conventional heterogeneity indices alone within the clustered multilevel structure.

Because multiple algorithm-specific estimates were often contributed by the same study, dependence was handled using multilevel mixed-effects models with study-level clustering rather than by systematic exclusion of multi-arm studies. This approach was chosen to preserve the maximum amount of eligible algorithm-specific information while reducing arbitrariness and avoiding the information loss that would have resulted from selecting a single arm per study. In this review, preserving multi-arm structure was considered especially important because direct comparisons between algorithms within the same study population constituted a meaningful part of the available evidence base.

To improve statistical inference in the presence of a relatively small number of independent study clusters, cluster-robust variance estimation with a small-sample correction was applied. Specifically, robust inference was based on the clubSandwich CR2 estimator with Satterthwaite-adjusted degrees of freedom, as implemented in JASP using the metafor computational engine. This approach was considered appropriate because it provides bias-reduced standard errors and more reliable hypothesis testing in multilevel meta-analytic settings with limited cluster numbers.

Bivariate diagnostic test accuracy (DTA) models were not selected as the primary analytical framework for this review. Although such models can jointly estimate sensitivity and specificity while accounting for their correlation, the present evidence base posed two important challenges. First, many included studies contributed multiple correlated algorithm-specific estimates within the same study, such that a conventional bivariate model would either require correlated within-study arms to be treated as independent observations or require the multi-arm structure to be discarded. Second, there was substantial heterogeneity in the reference standards used to define patient acuity. Because this heterogeneity introduced conceptual inconsistency in the target construct rather than providing a stable benchmarking framework, the core bivariate assumption of a sufficiently common underlying reference construct was not adequately satisfied.

For these reasons, separate multilevel proportion meta-analyses with moderator exploration were considered a more transparent and structurally appropriate primary approach to the review question. This strategy was also more closely aligned with the principal objective of the review, which was to estimate pooled absolute undertriage and overtriage proportions across prehospital MCI triage systems rather than to derive relative comparative accuracy metrics between a fixed pair of index tests. In other words, the central methodological challenge in this review was not simply paired test-accuracy synthesis, but synthesis under correlated multi-arm structure together with heterogeneous acuity definitions.

Alternative comparative strategies, including within-study paired comparisons between algorithm pairs sharing the same study population and reference standards, were considered conceptually relevant. However, these approaches were not adopted as the primary synthesis method because not all included studies contributed directly comparable multi-arm data, available algorithm pairings varied substantially across studies, and the evidence base was not sufficiently homogeneous to support a unified comparative framework without substantial loss of data or

interpretability. Such approaches may become more informative in future reviews if a larger and more methodologically consistent body of head-to-head evidence becomes available.

In the updated moderator analyses, the revised collapsed 2-category reference-standard variable did not show statistically significant moderation for either undertriage or overtriage. This does not imply that reference-standard choice is clinically unimportant; rather, it suggests that a broad dichotomized grouping was insufficient to explain a statistically detectable proportion of the substantial between-study heterogeneity in the present 18-cluster evidence base. Accordingly, reference-standard heterogeneity is interpreted primarily as a challenge to cross-study comparability rather than as a confirmed dominant statistical driver of variability in the updated models.