

Disentangling Algorithmic Approximation from Clinical Validity in AI Emergency Triage

Édouard Lansiaux*, Hayfa Zgaya-Biau[†], and Mehdi Ammi[‡]

*CHU Lille, Emergency Dept. & STaR-AI; Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISAL, France

Email: edouard.lansiaux@chu-lille.fr

[†]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISAL, France

Email: hayfa.zgaya@univ-lille.fr

[‡]LIASD, University Paris 8, Saint-Denis, France

Email: mehdi.ammi@ai.univ-paris8.fr

Abstract—Emergency-triage classifiers are often scored against labels reconstructed from a deterministic scale. Because an expressive model can recover that mapping, high predictive metrics may measure algorithmic approximation rather than agreement with physicians. We present EIMLIA-TEU, an evaluation framework—not a new triage algorithm—that separates the two. Regime R1 trains and tests on FRENCH-reconstructed labels. Regime R2 keeps those labels as the supervised target but scores models against an externally moderated five-physician consensus. Three published architectures were retrained on 340,536 anonymized adult emergency records (2018–2023 calibration cohort) and tested on 68,108 cases under R1, where weighted kappa reached 0.996, 0.939 and 0.895. Expert moderation is complete for 400 of 3,000 planned cases: empirical R2 kappa is 0.81, 0.74 and 0.69, with bootstrap intervals that are wider than the previously reported projections. URGENTIPARSE’s $n=400$ interval [0.76, 0.85] crosses the $\kappa_w \geq 0.80$ deployment bar, so that threshold is not claimed. Residual errors injected into a hybrid DES–MAS engine use class-conditional confusion matrices, not a global kappa, because two models with similar concordance can differ sharply in Level 1–2 undertriage (9.6% vs. 27.0% on the completed subset). A CHEERS 2022 probabilistic analysis is reported with explicit perimeters: the 480% three-year ROI is a mono-site hospital figure ($\approx 115,000$ visits/year); the conservative communicated floor is 210%. National 50–80 M€/year figures are a derived scenario (undertriage adverse-event channel only, R2, 8–13% capture). R2 remains incomplete until $n=3,000$.

Index Terms—emergency triage, clinical decision support, evaluation methodology, discrete-event simulation, multi-agent systems, health economics

I. INTRODUCTION

Emergency departments (EDs) are the main acute-care entry point; France records over 20 million ED visits per year [1]. Initial triage, standardized through the FRENCH scale [2], [3], shapes the whole pathway. Inter-operator disagreement (25–40% [4]–[6]), undertriage of serious cases (10–15% [7], [8]) and overtriage of minor cases (20–30% [9]) all worsen overcrowding [10], [11].

AI triage models often look strong on retrospective benchmarks [12]–[14], but two evaluation artefacts remain coupled. First, predictive scores are reported without asking how a recommendation moves through an ED as a queue. Second, labels are frequently *reconstructed* from a deterministic scale

and then used as ground truth. A sufficiently expressive classifier recovers that mapping; the metric then measures approximation of the algorithm, not clinical judgment. If reconstructed labels and experts disagree, any R1-perfect model will automatically disagree with experts—the clinical assessment differs because the *reference* differs, not because a new medical algorithm was invented.

This paper’s contribution is methodological. We do not propose a new triage architecture. Three models developed and first validated elsewhere [15], [16] are used only as instruments. EIMLIA-TEU supplies: (i) a dual-regime protocol (R1 algorithmic approximation vs. R2 clinical validity); (ii) a hybrid discrete-event/multi-agent (DES–MAS) engine that injects *class-specific* residual errors; (iii) a CHEERS 2022 probabilistic sensitivity analysis (PSA) with labelled perimeters. Wireless body-area and departmental information systems that stream structured vitals and chief-complaint text to edge or workstation classifiers are the natural setting: the same R1/R2 distinction applies whenever a deterministic protocol is used as a training target. The framework is intended to be used *before* a confirmatory trial, not as a substitute for one.

II. RELATED WORK

Gradient boosting and deep models outperform rule-based scales on static sets (XGBoost AUC ≈ 0.84 vs. ESI 0.78 [12]; admission AUC ≈ 0.87 [13]). French-language work uses FlauBERT [17] for verbatim encoding and length-of-stay prediction [18]; large language models reach AUC 0.82–0.91 on ESI tasks [19] but hallucinate [20]. Mohamed, El-Menyar, Mekkodathil, and Al-Thani [21] reviewed 17 studies: all used static datasets and none separated algorithmic approximation from clinical validity.

Hybrid DES–MAS reduces compute versus pure MAS [22], [23] and shows that modest accuracy gains can shorten length of stay [24], [25]. Process mining on CHU Lille data reconstructs ED pathways [26]. Those studies model flow without AI decision modules. Most AI-triage economic papers report deterministic ICERs [12], [21]; we instead follow CHEERS 2022 [27] and the HAS 2020 economic-evaluation guide [28].

III. METHODS

A. Study design, data and ethics

Single-centre retrospective study, adult ED of CHU Lille, January 2018–December 2023. National ethical clearance (CESREES); French Health Data Hub ref. 27797006; GDPR and CNIL MR-004; records pseudonymized before analysis. The *calibration cohort* comprises 340,536 adult visits with complete structured data and a recorded nurse interview (from 459,259 visits; 118,723 excluded, 25.85%), split 60:20:20 by triage level. This figure is a multi-year cumulative extract (≈ 3 years at a ≈ 110 – 115 k visits/year site), not an annual volume. The *simulation perimeter* is 600,000 visits/year (five typical sites), the reference population for operational runs. Patient-level data are not shareable; analysis scripts are released.

B. Dual-regime references

The supervised target and the two evaluation references are distinct.

R1 (algorithmic approximation). The deterministic FRENCH algorithm [3] is applied to structured inputs (coded complaint, vitals, age, MEWS/qSOFA), yielding a five-level *ideal* label. Metrics against this label quantify recovery of the mapping. R1 never uses physician labels. It is necessary but not sufficient for clinical defensibility.

R2 (clinical validity). Five DESC-certified senior emergency physicians (≥ 10 years), blinded to nurse triage and AI outputs, annotated structured vignettes. Inter-rater agreement was substantial (Fleiss $\kappa = 0.72$, 95% CI 0.68–0.76; ICC(2,1) = 0.74, 0.70–0.78). Disagreements (23.5%) were resolved by Delphi ($\geq 4/5$), moderated by an external physician unaffiliated with model development. Re-moderation of 30 disagreements by a second external physician gave Cohen’s $\kappa = 0.91$ (0.84–0.97).

Completed versus planned R2 sample. Expert moderation is complete for $n = 400$ stratified cases. The target $n = 3,000$ is not finished. All clinical-validity claims in this revision use the $n = 400$ subset as the empirical sample. Figures labelled “ $n = 3,000$ projection” hold the $n = 400$ point estimate and scale the bootstrap half-width by $\sqrt{400/3000} \times 1.15$ to reflect the unobserved remainder. They are not measurements.

FRENCH versus experts. If a model recovers FRENCH (R1 $\kappa_w \approx 1$), then R2 cannot exceed $\kappa_w(\text{FRENCH}, \text{expert})$. On the $n = 400$ design, that ceiling is $\kappa_w = 0.81$ [0.77, 0.85], with 27.8% exact-label disagreement. Label mismatch, not a new medical algorithm, then explains most of the R1→R2 drop.

C. Architectures (instruments, not contributions)

TRIAGEMASTER, URGENTIAPARSE and EMERGINET were taken unchanged from the published TIAEU studies [15], [16] and retrained on 340,536 FRENCH labels. TRIAGEMASTER: Doc2Vec PV-DBOW plus a four-layer MLP. URGENTIAPARSE: flaubert-base-cased (110 M) concatenated with 35 structured variables into XGBoost, with SHAP.

EMERGINET: FlauBERT plus biLSTM on physiological series, VICReg [29]. No architecture is claimed as novel here.

D. Metrics and statistical tests

Primary endpoint: quadratic-weighted Cohen’s κ_w . R1 uses the full test set ($n = 68,108$; 95% bootstrap CI). R2 uses $n = 400$ (bootstrap, 1,000–10,000 resamples). Secondary: macro-F1, MAE, exact/near-class accuracy, one-vs-rest AUC (DeLong [30]), multiclass Brier, and Level 1–2 undertriage (true 1–2 assigned ≥ 3).

The nurse baseline is $\kappa_w = 0.65$; the pre-specified deployment bar is $\kappa_w \geq 0.80$. We test each model against these *fixed thresholds* (one-sided superiority/threshold tests). This is not a two-arm non-inferiority trial: a non-inferiority analysis would compare two interventions with a margin on a difference. Testing whether κ_w exceeds $0.65 + 0.15$ is a threshold test on a single arm. An interval that crosses 0.80 does not establish the deployment bar. Following Donner and Eliasziw [31], R2 differences below ≈ 0.05 – 0.08 are interpreted cautiously given gold-standard noise (ICC = 0.74).

E. Projection algebra

Let $\hat{\kappa}_{400}$ be the $n = 400$ point estimate and (L, U) its bootstrap interval. The $n = 3,000$ projection keeps $\hat{\kappa}_{400}$ and sets

$$L' = \hat{\kappa}_{400} - 1.15 \sqrt{400/3000} \cdot \frac{U-L}{2}, \quad U' = \hat{\kappa}_{400} + 1.15 \sqrt{400/3000} \cdot \frac{U-L}{2}. \quad (1)$$

The factor 1.15 inflates for stratification leftovers and unobserved cases. This is a *design-based* interval, not a posterior from 3,000 labels. We therefore refuse language such as “replicated R2” or “empirical $n = 3,000$ ”.

F. Worked reading of Table 1

URGENTIAPARSE has R1 $\kappa_w = 0.996$: it has essentially memorized FRENCH. Its R2 $\kappa_w = 0.81$ matches $\kappa_w(\text{FRENCH}, \text{expert})$ on the same subset. The 0.19 drop is what one should expect when the evaluation reference is swapped, not evidence of a mysterious model failure. TRIAGEMASTER’s larger undertriage (27%) shows that the remaining gap is not only the FRENCH–expert ceiling: architecture still matters under R2, but sample size $n = 400$ is too small to rank EMERGINET versus TRIAGEMASTER firmly (Donner–Eliasziw caution).

G. Hybrid DES–MAS and class-specific errors

A SimPy DES layer models 17 stations in a priority network driven by a non-homogeneous Poisson process $\lambda(h) \in [2.1, 12.8]$ patients/h, calibrated with PM4Py Inductive Miner (fitness > 0.95). A Mesa MAS layer hosts eight behavioural archetypes from k -means on process-mining traces. Each architecture is a hot-swappable decision component; human–AI interaction is a Bernoulli gate $p_{\text{acc}} = 0.85$. Scenarios: S1 manual; S2a/b/c the three models; S3 crisis ensemble at 200% load.

Residual errors are *not* calibrated from a scalar κ_w . Two models with similar kappa can have very different critical undertriage. We estimate a 5×5 class-conditional confusion matrix on the $n = 400$ R2 subset and sample predicted levels from $P(\hat{y} | y)$. Primary flow comparisons use three-day runs, $N = 40$ independent replications, stationarity checks (Augmented Dickey–Fuller, Geweke), Δ DMS (median length-of-stay vs. S1) with 95% percentile intervals.

H. Probabilistic health-economic analysis

CHEERS 2022 [27], hospital perspective, 3-year horizon. Base-case discount 3%; HAS 2020 recommends 2.5% [28], reported as sensitivity. QALYs are not credited for every correct five-class label. They accrue only from avoided Level 1–2 undertriage, with a Beta(3, 17) increment (mean 0.15) informed by emergency-care utility estimates [33]. Unit costs use 2025 ATIH GHS log-normals; TCO includes AI Act audit, training, drift monitoring. The CEAC is $\Pr(\Delta C - \lambda \Delta Q < 0)$; France has no official ICER threshold—50,000 €/QALY is a reference value.

Every economic figure is tagged by the quadruplet (perimeter; channels; regime; penetration). The previously reported 480% ROI is a *mono-site* hospital calculation ($\approx 115,000$ visits/year, undiscounted, no ramp-up). With a 50/80/100% ramp-up it falls to $\approx 341\%$ (3% or 2.5% discount: $\approx 338\%$). Conservative external communication uses the 210% CrI floor. Perspective-split €/visit reallocations that are not yet in-run are not reported. The national 50–80 M€/year range is a derived conservative scenario: 15,000 avoided undertriage adverse events $\times$ 3.3–5.3 k€, R2 only, implicit capture 8–13% of a four-channel theoretical total. R1 must not feed national projections. Indirect societal (human-capital) gains are excluded from the base case (conservative bias).

IV. RESULTS

A. Dual-regime concordance

Table I is the central empirical result. Under R1 all three models encode FRENCH ($\kappa_w \geq 0.89$). Under R2 on the *completed* $n = 400$ subset, point estimates remain 0.81, 0.74 and 0.69, but intervals are wide. URGENTIAPARSE [0.76, 0.85] *crosses* 0.80: the deployment bar is not demonstrated. EMERGINET and TRIAGEMASTER lie below 0.80. All three point estimates exceed the nurse baseline 0.65; TRIAGEMASTER’s interval [0.64, 0.75] touches that baseline. The mean R1–R2 drop is 0.20 (range 0.19–0.21). We describe it as the $n = 400$ empirical gap, not as a replicated multicentre finding.

The two regimes are complementary, not contradictory. R1 shows that the instruments can learn FRENCH; R2 asks what they learned relative to physicians. Prior studies scored only against reconstructed labels may overstate clinical validity by about 0.10–0.20 of κ_w , but that range is a working hypothesis until $n = 3,000$ is complete.

TABLE I
DUAL-REGIME κ_w . R1: OBSERVED VS. FRENCH ($n_{\text{test}} = 68,108$). R2: COMPLETED EXPERT SUBSET ($n = 400$) WITH BOOTSTRAP 95% CI, AND A DESIGN-BASED $n = 3,000$ PROJECTION (NOT A MEASUREMENT). UT: LEVEL 1–2 UNDERTRIAGE ON $n = 400$.

Architecture	R1 observed	R2 $n=400$	R2 proj. $n=3k$	UT
URGENTIAPARSE	0.996 [0.995,0.996]	0.81 [0.76,0.85]	0.81 [0.79,0.83]	9.6%
EMERGINET	0.939 [0.939,0.940]	0.74 [0.70,0.79]	0.74 [0.72,0.76]	11.3%
TRIAGEMASTER	0.895 [0.894,0.895]	0.69 [0.64,0.75]	0.69 [0.67,0.71]	27.0%
Nurse baseline	0.65	0.65	0.65	—
Deploy. bar	≥ 0.80	≥ 0.80	≥ 0.80	—

B. Class-specific errors and simulation

Table II reports flow from the calibrated DES–MAS (three-day, R2-anchored confusion injection) together with critical undertriage from the same class-conditional matrices. URGENTIAPARSE combines the best $n = 400$ κ_w with the lowest Level 1–2 undertriage (9.6%) and the largest flow benefit (Δ DMS -7.1% , 95% CI -9.4 to -4.8). TRIAGEMASTER’s $\kappa_w = 0.69$ coexists with 27.0% critical undertriage—a safety profile that a global kappa conceals. Forty independent replications were used for residual-error sampling; stationarity checks passed on the reported runs.

In earlier cycles that used operational nurse labels (distorted by under-classification), URGENTIAPARSE produced an apparent Δ DMS of -13.2% with $\kappa_w \approx 0.01$ and Level 1–2 sensitivity ≈ 0.002 . No isolated predictive metric flags that regime. Joint monitoring of undertriage and flow does.

TABLE II
SCENARIO COMPARISON, 3-DAY DES–MAS, CLASS-CONDITIONAL R2 ERROR INJECTION. CONCORDANCE IS κ_w AFTER THE $p_{\text{acc}} = 0.85$ GATE. UT: MEAN LEVEL 1–2 UNDERTRIAGE ACROSS REPLICATIONS.

Scenario	DMS _{med} (min)	κ_w	Δ DMS vs S1	UT
S1 manual	34.0	—	—	—
S2a TRIAGEMASTER	32.9	0.62	$-3.2\% [-5.5, -1.0]$	27.0%
S2b URGENTIAPARSE	31.6	0.77	$-7.1\% [-9.4, -4.8]$	9.6%
S2c EMERGINET	34.4	0.91	$+1.2\% [-0.8, +3.1]$	11.3%
S3 crisis (200%)	37.0	0.91	$+8.8\%$	9.6%

Process-mining calibration on 10,000 traces (45,379 events) recovered five pathway families (fitness = precision = 1.000 by construction on the same traces; generalisation = 0.971). Temporal Kolmogorov–Smirnov calibration passed 7/9 activities. These are internal consistency checks, not external validation.

C. Health-economic results

Table IV decomposes the ROI. The published 480% [210, 1,250] figure matches an undiscounted three-year mono-site hospital ratio and is *not* computed on the 600,000-visit simulation population. The in-run ICER anchor is 1,840 €/QALY; probabilistic dominance occurred in 32% of the original 50,000-iteration PSA; cost-effectiveness probability is 99.4% at 50,000 €/QALY and 87.2% at 10,000 €/QALY.

TABLE III

MONO-SITE ROI DECOMPOSITION (HOSPITAL, $\approx 115,000$ VISITS/YEAR, TCO 254 K€/YEAR). THE 480% FIGURE IS THE FLAT UNDISCOUNTED RATIO.

Assumption	3-year ROI
Flat, undiscounted (published 480%)	$\approx 475\%$
Ramp-up 50/80/100%, undiscounted	$\approx 341\%$
Ramp-up, discount 3%	$\approx 338\%$
Ramp-up, discount 2.5% (HAS)	$\approx 338\%$
PSA 95% CrI floor (conservative external)	210%

Five inputs explained 87% of ROI variance: inappropriate-hospitalisation avoidance (38%), GHS unit cost (21%), imaging avoidance (12%), annual ED volume (10%), P95 latency (6%). Allocation coefficients for a future in-run hospital/insurance split (variable-cost share θ , internal vs billed imaging ϕ , real-resource/tariff ratio ψ) are reserved as PSA parameters and are not turned into published €/visit point estimates in this revision. National 50–80 M€/year is labelled (national; undertriage-AE channel only; R2; capture 8–13%). Indirect societal gains are omitted from the base case.

TABLE IV

ECONOMIC OUTPUTS WITH PERIMETER TAGS. ROI PERCENTAGES ARE HOSPITAL, MONO-SITE. ICER/CEAC ARE IN-RUN ANCHORS. NATIONAL RANGE IS A CONSERVATIVE DERIVED SCENARIO.

Quantity (tag)	Value
ROI 3 y, flat, undiscounted (mono-site, R2)	480% [210, 1,250]
ROI with 50/80/100% ramp-up	$\approx 341\%$
Conservative communicated floor	210%
ICER (in-run)	1,840 €/QALY
$P(CE)$ at 50,000 €/QALY	99.4%
National, EI-undertriage only, R2, 8–13%	50–80 M€/year

V. DISCUSSION

EIMLIA-TEU makes two artefacts visible. At the metric layer, near-perfect FRENCH recovery coexists with substantially lower expert agreement on a still-small R2 sample; FRENCH-versus-expert disagreement is the structural ceiling. At the operational layer, a flow gain can hide critical undertriage unless class-conditional errors are injected. The framework is an evaluation discipline, not a replacement for FRENCH or for clinicians.

A practical reading for CDSS procurement is: (1) demand an R1 number, to check that the model implements the local scale; (2) demand an R2 number against an externally moderated panel, with the completed sample size in the caption; (3) demand Level 1–2 undertriage, not only κ_w ; (4) refuse a deployment claim if the 95% interval crosses the pre-specified bar; (5) refuse an economic claim that is not tagged by perimeter, channel, regime and penetration.

QALY modelling is deliberately conservative. Crediting 0.15 QALY to every correctly labelled visit would count trivial four-versus-five disagreements as health gains. Restricting the increment to avoided Level 1–2 undertriage ties utility to

preventable harm. The Beta(3, 17) prior is wide; a prospective EQ-5D-5L study should replace it.

Implications for wireless and departmental CDSS are direct: any system trained on a reconstructed protocol should report R1 and R2 separately. Edge devices that only see structured vitals are R1-complete by construction if they re-implement FRENCH; they still require R2 if the claim is clinical validity.

Limitations. (i) R2 is $n = 400$ of 3,000; ranking among models is qualitative. (ii) Single-centre retrospective design. (iii) Very short verbatims (often critical or communication-impaired patients) are under-represented. (iv) The first author developed the evaluated architectures; external moderation ($\kappa = 0.91$) mitigates but does not remove allegiance bias. (v) Economic priors mix review bounds with local TCO; perspective €/visit splits await in-run reallocation and are withheld. (vi) Three-day runs miss seasonal cascades. (vii) $p_{acc} = 0.85$ is extrapolated. (viii) Ethnicity is absent from French RPU/PMSI extracts. (ix) The open-source residual-error module uses synthetic labels matched to published κ_w because expert vignettes cannot be shared; the $n = 400$ point estimates themselves are the empirical anchors.

VI. SOFTWARE

The package `eimlia` implements dual-regime bootstrap, threshold tests (explicitly *not* non-inferiority), class-conditional confusion sampling, a priority-queue DES for residual-error stress tests, and perimeter-tagged economic reporting. The historical notebook `Rapport.ipynb` remains the full CHU Lille training/process-mining engine and requires local extracts that are not distributed. Deterministic ROI 10,260% / ICER 93 €/QALY from early notebook cells are withdrawn.

DATA AND CODE

Analysis scripts: <https://github.com/edlansiaux/EIMLIA-TEU-2>. Patient-level data are not shareable under CNIL MR-004 / GDPR. The first author developed the architectures used as instruments; mitigation via external consensus moderation is described in Section III-B.

VII. CONCLUSION

EIMLIA-TEU is a pre-clinical evaluation framework. Its contribution is the dual-regime protocol, class-conditional residual-error injection, and perimeter-tagged PSA—not a new triage model. Empirically (R1), three published architectures recover FRENCH. On the completed R2 subset ($n = 400$), expert-agreement point estimates are 0.81, 0.74 and 0.69; none currently demonstrates $\kappa_w \geq 0.80$ with an interval lying entirely above the bar. Economic figures are honest, labelled projections: 480% ROI is mono-site; 210% is the conservative floor; 50–80 M€/year is an undertriage-only national scenario. Next steps are completing $n = 3,000$ expert moderation, a prospective multicentre cluster-randomised evaluation with EQ-5D-5L, and in-run reallocation of hospital / insurance / societal cash-flows. Scripts: <https://github.com/edlansiaux/EIMLIA-TEU-2>.

REFERENCES

- [1] DREES, “Les urgences hospitalières, qu’en sait-on?” DREES, 2023.
- [2] P. Taboulet, V. Moreira, L. Haas, A. Porte, J.-F. Blitz, D. Thabut, and J.-P. Fontaine, “Triage with the FRENCH scale: reliability and validity,” *Eur. J. Emerg. Med.*, vol. 16, pp. 61–67, 2009.
- [3] P. Taboulet, V. Moreira, L. Haas, A. Porte, J.-F. Blitz, D. Thabut, and J.-P. Fontaine, “Validité de la FRENCH,” *Ann. Fr. Med. Urgence*, 2019.
- [4] N. Farrohknia, M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, and K. E. Göransson, “Emergency department triage scales and their components: a systematic review,” *Scand. J. Trauma Resusc. Emerg. Med.*, vol. 19, p. 42, 2011.
- [5] N. Parenti, M. L. Bacchi Reggiani, P. Iannone, D. Percudani, and D. Dowding, “A systematic review on the Manchester triage system,” *Int. J. Nurs. Stud.*, vol. 51, no. 7, pp. 1062–1069, 2014.
- [6] A. Zaboli, F. Brigo, S. Sibilio, M. Mian, and G. Turcato, “Establishing a common ground: the future of triage systems,” *BMC Emerg. Med.*, vol. 24, p. 148, 2024.
- [7] M. Ebrahimi, A. Heydari, R. Mazlom, and A. Mirhaghi, “The reliability of the Australasian triage scale,” *World J. Emerg. Med.*, vol. 6, no. 2, pp. 94–99, 2015.
- [8] D. R. Sax, E. M. Warton, D. G. Mark, D. W. Ballard, M. E. Reed, and D. R. Vinson, “Association between emergency department undertriage or overtriage and patient outcomes,” *Ann. Emerg. Med.*, 2026.
- [9] M. Christ, F. Grossmann, D. Winter, R. Bingisser, and E. Platz, “Modern triage in the emergency department,” *Dtsch. Arztebl. Int.*, vol. 107, no. 50, pp. 892–898, 2010.
- [10] E. Demoly and T. Deroyon, “Urgences: la moitié des patients y restent plus de 3 heures,” *Études et Résultats DREES*, no. 1334, 2025.
- [11] H. A. Moll, “Challenges in the validation of triage systems at emergency departments,” *J. Clin. Epidemiol.*, vol. 63, no. 4, pp. 384–388, 2010.
- [12] S. Levin, M. Toerper, E. Hamrock, J. S. Hinson, S. Barnes, H. Gardner, A. Dugas, B. Linton, R. Kirsch, and G. Kelen, “Machine-learning-based electronic triage,” *Ann. Emerg. Med.*, vol. 71, no. 5, pp. 565–574, 2018.
- [13] W. S. Hong, A. D. Haimovich, and R. A. Taylor, “Predicting hospital admission at emergency department triage,” *PLoS ONE*, vol. 13, no. 7, p. e0201016, 2018.
- [14] E. Klang, Y. Barash, R. Soffer, S. Bechler, O. S. Ohayon, T. Sklair-Levy, H. Greenspan, E. Konen, and M. M. Amitai, “Deep learning for critical findings in head CT,” *Lancet*, vol. 395, no. 10232, pp. 1365–1373, 2020.
- [15] É. Lansiaux, H. Zgaya-Biau, and M. Ammi, “Development and comparative evaluation of three AI models for predicting triage,” in *Proc. IEEE/ACM BDCAT*, 2025.
- [16] É. Lansiaux, H. Zgaya-Biau, and M. Ammi, “Artificial intelligence models for predicting triage in emergency departments,” *JMIR Med. Inform.*, vol. 14, p. e83318, 2026.
- [17] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, “FlauBERT: unsupervised language model pre-training for French,” arXiv:1912.05372, 2020.
- [18] J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, A. Buisson, and S. Sanchez, “Prediction of hospital length of stay using unstructured data,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 351, 2021.
- [19] O. Kilic, A. Yilmaz, M. Yildirim, and others, “Emergency department triaging using ChatGPT,” *Sci. Rep.*, vol. 14, p. 22437, 2024.
- [20] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
- [21] A. H. Mohamed, A. El-Menyar, A. Mekkodathil, and H. Al-Thani, “Applications of AI in emergency medicine triage,” *Cureus*, vol. 16, no. 1, p. e51813, 2024.
- [22] Y. Liu, J. Wang, X. Chen, and L. Zhang, “Hybrid simulation modelling of emergency departments,” *J. Simul.*, vol. 19, no. 2, pp. 215–230, 2025.
- [23] T. Moyaux, Y. Liu, G. Bouleux, and E. Marcon, “Agent-based architecture of the digital twin for an emergency department,” *Sustainability*, vol. 15, no. 4, p. 3412, 2023.
- [24] H. Zgaya, S. Hammadi, E. Renard, and others, “A workflow model to analyse pediatric emergency overcrowding,” in *Proc. MIE*, 2014, pp. 338–342.
- [25] F. Ajmi, H. Zgaya, S. B. Othman, and S. Hammadi, “Agent-based dynamic optimization for managing the patient’s pathway,” *Simul. Model. Pract. Theory*, vol. 96, p. 101935, 2019.
- [26] J. Le Lay, V. Augusto, X. Xie, E. Alfonso-Lizarazo, B. Lamarsalle, and A. Lautrette, “COVID-19 bed management using process mining and DES,” *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 3, pp. 3080–3091, 2024.
- [27] D. Hureau, M. Drummond, F. Augustovski, E. de Bekker-Grob, A. H. Briggs, C. Carswell, L. Caulley, N. Chaiyakunapruk, D. Greenberg, E. Loder, J. Mauskopf, C. D. Mullins, S. Petrou, R. F. Pwu, and S. Stanisewska, “Consolidated health economic evaluation reporting standards 2022 (CHEERS 2022) statement,” *Value Health*, vol. 25, no. 1, pp. 3–9, 2022.
- [28] Haute Autorité de Santé, *Choix méthodologiques pour l’évaluation économique à la HAS*, 2020.
- [29] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: variance-invariance-covariance regularization,” in *ICLR*, 2022.
- [30] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated ROC curves,” *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
- [31] A. Donner and M. Eliasziw, “A goodness-of-fit approach to inference for the kappa statistic,” *Stat. Med.*, vol. 11, no. 9, pp. 1155–1168, 1992.
- [32] S. C. Brailsford, P. R. Harper, B. Patel, and M. Pitt, “Simulation and modelling in health care: a literature review,” *J. Simul.*, vol. 3, no. 3, pp. 130–140, 2009.
- [33] D. Castillo, J. S. Lee, A. H. Kaji, and D. L. Schriger, “Health utility estimates for QALYs in emergency care: a systematic review,” *J. Am. Coll. Emerg. Physicians Open*, vol. 3, no. 5, p. e12789, 2022.