Scientific Reports

https://doi.org/10.1038/s41598-026-65349-z

Article in Press

Intensity–complexity post-triage prioritization to shorten waiting times in emergency departments

Received: 29 January 2026

Accepted: 30 July 2026

Published online: 22 August 2026

Cite this article as: Marchese D.N., Cattaneo M., Morlotti C. *et al.* Intensity–complexity post-triage prioritization to shorten waiting times in emergency departments. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-65349-z

Dario Nicola Marchese, Mattia Cattaneo, Chiara Morlotti, Stefano Paleari, Franco Cammarota, Francesco Paleari, Sara Fascendini, Patrizio Sale, Francesco Locati & Filippo Manelli

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

ARTICLE IN PRESS

Intensity–Complexity Post-Triage Prioritization to Shorten Waiting Times in Emergency Departments

Dario Nicola Marchese[1,*], Mattia Cattaneo[1], Chiara Morlotti[1], Stefano Paleari[1], Franco Cammarota[2,3], Francesco Paleari[4], Sara Fascendini[2], Patrizio Sale[2], Francesco Locati[5], and Filippo Manelli[6]

[1]Department of Management, Information, and Production Engineering, University of Bergamo, Dalmine, 24044, Italy
[2]Fondazione Europea di Ricerca Biomedica (FERB Onlus), Milano, 20123, Italy
[3]RigenerAI, Bari, 70132, Italy
[4]Department of Neurology, Vita-Salute San Raffaele University, Milan, 20132, Italy
[5]ASST Papa Giovanni XXIII, Bergamo, 24127, Italy
[6]ASST Bergamo Est, Seriate, 24068, Italy
[*]dario.marchese@unibg.it

ABSTRACT

Emergency department triage reliably identifies patients at immediate risk, although it compresses heterogeneous clinical profiles into a small number of urgency categories, leaving discretion in how patients are prioritized within the same category. We propose a second-stage prioritization framework that makes two complementary clinical dimensions explicit: intensity (acute physiological severity) and complexity (longer-term vulnerability and management burden). Using ED triage data from a medium-large hospital, we derive patient-level intensity and complexity measures for patients assigned to a high-volume intermediate-acuity class through a domain-adapted large language model pipeline that integrates structured variables with free-text triage documentation that we validate against independent physicians assessment on a representative 10% subsample. We then evaluate the operational impact using a discrete-event simulation of ED flow under congestion, comparing first-come–first-served sequencing with an intensity–complexity–aware prioritization rule. The proposed approach reduces average waiting time in the targeted class by 23-39 minutes per patient and yields spillover reductions in lower-acuity classes of approximately 24-44 minutes per patient. Robustness analyses indicate that these gains persist under alternative modeling assumptions, including a context-specific empirical non-FCFS comparator, and validation-calibrated LLM label noise. These improvements arise by limiting delay for patient profiles whose service requirements are most sensitive to waiting time, thereby mitigating downstream congestion effects.

ARTICLE IN PRESS

ARTICLE IN PRESS

1 Introduction

Emergency departments (EDs) rely on triage systems designed to rapidly identify patients at risk of clinical deterioration and to prioritize care primarily on the basis of acute physiological severity[1,2]. However, the assignment of patients to a limited number of urgency categories limits the ability of triage systems to capture heterogeneity. EDs increasingly serve populations whose vulnerability may differ not only in acute clinical intensity, but also in clinical complexity, including multimorbidity, accumulated physiological burden, frailty, cognitive impairment, psycho-social vulnerability, and functional dependence—features that are particularly prevalent among older adults and strongly associated with adverse outcomes[3–5]. These dimensions can influence care trajectories, resource utilization, and clinical outcomes, often independently of acuity-based triage classification[6]; however, these aspects are only partially reflected in routine triage practice. Emerging evidence suggests that incorporating complexity-informed approaches could improve ED efficiency and patient outcomes[7]. As a result, patients sharing the same triage category may face markedly different operational needs, creating ambiguity in within-class prioritization.

Once triage is completed, patients are assigned to acuity categories that drive initial ED assessment, but leave discretion in how patients are prioritized within each category. In practice, existing triage systems offer limited guidance on patients sequencing within the same urgency class and do not provide a structured way to account for differences in clinical intensity or clinical complexity beyond the initial urgency assignment. As a result, clinicians may often default to first-come-first-served ordering, but observational studies suggest that this principle is frequently–and informally–relaxed as waiting times increase, with forms of delay-dependent prioritization in response to perceived vulnerability and operational pressure[8,9]. Consequently, within-class prioritization is largely driven by individual clinical intuition and experience rather than by systematic, reproducible criteria.

In this work, we focus on the post-triage operational decision of determining which patient should be seen next among those sharing the same acuity level, by incorporating clinical complexity. Triage categorization itself is left unchanged, as it remains essential for rapid risk stratification. This downstream sequencing decision represents a critical yet underexplored opportunity to improve both efficiency and equity in emergency care. Operationalizing such a second-stage prioritization requires patient-level indicators that go beyond acuity labels and can be computed in real time using information already available at triage. In principle, these indicators could be derived from structured variables, rule-based scores, or statistical and machine-learning models trained to predict downstream outcomes or resource use. However, many clinically salient aspects of vulnerability (particularly those related to complexity) are predominantly documented in unstructured free-text fields and are difficult to capture through predefined rules without substantial information loss.

Recent advances in natural language processing provide means to formalize such decision process without altering established workflows. In particular, domain-adapted large language models (LLMs) can extract indicators of both acute clinical intensity and longitudinal clinical complexity from information routinely collected at triage, including brief medical histories and free-text nursing documentation. Prior work in digital medicine has demonstrated the potential of AI-based approaches in ED settings to characterize severity and predict deterioration[10–14], while also highlighting the limitations of existing digital triage tools[15,16]). Together, these findings underscore the potential of AI systems that augment clinician-led prioritization rather than bypass it. This aligns with evidence from the literature suggesting that digital triage tools are best positioned as decision support systems rather than replacements for clinician judgment, reflecting healthcare professionals' preferences and the need for human-in-the-loop decision support[17,18].

Within this context, we introduce a second-stage prioritization framework that leverages LLM-derived measures of clinical intensity and clinical complexity to support within-class prioritization at the ED. These measures are developed and validated in collaboration with physicians, ensuring that they are interpretable and consistent with established clinical reasoning. We assess the operational implications of adopting this approach using a discrete-event simulation that captures key ED dynamics, including congestion, waiting times, service interactions, and heterogeneity in patient service requirements. By jointly representing clinical intensity and clinical complexity, the proposed framework provides a structured, data-driven basis for fine-grained prioritization among patients assigned to the same acuity category, complementing existing triage systems rather than replacing them.

2 Results

2.1 Data and outcome modeling

Our analysis is based on an Italian ED that records approximately 46,500 admissions per year, corresponding to an average of 127 patient arrivals per day. Patients are triaged into 4 urgency levels, ranging from level 1 (UL1, the most urgent) to level 4 (UL4, the least urgent). Specifically, UL1 patients account for 1% of arrivals, UL2 for 15%, UL3 for 77% and UL4 for the remaining 7%. Each record comprises structured anonymized clinical information—triage category, vital signs, age, sex—with unstructured free-text fields, including the presenting complaint and the brief anamnesis routinely collected at triage. These

ARTICLE IN PRESS

data collectively reflect the information available to clinicians at first assessment and constitute the empirical basis for modeling post-triage prioritization decisions among patients sharing the same urgency level.

Patients assigned the same triage category may differ in both the severity of their current presentation and the broader clinical burden associated with their care. Evidence from the literature shows that patients classified within the same triage urgency level can exhibit markedly different downstream risks, underscoring heterogeneity that is not captured by acuity alone, including variation in hospitalization risk and unplanned revisits with hospitalization[19–21]. To represent this heterogeneity in a structured and operationally meaningful way, we introduce two distinct dimensions of patient health status: *clinical intensity* and *clinical complexity*.

Clinical intensity captures the severity and immediacy of a patient's condition at arrival, reflecting short-term physiological risk and the urgency with which medical intervention is required. Intensity is driven by acute abnormalities observable at triage and summarizes features indicative of imminent clinical deterioration rather than longer-term care needs. In this study, intensity assessment is guided by markers of physiological derangement—heart rate, systolic blood pressure, oxygen saturation, and body temperature—as well as the presence of trauma, acute symptom burden, or exacerbations of chronic disease. Based on these elements, clinical intensity is represented using an ordinal scale with three ordered levels, from 1 (low) to 3 (high), reflecting increasing degrees of acute severity.

Clinical complexity, in contrast, characterizes the anticipated management burden associated with the patient, independently of immediate acuity. It reflects vulnerability arising from clinical characteristics that influence diagnostic effort and care coordination rather than urgency. Factors informing complexity include age, sex, multimorbidity, chronic disease burden, disability or loss of autonomy, pregnancy, and medication patterns indicative of long-term conditions. As with intensity, clinical complexity is represented using three-level ordinal scale, from 1 (low) to 3 (high), capturing increasing degrees of management complexity.[1]

Clinical intensity and clinical complexity represent distinct but interacting dimensions of patient health status. Patients with similar acute presentations may differ substantially in complexity, while highly complex patients may present with relatively low immediate intensity. This distinction is central to our framework and motivates a second-stage prioritization that structures sequencing decisions within a given triage category, where immediate clinical risk has already been assessed. Accordingly, the ordinal representation of intensity and complexity are designed to capture clinically meaningful differences within-class heterogeneity and inform downstream prioritization decisions, without replacing established triage classifications or override acuity-based principles.

From an operational perspective, clinical intensity and clinical complexity are introduced at triage using a large language model–based pipeline that integrates unstructured free-text documentation with structured clinical variables. This approach yields patient-level measures of intensity and complexity that are subsequently used for validation analyses and simulation experiments. In this study, clinical intensity and clinical complexity are inferred exclusively for UL3 triage category, reflecting the operational setting in which post-triage sequencing decisions are both most frequent and least formally structured.

ARTICLE IN PRESS

2.2 LLM-based evaluation performance

Although clinically intuitive, neither intensity nor complexity is directly observable as a single structured variable at triage. Instead, both constructs emerge from the synthesis of heterogeneous information sources, including physiological measurements, demographic characteristics, and, critically, free-text documentation such as presenting complaints and brief anamneses. These unstructured texts encode important contextual elements that are central to critical reasoning, such as symptoms, chronic diseases, care dependencies, medication use, allergies, and social or functional vulnerability. At the same time, triage documentation is highly heterogeneous. It is produced under time pressure, by different ED professionals, and frequently relies on abbreviations, acronyms, shorthand expressions, and non-standardized phrasing. As a result, operationalizing intensity and complexity in a way that is scalable, reproducible, and aligned with clinical reasoning requires methods capable of jointly interpreting unstructured clinical language and structured triage data, together with a systematic process for text normalization and standardization before high-level understanding and inference. Accordingly, our approach combines rule-based preprocessing with natural language semantic enrichment of clinical text, followed by calibrated LLM-based clinical inference (further details are provided in subsection 4.2).

LLM results-validation. We evaluate the LLM-derived intensity and complexity measures against physician annotations on a validation set of 2,000 admissions, corresponding to 10% of the analysis sample and drawn to be representative of

[1] The use of a three-level ordinal scale is chosen to balance clinical expressiveness with annotation stability[22]. The objective of the labels is not to reproduce fine-grained diagnostic judgments, but to distinguish low, intermediate, and high levels of intensity and complexity in a way that can be applied consistently to brief triage documentation. A more granular scale would imply a degree of precision that is unlikely to be reliable given the concise and heterogeneous nature of triage notes. The validation results support this design choice: disagreements between the LLM and physician annotations are concentrated primarily between adjacent categories, while extreme misclassification is rare, indicating that the three-level scale preserves clinically meaningful ordering without requiring unstable fine distinctions.

ARTICLE IN PRESS

the approximately 20,000 admissions retained after filtering for sufficient free-text triage documentation and removal of extreme operational outliers from the original dataset (see subsection 4.1 for details). To further support interpretability, Appendix F presents complementary regression-based analysis. This analysis is not used to generate, predict, or benchmark the LLM-derived clinical intensity and complexity, but rather to characterize how observable clinical features relate to the LLM-inferred constructs.

Because both clinical intensity and clinical complexity are ordinal constructs, we use quadratic-weighted Cohen's kappa as the primary validation metric. The LLM achieves moderate ordinal agreement for clinical intensity ($\kappa_w = 0.49$, 95% CI: 0.452–0.527), and substantially stronger agreement for clinical complexity ($\kappa_w = 0.68$, 95% CI: 0.648–0.712).

Additionally, Figure 1 reports the confusion matrices comparing physician-assigned and LLM-inferred levels for clinical intensity (I) and complexity (C), for which an overall accuracy of 60.5% and 83% is obtained, respectively.[2]

For clinical intensity, as shown in Figure 1a the LLM correctly identifies physician-labeled $I = 1$ cases with 40.3% recall. Among these cases, 55.1% are assigned to the adjacent intermediate category $I = 2$, while only 4.6% are assigned to the highest category $I = 3$. The LLM also correctly identifies $I = 2$ in 694/911 cases (76.2% recall) and $I = 3$ with 62.3% recall. Misclassification involving the two extreme categories is limited: only 4.6% of physician-labeled $I = 1$ cases are inferred as $I = 3$, and 4.6% of physician-labeled $I = 3$ cases are inferred as $I = 1$. Therefore, while the model preserves the broad ordinal separation between low and high intensity, it tends to concentrate ambiguous cases around the intermediate category and, specifically, to compress the distinction between low and intermediate intensity by over-assigning some physician-labeled low-intensity patients to the intermediate category.

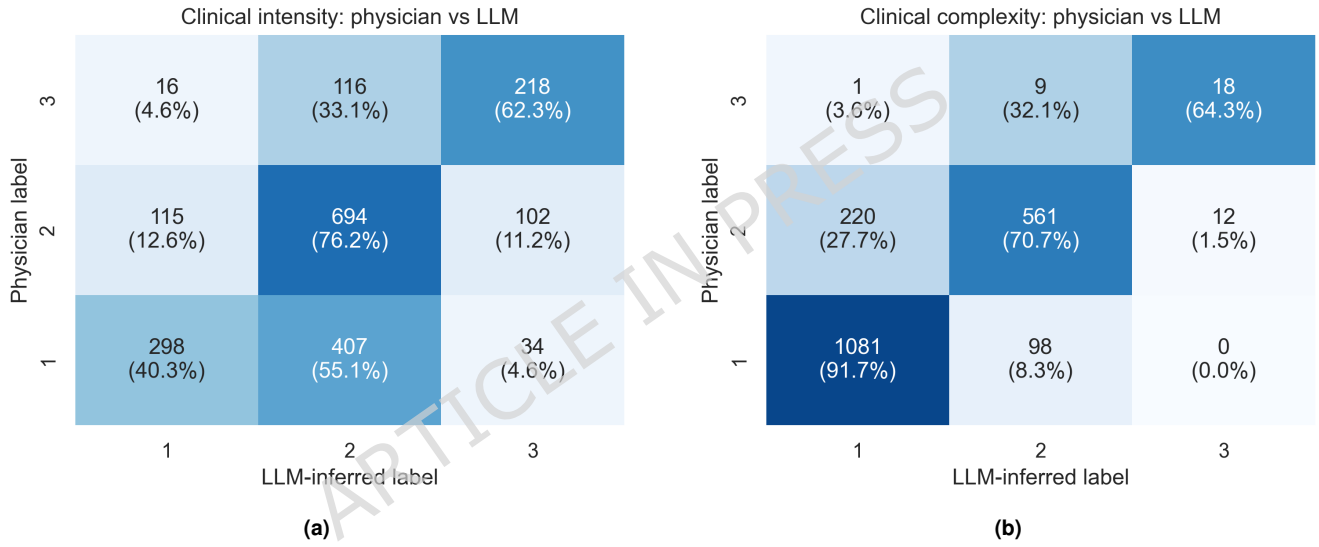


Figure 1. Physician vs LLM-inferred labels of clinical intensity (a) and clinical complexity (b) on a validation set of 2,000 admissions.

For clinical complexity, agreement is substantially stronger as reported in Figure 1b. $C = 1$ is correctly identified in 1,081/1,179 cases (91.7% recall), and $C = 2$ with 70.7% recall. Although $C = 3$ is rare (1.4% of the sample), it is correctly identified in 18/28 cases (64.3% recall). Notably, the LLM does not escalate the physician-labeled low-complexity cases to the highest complexity level.

As for intensity, disagreements occur primarily between adjacent categories (i.e., $1 \leftrightarrow 2$ or $2 \leftrightarrow 3$). Extreme misclassification is virtually absent for complexity, highlighting that errors are confined to neighboring levels rather than reflecting incorrect clinical characterizations of patients. Instead, the direction of disagreement differs across the two dimensions. For clinical intensity, misclassifications are predominantly upward, with inferred values concentrated at the intermediate level, $I = 2$. For clinical complexity, errors are more often downward, with 27% of physician-labeled $C = 2$ cases inferred as $C = 1$.

In summary, these results indicate that the LLM-based pipeline produces intensity and complexity assignments that are consistent with physician annotations in their ordinal structure. At the same time, the model preserves a clear separation between extreme categories and exhibits disagreement primarily in intermediate cases where clinical judgment is inherently variable.

[2]Given the ordinal nature of the labeling and the documented variability in clinical assessments based on triage data, this level of exact agreement is comparable with prior digital triage studies reporting prediction of urgency categories from routine clinical data[23].

ARTICLE IN PRESS

NEWS2-based baseline for clinical intensity. In addition, we evaluate whether the clinical intensity construct could be approximated using structured physiological information alone. To this end, we construct a NEWS2-based baseline from vital signs and altered mental status, when available, and compare it with the proposed LLM-derived intensity measure[3].

This comparison should be interpreted in light of the scope of the present study: NEWS2 is designed to detect acute physiological deterioration across the broader emergency-care population and it has been shown to correlate with 30-days mortality within in-hospital settings[24]. Our analysis, instead, focuses on patients already assigned to the same acuity triage class (UL3). Therefore, high NEWS2 values are not expected to be frequent in the validation sample. The question we address is not whether NEWS2 can immediately identify unstable patients, but whether a such score provides sufficient within-class variation to reproduce physician-assigned post-triage intensity among UL3 patients.

Because NEWS2 is defined on a numerical risk scale rather than on the three-level intensity scale used in this study, we evaluate it in two complementary ways. First, we assess the ordinal association between the NEWS2 score on its original numerical scale and physician-assigned intensity labels. Second, we map NEWS2 into the three-level intensity scale using pre-specified NEWS2 thresholds and compare this mapped score with physician intensity labels using ordinal agreement metrics. The mapped NEWS2 baseline classifies almost all validation-set patients as low physiological risk on the three-level scale, with 99.95% assigned to level 1 and 0.05% assigned to level 2. This concentration is consistent with the fact that the validation set consists of UL3 patients, but also indicates that NEWS2 provides limited discrimination for within-class intensity stratification.

Consistently, the NEWS2 score on its original numerical scale shows only weak monotonic association with physician intensity labels (Spearman $\rho = 0.116$, 95% CI 0.074–0.16; Kendall $\tau_b = 0.107$, 95% CI 0.068–0.147). In contrast, the LLM-derived intensity label shows substantially stronger ordinal association with physician intensity (Spearman $\rho = 0.496$, 95% CI 0.46–0.529; Kendall $\tau_b = 0.465$, 95% CI 0.43–0.497). After mapping NEWS2 to the three-level intensity scale, the NEWS2 baseline achieves negligible ordinal agreement with physician labels ($QWK = 0.001$) and exact agreement of 36.95%. By contrast, the LLM-derived intensity achieves $QWK = 0.49$ and exact agreement of 60.5%.

These results indicate that structured vital signs and altered mental status capture acute physiological instability, but do not reproduce the physician-assigned post-triage intensity construct used in this study. In the UL3 setting, physician intensity also reflects symptoms, trauma, exacerbations, age, functional status, comorbidities, medications, and contextual information available in triage documentation. Therefore, the results suggest that the LLM-derived intensity construct captures clinically relevant information beyond structured physiological scoring alone within the UL3 setting.

2.3 Simulation-based analysis

To assess the operational consequences of incorporating LLM-derived clinical intensity and clinical complexity into post-triage sequencing decisions, we conduct a discrete-event simulation of ED patient flow. The simulation is calibrated using empirical arrival processes and service-time distributions estimated from the ED data, and is constructed to hold demand and capacity constant while varying only the within-class sequencing rule. Details of the simulation setup are reported in Appendix C. This design allows differences in system performance to be attributed to alternative sequencing policies rather than to changes in arrivals, capacity, or other operational conditions.

Our primary benchmark is a first-come-first-served (FCFS) sequencing discipline within urgency class. FCFS is not intended to be a literal representation of real ED practice. Rather, it provides a transparent, reproducible, and policy-neutral structural benchmark against which the incremental effect of intensity- and complexity-informed prioritization can be evaluated under identical demand, capacity, and service-time conditions. Because real ED sequencing may deviate from strict FCFS, we also construct an additional empirical non-FCFS comparator based on the observed within-UL3 service order in the historical data. This comparator is not used as the main benchmark, nor is it intended to define a generalizable policy. Instead, it provides a context-specific approximation of the local "as-is" sequencing pattern observed in the studied ED. Unless otherwise specified, performance effects are therefore reported relative to FCFS, while the empirical dynamic-rank comparator is reported as an additional context-specific analysis.

To further isolate the role of sequencing decisions, we adopt an aggregate representation of service capacity, modeled as a single equivalent service process capturing the joint availability of clinical space and personnel. This abstraction avoids institution-specific assumptions and ensures that observed performance differences arise from prioritization rules rather than from detailed resource configurations.

The simulation is run over multiple independent replications to account for stochastic variability in patient arrivals and service times. We compare two main within-class sequencing policies: i) the neutral benchmark discipline, namely FCFS, applied uniformly within each urgency class, and ii) the intensity–complexity–aware prioritization rule, the Priority-IC, which

[3]Note that NEWS2 score is considered only to individuals with age greater than 16 and no pregnant, who constitute the 95% of the validation sample. Clinicians calculating NEWS2 assumed consciousness, room air and parameters to be normal when the latter clinical information were not reported on triage patient data

ARTICLE IN PRESS

is applied exclusively to patients in the target class, UL3, while all other urgency classes continue to follow FCFS sequencing. Priority-IC ranks UL3 patients using a composite index that combines inferred clinical intensity and clinical complexity. The relative weight assigned to the two dimensions is governed by the parameter α, with lower values emphasizing complexity and higher values emphasizing intensity (i.e., $\alpha \times I + (1-\alpha) \times C$—see subsection 4.2 for further details). Higher values of the index correspond to higher priority for service. Varying α allows the simulation to explore different prioritization regimes, ranging from complexity-focused to intensity-focused ordering within the same urgency class.

For UL3 patients, service times are modeled endogenously and depend on both patient characteristics and experienced waiting time. Specifically, service times increase with waiting time at different rates across combinations of clinical intensity and clinical complexity, capturing heterogeneous deterioration and empirically observed interactions between congestion, patient profile, and care workload. These relationships are estimated from ED data using regression analysis and are held fixed across simulation runs (details in Appendix B). As a result, differences in system performance under FCFS and Priority-IC arise from how sequencing decisions interact with heterogeneous patient profiles, rather than from changes in arrivals or service capacity.

Waiting time - results. Table 1 summarizes performance under FCFS and Priority-IC across urgency levels and values of the prioritization parameter α, assuming a fixed number of available resources $n = 8$. Sensitivity analysis with respect to the number of resources is reported in Appendix D. For each urgency level and value of α, we report mean waiting time under Priority-IC and FCFS, absolute and relative differences with respect to FCFS (Δ, $\Delta_{\%}$), and the implied time saved relative to the FCFS benchmark. Across simulation replications, Priority-IC consistently reduces average waiting times relative to FCFS, with

Urgency levels (UL)	α	Priority-IC wait [min]	FCFS wait [min]	Δ [min]	$\Delta_{\%}$ [%]	Time saved [h]
1	0.0	4.53	4.8	-0.27	-5.55	0.01
1	0.4	4.5	4.8	-0.3	-6.24	0.01
1	0.7	4.63	4.8	-0.17	-3.44	0.0
1	1.0	4.57	4.8	-0.23	-4.77	0.0
2	0.0	8.65	9.14	-0.49	-5.37	0.15
2	0.4	9.39	9.14	0.25	2.69	-0.07
2	0.7	9.16	9.14	0.02	0.18	-0.0
2	1.0	9.04	9.14	-0.10	-1.09	0.03
3	0.0	112.7	136.34	-23.64	-17.34	35.32
3	0.4	96.85	136.34	-39.49	-28.96	59
3	0.7	97.31	136.34	-39.03	-28.62	58.31
3	1.0	101.89	136.34	-34.45	-25.27	51.48
4	0.0	146.38	170.29	-23.91	-14.04	1.86
4	0.4	126.39	170.29	-43.90	-25.78	3.41
4	0.7	128.03	170.29	-42.26	-24.82	3.29
4	1.0	135	170.29	-35.29	-20.73	2.74

Table 1. Average waiting time and comparison with FCFS baseline by urgency level and prioritization parameter α. Number of resources 8.

improvements concentrated in the high-volume, lower-acuity urgency classes. Importantly, the benefits are not limited to the urgency class to which Priority-IC is directly applied (UL3), but also extend to UL4 through the shared service process.

The largest improvements occur for UL3 patients whose average waiting times decrease by 23-to-39 minutes (17-to-28%) per patient relative to FCFS, depending on the value of α. UL3 patients account for the largest share of arrivals in the system, thus reductions in their waiting times directly affect system congestion and queue dynamics. In practice, because all urgency classes draw from the same aggregate service capacity, reductions in waiting times for one large patient group (UL3) translate into shorter queues and lower system-wide congestion, benefiting patients in other urgency class as well. As a consequence, waiting times for UL4 patients also decrease substantially under Priority-IC, with relative reductions ranging between 14% (23 minutes) and 25% (43 minutes). Under FCFS sequencing, UL4 patients—who constitute a smaller fraction of total arrivals—tend to accumulate at the back of the queue during periods of congestion. When prioritization within UL3 reduces the buildup of long queues, UL4 patients experience shorter waiting times even though their within-class prioritization is unchanged. By contrast, UL1 or UL2 patients exhibit small absolute changes in waiting time. These patients already receive prompt service under FCFS due to their high urgency, and their baseline waiting times are therefore low.

System level performance is sensitive to the choice of α. Although differences across values are modest, intermediate configurations ($\alpha = 0.4 - 0.7$) consistently yield the largest reductions in waiting times and the greatest total time saved. In contrast, placing all weight on either complexity (i.e., $\alpha = 0$) or intensity ($\alpha = 1$) results in smaller—though still positive—gains.

The positive gains observed at the extreme values of α have a direct operational interpretation. Although $\alpha = 0$ and $\alpha = 1$

ARTICLE IN PRESS

ignore one of the two dimensions, they still replace arrival-order sequencing, represented by FCFS policy, with a clinically ordered rule along a dimension that is correlated with delay sensitivity. When $\alpha = 0$, high-complexity patients are served earlier; when $\alpha = 1$, high-intensity patients are served earlier. Because the estimated service-time response to waiting increases with either clinical complexity or clinical intensity, advancing these patients reduces waiting-induced service-time inflation for at least part of the UL3 population. This explains why both single-dimension policies improve upon FCFS. Their gains are smaller than those obtained with intermediate values of α because patients who are vulnerable along the non-prioritized dimension may still experience prolonged waits, partially offsetting the benefits obtained for the prioritized subgroup.

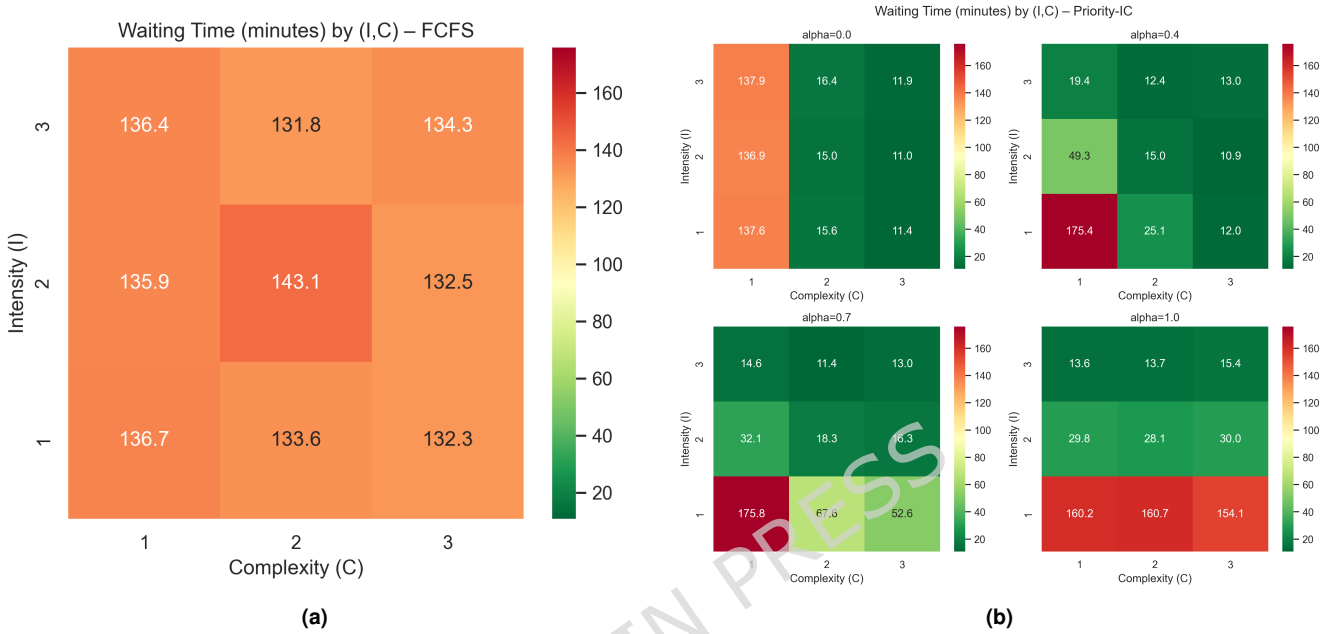


Figure 2. Average waiting times across (I, C) profiles of UL3 patients under both FCFS and Priority-IC prioritization rules.

To better understand how Priority-IC generates improvements reported above, we examine waiting times for UL3 patients stratified by clinical intensity (I) and clinical complexity (C). Figure 2 reports mean waiting times across all simulation replications for each (I, C) profile under FCFS and under Priority-IC for different values of α. As depicted in Figure 2a, average waiting times under FCFS for UL3 patients are largely homogeneous across intensity-complexity profiles, ranging between approximately 132 and 143 minutes. No systematic pattern emerges, indicating that FCFS treats UL3 patients as exchangeable within the urgency class, with observed differences reflecting stochastic variability. Consequently, patients with higher clinical intensity or complexity experience waiting times comparable to those of lower-risk profiles under FCFS.

In contrast, Priority-IC, induces a structured redistribution of waiting times across intensity-complexity profiles, as represented in Figure 2b. Across all values of α, patients with higher clinical complexity and/or higher clinical intensity experience substantially shorter waiting times relative to FCFS, while patients with low intensity and low complexity experience longer waits. The magnitude and direction of this redistribution depend on the value of α, which governs the relative emphasis placed on intensity versus complexity in the prioritization index and, in turn, which UL3 subgroups are prioritized.

When $\alpha = 0$, prioritization is driven entirely by clinical complexity. Highly complex patients ($C = 3$) experience waiting times of approximately 11 minutes minutes across all intensity levels, corresponding to reductions of approximately 91% relative to FCFS. Patients with moderate complexity ($C = 2$) experience waiting times of roughly 15-16 minutes, compared to more than 131 minutes under FCFS. These gains come at the expense of low-complexity patients ($C = 1$), whose waiting times only slightly increase relative to FCFS, reflecting a shift of delay toward less vulnerable profiles within UL3 under Priority-IC.

On the other side, when $\alpha = 1$, prioritization is driven exclusively by clinical intensity. High-intensity profiles ($I = 3$) experience waiting times of approximately 13-16 minutes, while low-intensity profiles ($I = 1$) incur substantially longer delays, exceeding 150 minutes, compared to 132-136 minutes under FCFS. Notably, under $\alpha = 1$, patients with high complexity but low intensity (i.e., along the non-prioritized dimension) are penalized more severely than patients with high-intensity but low-complexity do under $\alpha = 0$. As shown in subsection 4.2, this asymmetry arises from the interaction between prioritization decisions, estimated waiting-time sensitivity of service times, and the composition of arriving patient profiles. Accordingly, these results do not imply that clinical complexity should dominate intensity in prioritization decisions or vice versa. Rather, they illustrate how sequencing decisions that focus exclusively on a single dimension can interact with heterogeneous waiting-time

ARTICLE IN PRESS

sensitivity and composition of patient profiles in the arrival mix.

Intermediate values of α ($\alpha = 0.4$ and $\alpha = 0.7$) exhibit a different behavior. These configurations jointly account for both intensity and complexity in prioritization decisions. Patients with jointly moderate to high intensity and complexity–such as $(I = 2, C = 2)$, $(I = 3, C = 2)$, $(I = 2, C = 3)$ and $(I = 3, C = 3)$–experience between 11 and 18 minutes of waiting times, corresponding to reductions of 75-90% relative to FCFS. In addition, intermediate values of α mitigate the sharp penalties observed under extreme prioritization with $\alpha = 0$ or $\alpha = 1$ for patients who are vulnerable along the non-prioritized dimension. For instance, under $\alpha = 0$, moderate to highly intense but low complex patients, such as $(I = 3, C = 1)$ and $(I = 2, C = 1)$, wait on average more than 136 minutes, while under $\alpha = 0.4$ their waiting times decrease to approximately 20 and 49 minutes, respectively. Similarly, under $\alpha = 1$, highly complex but low-intensity patients experience long delays that are reduced when α takes intermediate values. Therefore, under intermediate configurations, delay is concentrated primarily on patients with both low intensity and low complexity $(I = 1, C = 1)$, whose average waiting times increase to nearly 175 minutes, compared to approximately 136 minutes under FCFS, reflecting a reallocation of delay toward patients with lower expected sensitivity to prolonged waiting.

Service time - results. The simulation also generates endogenous service times for UL3 patients, according to the empirically estimated dependence of service duration on realized waiting time and clinical profile. To examine how FCFS and Priority-IC affect service durations, we analyze the distribution of service times for UL3 patients stratified by intensity-complexity profiles. Figure 3 reports empirical cumulative distribution functions of service duration for UL3 patients under FCFS and Priority-IC, assuming a service capacity of $n = 8$, for different values of the parameter α, separately for each (I, C) and across all simulation replications.

For low-intensity, low-complexity patients $(I = 1, C = 1)$, service time distributions are similar across sequencing rules. Relative to FCFS, Priority-IC induces only a slight rightward shift, corresponding to marginal increase in service duration. Differences become more pronounced as either clinical complexity or clinical intensity increase. For patients with moderate to high complexity $(C = 2, C = 3)$ and low intensity $(I = 1)$, prioritization rules that emphasize complexity (i.e., $\alpha = 0$ or $\alpha = 0.4$) shift service-time distributions to the left relative to FCFS. For these patients, approximately 95% of services are completed within 67.5 minutes for $C = 2$ and 80 minutes for $C = 3$, compared to nearly 75 and 100 minutes under FCFS, respectively. In contrast, when intensity is prioritized ($\alpha = 1$) these same patients experience service time distributions that are comparable to–or slightly worse than–those observed under FCFS, reflecting longer realized waiting times when complexity is not prioritized.

A symmetric pattern emerges for patients with moderate to high intensity $(I = 2, I = 3)$ and low complexity $(C = 1)$. In this case, prioritization rules emphasizing intensity ($\alpha \geq 0.7$) reduce service times for these patient profiles relative to FCFS, while prioritization based on complexity alone yields distributions that resemble FCFS or exhibit modest rightward shifts. Therefore, under extreme values of α, patients who are vulnerable primarily along the non-prioritized dimension may experience service times that do not improve–and may even worsen–relative to FCFS benchmark, illustrating how prioritization decisions interact with heterogeneous waiting time-sensitivity to shape service times.

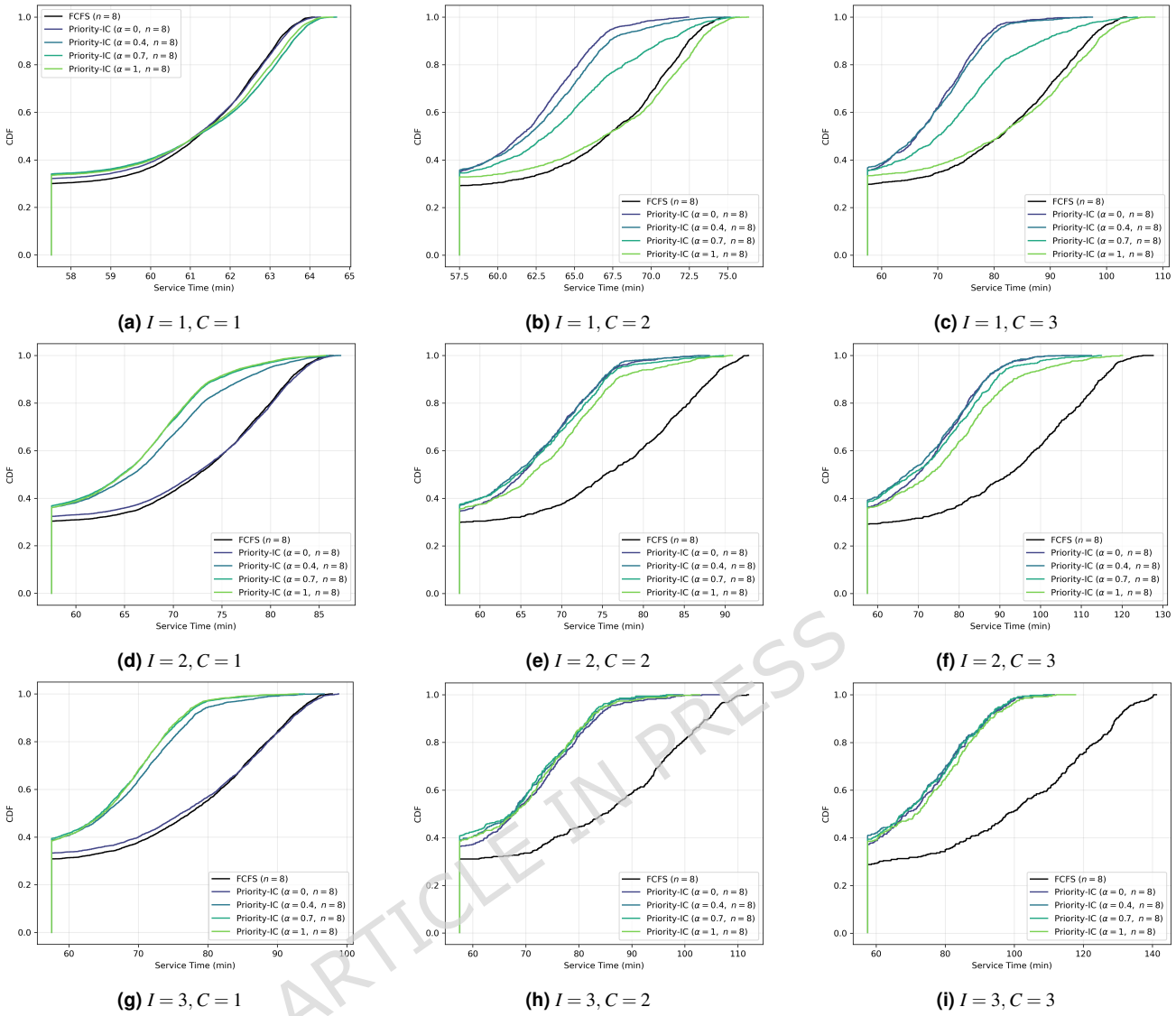
By contrast, intermediate values of α consistently outperform FCFS across nearly all intensity-complexity profiles. For patients with moderate to high levels of both intensity and complexity, service times are reduced by approximately 10-20 minutes for nearly 50% of patients. Moreover, patients who are vulnerable along only one dimension, such as those with high intensity but low complexity or vice versa, also experience reductions in service time relative to FCFS, an effect that does not arise under extreme configurations $\alpha = 0$ or $\alpha = 1$.

As for waiting time improvements, extreme values of α generate benefits for specific subgroups of UL3 patients, while leaving other delay-sensitive profiles exposed to prolonged delays. Intermediate values of α, instead, balance prioritization across both intensity and complexity dimensions, yielding reductions in service duration relative to FCFS for almost all UL3 profiles. The only exception is the low-intensity, low-complexity group $(I = 1, C = 1)$, for which service times increase slightly under intermediate configurations. However, the increase remains limited, reflecting the weak sensitivity of service time to waiting time for this UL3 profile. Although this group represents the largest share of arrivals (see Table 5), its limited delay sensitivity explains why intermediate values of α achieve the strongest overall performance reported in Table 1.

These findings clarify the role of intensity-complexity aware sequencing and quantify the potential benefits of considering both dimensions into second-stage prioritization. These results show that jointly accounting for clinical intensity and clinical complexity allows sequencing decisions to better adapt to heterogeneity in waiting-time sensitivity across patient profiles. In turn, they provide evidence for Priority-IC as a structured second-stage prioritization mechanism that complements triage decisions by guiding within-class sequencing using multiple clinically interpretable dimensions, rather than privileging a single one.

ARTICLE IN PRESS

ARTICLE IN PRESS



(a) $I = 1, C = 1$ **(b)** $I = 1, C = 2$ **(c)** $I = 1, C = 3$

(d) $I = 2, C = 1$ **(e)** $I = 2, C = 2$ **(f)** $I = 2, C = 3$

(g) $I = 3, C = 1$ **(h)** $I = 3, C = 2$ **(i)** $I = 3, C = 3$

Figure 3. Cumulative Distribution Function (CDF) of service time for patients within UL3 urgency class across intensity I and complexity C (resources $n = 8$).

3 Robustness and sensitivity analyses

The preceding results show that Priority-IC improves ED performance by reallocating delay away from UL3 profiles whose service duration is more sensitive to waiting. We therefore conduct additional analyses to assess whether this conclusion depends on key modeling choices. Specifically, we examine: i) the sensitivity of results to the estimated waiting time to service time elasticities in Appendix B; ii) a structured-only intensity counterfactual based on NEWS2; iii) a context-specific empirical dynamic-rank comparator that approximates observed non-FCFS sequencing in the studied ED; and iv) the robustness of Priority-IC to validation-calibrated LLM label misclassification.

Sensitivity of waiting time to service time elasticities. The first robustness check concerns the estimated relationship between waiting time and service time reported in Appendix B. In the simulation, delaying a UL3 patient with profile $(I = i, C = c)$ increases expected service duration according to the profile-specific elasticity $\beta_{i,c}$. Because this feedback mechanism is central to the operational interpretation of Priority-IC, we re-ran the simulation under three alternative specifications to investigate whether intermediate values of α still obtain better performance: a no-feedback case with $\beta_{i,c} = 0$, a conservative case in which all elasticities are halved, and an upper-bound case in which each elasticity is replaced by the upper limit of its 95% confidence interval.

Table 2 reports the resulting mean waiting time for UL3 patients under Priority-IC only, for each value of α, with $n = 8$.

ARTICLE IN PRESS

The results show that the absolute level of UL3 waiting time depends on the strength of the waiting time to service time feedback, as expected. However, whenever a positive feedback mechanism is present, the ranking of Priority-IC regimes is preserved: intermediate values of α, which jointly account for intensity and complexity, continue to outperform single-dimension prioritization based only on complexity ($\alpha = 0$) or only on intensity ($\alpha = 1$). Under halved elasticities, mean UL3 waiting time is lowest for $\alpha = 0.4$ and $\alpha = 0.7$, with values around 68.6 minutes, compared with 72.2 minutes for $\alpha = 0$ and 69.8 minutes for $\alpha = 1$. Under upper-bound elasticities, the same pattern is amplified: intermediate values yield approximately 112.4 minutes, compared with 139.9 minutes for $\alpha = 0$ and 120.2 minutes for $\alpha = 1$.

In the no-feedback case, differences across values of α become negligible. This is expected because setting $\beta_{i,c} = 0$ removes the mechanism through which prioritizing delay-sensitive profiles reduces downstream service-time inflation.

Table 2. Sensitivity of UL3 mean waiting time to the waiting time to service time elasticity. The table reports mean waiting time, in minutes, for UL3 patients under Priority-IC, by value of the prioritization parameter α, with $n = 8$.

Sensitivity	$\alpha = 0$	$\alpha = 0.4$	$\alpha = 0.7$	$\alpha = 1$
No feedback, $\beta_{i,c} = 0$	54.59	53.79	53.77	54.16
Halved sensitivity	72.19	68.64	68.68	69.84
Upper 95%	139.93	112.44	112.49	120.24

NEWS2-based intensity counterfactual. The second robustness analysis examines whether the intensity component of Priority-IC could be replaced by a structured vital-sign-based score. We therefore evaluate a counterfactual in which the LLM-derived intensity used by the prioritization rule is replaced by the mapped NEWS2-derived intensity score, while the LLM-derived complexity dimension is kept unchanged, so that the experiment isolates the effect of replacing the policy-observed intensity signal rather than changing the service-time-generating mechanism.

Formally, the reference profile determining service-time sensitivity remains (I^{LLM}, C^{LLM}), while the priority score is computed using (I^{NEWS2}, C^{LLM}).

Because the NEWS2-derived score assignes almost all UL3 patients to the lowest intensity level, it contributes almost no within-class variation to the priority index. Consequently, for $\alpha < 1$, the priority score is driven almost entirely by complexity; for $\alpha = 1$, all patients have nearly identical priority scores and the policy collapses to FCFS-like sequencing. This behavior is reflected in the simulation results. With $n = 8$, FCFS produces a mean UL3 waiting time of 136.34 minutes. Replacing LLM intensity with NEWS2 intensity yields mean UL3 waiting times of 112.7 minutes for $\alpha = 0$, 0.4, and 0.7, respectively. When $\alpha = 1$, the NEWS2-based policy produces a mean UL3 waiting time of 136.34 minutes, essentially identical to FCFS. By contrast, the full LLM-based Priority-IC policy achieves lower UL3 waiting times of 96.85 and 97.31 minutes for $\alpha = 0.4$ and $\alpha = 0.7$, respectively.

These results show that NEWS2-based intensity alone does not provide sufficient within-UL3 discrimination to support intensity-driven prioritization. Complexity still generates operational gains when included in the priority score, but replacing the LLM-derived intensity with NEWS2 removes the additional benefit obtained by jointly accounting for both clinical intensity and clinical complexity. The simulation therefore reinforces the validation results: NEWS2 is appropriate for detecting acute physiological instability, but it is not designed to stratify post-triage intensity within a single intermediate-acuity class.

Table 3. Simulation counterfactual replacing LLM-derived intensity with NEWS2-derived intensity in the priority score. Waiting times are reported in minutes for $n = 8$.

Policy	α	UL3 wait	All-class wait	UL3 service time
FCFS	–	136.34	113.78	66.62
Priority-IC, LLM	0.0	112.7	94.32	64.32
Priority-IC, LLM	0.4	96.85	81.14	62.25
Priority-IC, LLM	0.7	97.31	81.5	62.24
Priority-IC, LLM	1	101.89	85.31	62.87
Priority-IC, NEWS2 intensity	0.0	112.7	94.32	64.32
Priority-IC, NEWS2 intensity	0.4	112.63	101.07	64.32
Priority-IC, NEWS2 intensity	0.7	112.63	101.08	64.32
Priority-IC, NEWS2 intensity	1.0	136.38	113.81	66.62

Context-specific comparator: Empirical dynamic rank. The third analysis addresses the choice of FCFS as the main structural benchmark. Although FCFS provides a transparent policy-neutral comparator, real ED sequencing may deviate from strict arrival order. We therefore compare Priority-IC with the empirical dynamic-rank baseline which approximates

ARTICLE IN PRESS

the observed within-UL3 sequencing pattern in the historical data. This comparator should be interpreted as a data-driven approximation of local observed sequencing, not as a transferable scheduling rule. Its purpose is to assess whether the gains from Priority-IC persist when the comparator is allowed to deviate from strict FCFS in a manner consistent with the studied ED. Details on the empirical dynamic rank calibration are available in the Appendix C.

The observed within-UL3 sequencing is substantially non-FCFS but diffuse rather than deterministic. The probability that the selected patient is the first FCFS-ranked patient decreases sharply as the UL3 queue length increases: it is 59.7% when two UL3 patients are waiting, 32.5% for queues of length 3–5, 19.8% for queues of length 6–10, 15.6% for queues of length 11–20, and 17.6% for queues of length 21–50. This pattern indicates that informal prioritization is present in practice, but does not correspond to a transparent or reproducible clinical priority rule.

To evaluate whether this observed non-FCFS behavior already captures part of the operational gain delivered by Priority-IC, we compare three sequencing rules under the same simulated demand, capacity, and service-time model: FCFS, the empirical dynamic-rank comparator, and Priority-IC with $\alpha = 0.7$. The comparison is reported in Table 4. In addition to mean waiting times for UL3 and UL4 patients, we report the total waiting burden for UL3 and UL4 patients combined, measured in patient-minutes per replication. We also report the fraction of the FCFS-to-Priority-IC improvement that is already recovered by the empirical dynamic-rank comparator.

FCFS corresponds to 0% recovered by construction, while Priority-IC corresponds to 100%. As shown in Table 4, the empirical dynamic-rank comparator improves upon strict FCFS, reducing mean UL3 waiting time from 136.34 to 120.42 minutes and reducing the combined UL3+UL4 waiting burden from 13451 to 11903 patient-minutes per replication. This corresponds to 1548 patient-minutes saved relative to FCFS, or 41.4% of the total FCFS-to-Priority-IC improvement. Priority-IC with $\alpha = 0.7$ achieves a larger reduction, lowering mean UL3 waiting time to 97.35 minutes and reducing the combined UL3+UL4 waiting burden to 9714 patient-minutes per replication. Using UL3 mean waiting time rather than the combined UL3+UL4 waiting burden gives a similar recovered fraction of 40.8%. Similar outcomes are obtained when $\alpha = 0.4$.

These results suggest that the observed local sequencing behavior already departs from FCFS in a way that reduces waiting burden relative to a strict structural benchmark. However, Priority-IC retains a substantial additional improvement by replacing diffuse, context-specific sequencing patterns with an explicit rule based on clinical intensity and clinical complexity.

Table 4. Empirical non-FCFS comparator for $n = 8$. Waiting times are reported in minutes. The combined UL3+UL4 waiting burden is reported in patient-minutes per replication.

Policy	α	UL3 wait [min/patient]	UL4 wait [min/patient]	UL3+UL4 burden [patient-min/rep.]	Saving vs FCFS [patient-min/rep.]	Fraction recovered [%]
FCFS	–	136.34	170.29	13451	0.0	0.0%
Empirical dynamic rank	–	120.42	159.63	11903	1548	41.4%
Priority-IC	0.7	97.31	128.05	9714	3737	100.0%

Sensitivity to LLM misclassification. Finally, we assess the robustness of Priority-IC to LLM label misclassification. To this end, we calibrate a validation-based label-noise model from the physician-vs-LLM confusion matrices reported in Figure 1 and re-run the simulation allowing the prioritization rule to observe perturbed intensity-complexity profiles. The reference patient profile continues to determine the service-time response to waiting, while the noisy observed profile determines the Priority-IC sequencing score. The results, reported in Appendix E, show that the aggregate waiting-time gains of Priority-IC are retained under validation-calibrated label noise. However, the profile-level allocation of these gains changes: misclassification compresses the priority signal toward intermediate classes, promotes part of the large low-intensity, low-complexity group ($I = 1, C = 1$), and reduces the precision with which waiting-time reductions are directed toward profiles with higher service-time sensitivity. Thus, LLM-derived labels remain operationally useful for post-triage prioritization, but their misclassification structure attenuates the clinical targeting precision of the sequencing rule.

4 Methods

This study was conducted in accordance with all relevant guidelines and regulations, including the Declaration of Helsinki. The requirement for informed consent was waived by the data protection owner of the healthcare institution and the Ethics Committee of the University of Bergamo due to the retrospective nature of the study, the impracticability of obtaining individual consent given the large sample size, and the use of fully anonymized data. The study was reviewed and approved by the Research Integrity and Ethics Committee of the University of Bergamo.

4.1 LLM

LLM data used. The LLM-based inference procedure is implemented on a subset of two-years of ED admissions for patients assigned to the UL3 urgency class. Dataset records are filtered to ensure the availability of sufficient clinical information

ARTICLE IN PRESS

for reliable interpretation and validation. Because both clinical intensity and clinical complexity rely critically on contextual information embedded in free-text triage documentation, we restrict the analysis to patient records in which the brief anamnesis contains at least 100 characters. Shorter entries typically correspond to sparse clinical notes that do not meaningfully describe symptoms, comorbidities or care dependencies, and would therefore undermine the validity of downstream semantic inference. To limit the influence of extreme observations on subsequent simulation modeling, we additionally apply conservative outlier filtering based on observed service and waiting times. Specifically, we exclude admissions below the 5th percentile and above the 95th percentile of the service time, as well as observations above the 99th percentile of the waiting time. These exclusions eliminate atypical cases such as aborted visits or unusually prolonged stays driven by rare operational situations rather than routine clinical decision-making. The resulting dataset comprises approximately 20,000 admissions, which constitute the basis for the LLM inference and physician validation analysis.

LLM procedure. We first focus on preparing the free-text triage documentation for reliable downstream interpretation. Raw presenting complaints and brief anamnesis are initially processed using a rule-based normalization layer designed to remove non-informative artifacts and reduce variability across records. This normalization step includes the correction of frequent typographical errors and normalization of numeric expressions (e.g, different ways of reporting values or units). The objective is not semantic interpretation but the construction of a clean and consistent textual representation suitable for subsequent processing. A second, clinically critical preprocessing step addressed the pervasive use of acronyms and shorthand notation in triage documentation. ED texts often contain dense sequences of abbreviations that are locally understood by clinicians but may be inconsistently used across staff and potentially ambiguous for general-purpose language models. To mitigate this issue, we apply a systematic extraction procedure to identify recurrent acronyms and abbreviated expressions within the texts. These acronyms are then reviewed in collaboration with emergency physicians, who validate intended meanings and provide standardized long-form expansions. The validated expansions are subsequently reintegrated into the original free-text fields, replacing shorthand expressions with explicit clinical language, while preserving the surrounding context. This process yields a semantically enriched and more homogeneous textual representation that retains the original clinical content while making implicit information explicit. In turn, this improves interpretability, reduces ambiguity, and increases the robustness of downstream LLM-based inference.

Once standardized, the free-text triage documentation is combined with structured clinical variables including vital signs, trauma indicators and demographic characteristics such as age and sex. This combined representation defines the complete input provided to the LLM and reflects the information set typically available to clinicians during initial triage evaluation. An LLM-based inference layer is then applied to this input with the objective of producing assessments from heterogeneous clinical information. Specifically, for each patient record, the LLM is prompted to assign two ordinal assessments—clinical intensity and clinical complexity—based on explicit criteria provided in the prompt. These criteria specify the types of clinical evidence that should be considered relevant for each dimension and constrain the model to produce discrete, ordered outputs. Indicators of acute physiological derangement, abnormal vital signs, trauma and symptom severity primarily drive intensity, while features related to multimorbidity, functional limitations, and care dependencies are designated as the primary drivers of clinical complexity. This separation is intentional and reflects the purpose of the proposed framework. Both dimensions are made explicit to support downstream sequencing decisions among patients who share the same urgency class. By structuring these complementary aspects of patient health status, the framework enables a more systematic within-class prioritization.

Central to the inference layer is a domain-specific prompt that defines how the LLM interprets heterogeneous triage information into clinical intensity and complexity. The initial prompt structure was informed by established clinical deterioration scores and triage frameworks—such as NEWS2 for physiological instability and AVPU for level of consciousness—which provide reference points for identifying signs of acute risk at presentation. These sources are not used to replicate threshold-based scoring rules, but rather to anchor the prompt to clinically recognized indicators. Unlike threshold-based scores and in line with the proposed framework, the signals are interpreted in conjunction with contextual information contained in free-text documentation. Such combination is essential for structuring prioritization among patients with the same urgency class.

Prompt calibration is conducted through an iterative process with emergency physicians and triage nurses. In each iteration, the LLM is applied to a sample of triage records, and the resulting clinical intensity and clinical complexity assessments are independently reviewed by clinicians. Discrepancies between model outputs and clinical judgment are examined to identify systematic sources of misalignment, such as excessive reliance on specific symptoms, misinterpretation of stable chronic conditions, or failing to recognize functional limitations and care dependencies described only implicitly in the text. Based on the feedbacks, the prompt is refined to distinguish acute from chronic clinical signals, give greater salience to descriptions of functional status, autonomy, and care needs when assessing complexity, and better align model interpretations with real practices and clinical expectation. The final prompt, reported in full in Appendix A enforces a reasoning process that reflects triage judgment.

Following calibration, the finalized prompt is applied to the full corpus of filtered anonymized ED admissions. All computations are performed locally using the open-weight *gpt-oss-120b* model released by OpenAI on August 5, 2025. The

ARTICLE IN PRESS

model is a mixture-of-experts transformer with 117B total parameters, 5.1B active parameters per token, and native support for context lengths up to 128k tokens. In this study, inference is performed using a local quantized deployment of the model (MXFP4), with a maximum context window of 50,000 tokens. Decoding is deterministic, with temperature set to 0; no stochastic sampling was used, and no explicit random seed set because generation is performed under deterministic decoding. No identifiable data left the institutional computing environment at any stage.

The inference pipeline preserves all intermediate and final elements used in the analysis, including normalized texts, acronym expansions, prompt definitions, and the resulting patient-level intensity and complexity assignments. In turn, this allows the full inference process to be traced and ensures that the same procedure can be consistently applied or adapted to comparable ED datasets using the same specifications.

4.2 Simulation

Rationale of second-stage prioritization. The simulation framework is designed to examine how augmenting standard triage with a second-stage, intensity-complexity-based sequencing rule (Priority-IC) affects ED operations, while keeping the underlying triage categorization unchanged. The focus is on post-triage sequencing decisions within a single high-volume intermediate-acuity class (UL3), which typically accounts for the largest share of ED arrivals and therefore has substantial potential to affect overall patient flow.

Because the mechanism underlying Priority-IC acts on patient sequencing, waiting time is the primary operational quantity through which second-stage prioritization affects the system. Waiting time measures exposure to delay, is directly affected by the order in which patients are selected for service, and enters the simulation as the mechanism through which congestion can amplify downstream workload. In this sense, we waiting time as an operational outcome rather than as a direct measure of clinical harm. This choice reflects the structure of the sequencing problem: the simulation evaluates whether alternative sequencing rules reduce delay and congestion under identical demand and capacity conditions, but it does not claim that simulated waiting-time reductions translate one-for-one into clinical outcome improvements. At the same time, the clinical relevance of waiting time is supported by prior ED literature. Prolonged ED waiting time and crowding have been associated with adverse outcomes, including short-term mortality, subsequent hospital admission, morbidity, and delayed access to critical care[25–29].

Within this framework, each UL3 patient is characterized by two attributes—clinical intensity $I \in \{1,2,3\}$ and clinical complexity $C \in \{1,2,3\}$—whose empirical distributions are calibrated from the ED dataset using LLM-derived measures. Patients subject to Priority-IC (i.e., UL3 patients) are ranked according to a linear priority index: $S = \alpha I + (1-\alpha)C$ where $\alpha \in [0,1]$ controls the relative emphasis placed on acute intensity versus complexity. Higher values of S corresponds to higher service priority within UL3 class. When $\alpha = 0$ prioritization is driven exclusively by complexity: all patients with $C = 3$ are ranked ahead of those with $C = 2$, and all patients with $C = 2$ are ranked ahead of those with $C = 1$, regardless of intensity. Conversely, when $\alpha = 1$, prioritization depends only on intensity, and complexity plays no role. For intermediate values of α, priority reflects a weighted trade-off between the two dimensions, such that patients with jointly high intensity and high complexity are ranked ahead of those high in only one dimension, with the relative emphasis determined by the value of α. This distinction is operationally meaningful because intermediate values of α generate finer-grained sequencing than either single-dimension rule. For instance, under $\alpha = 0$, a patient with $(I = 1, C = 2)$ is prioritized equally to a patient with $(I = 3, C = 2)$, whereas under $\alpha = 0.4$ the latter receives higher priority. Therefore, intermediate values of α differentiate among clinically heterogeneous patients who would otherwise be grouped together under prioritization based on intensity or complexity alone.

Priority-IC governs only the sequencing of UL3 patients. Service times for these patients are modeled as endogenous and depend on realized waiting time and the associated intensity-complexity profile. Specifically, service duration increases with waiting time according to an empirically estimated regression model (see Appendix B), with the sensitivity of service time to waiting time varying across combinations of clinical complexity and clinical intensity and increasing with either dimension. This formulation captures the observation that prolonged waiting can increase downstream care requirements, and, in turn, affect system performance.

This modeling choice is grounded in both analytical and empirical evidence. Analytical queueing models have documented dependence between waiting time and service time——often formalized as delay-sensitive or congestion-dependent service times—particularly in settings with deteriorating arrivals (see[30,31]). Complementary empirical studies have shown that service times increase with time spent waiting beyond a reference point[30]. In healthcare settings, longer waiting times are associated with increased LOS[32], reflecting clinically intuitive mechanisms such as symptom progression, additional diagnostics or monitoring needs, and greater coordination effort once care is initiated.

The simulation therefore evaluates Priority-IC as a mechanism for structured redistribution of waiting times among UL3 patients relative to a neutral within-class discipline such as first-come-first-served (FCFS) in which UL3 patients are served strictly by arrival order. Under FCFS, UL3 patients are served strictly by arrival order and are therefore treated as operationally

ARTICLE IN PRESS

exchangeable. Patients with higher clinical intensity and/or higher clinical complexity may consequently experience waiting times comparable to those of lower-sensitivity profiles. Under Priority-IC, by contrast, sequencing decisions depend on clinical intensity, clinical complexity, or both, depending on the value of α.

In practice, Priority-IC modifies the dynamic of FCFS discipline. Under Priority-IC waiting time within the UL3 class is reallocated away from patients who are prioritized by the sequencing rule—whether based on clinical intensity, clinical complexity, or a combination of the two depending on the value of α—and toward patients with low intensity and low complexity. Relative to FCFS, prioritized patients therefore experience shorter waiting times, which in turn translates into shorter realized service durations, for profiles whose service times are sensitive to delay, according to the estimated regression model. Patients with low intensity and low complexity, by contrast, are not prioritized under Priority-IC and consequently absorb a larger share of waiting time. However, this group constitutes the majority of UL3 arrivals and exhibits weak sensitivity of service time to waiting time. As a result, even when their waiting time increases, the induced increase in their individual service times remains limited. Moreover, because prioritized patients represent a relatively small fraction of UL3 arrivals with respect to low complexity-intensity profiles, the reallocation of waiting time does not completely alter sequencing order, but instead produces moderate, profile-specific shifts in delay. Ultimately, this asymmetric reallocation of waiting time—conditioned on the chosen value of α—reduces waiting times for delay-sensitive profiles while inducing only negligible increases in service times for less sensitive ones. The net effect of the second-stage prioritization is a reduction in overall congestion and an improvement in patient flow relative to FCFS, under identical arrival processes and service capacity as documented in subsection 2.3.

Role of the prioritization parameter α. The effect of the prioritization parameter α on system performance depends on the interaction of three elements: i) the joint distribution of intensity-complexity profiles in the UL3 arrival stream, ii) the sensitivity of service time to waiting time for each profile, and iii) the congestion regime under which the ED operates, as determined by available service capacity relative to demand. In the simulation, the sensitivity of service time to delay is heterogeneous and non-linear across UL3 patients and increases either with higher clinical complexity or higher clinical intensity. At the same time, the empirical arrival mix of UL3 patients depicted in Table 5 is highly asymmetric across these dimensions. While the majority of UL3 patients exhibit low clinical complexity ($C = 1$, approximately 80%), clinical intensity is more evenly distributed.

This asymmetry suggests that the system-level effects of second-stage prioritization also depend on how sequencing decisions interact with the distribution of arriving patients across intensity-complexity profiles and their associated delay-sensitive service times. Postponing patients whose service times increase more sharply with waiting time leads to greater downstream workload accumulation than postponing patients whose service times are less delay-sensitive. When prioritization is driven exclusively by one dimension ($\alpha = 0$ or $\alpha = 1$), patients who are highly delay-sensitive along the non-prioritized dimension tend to accumulate substantial waiting time, which translates into inflated realized service durations and partially offsets the gains from prioritizing the emphasized dimension. The magnitude of this effect depends on the empirical arrival mix: highly complex patients are relatively rare but exhibit steep service-time inflation when delayed, whereas low-complexity but moderately to highly intense patients are more prevalent and still display non-negligible delay sensitivity.

By contrast, intermediate values of α distribute prioritization across both clinical intensity and clinical complexity. As a result, delay-sensitive patients along either dimension are less likely to experience prolonged waiting, reducing service time inflation across a broader set of profiles. Although intermediate prioritization still reallocates waiting time toward low-intensity, low-complexity UL3 patients, this group—while numerically dominant—exhibits weak sensitivity of service duration to waiting time. Consequently, the additional delay absorbed by these patients generates only limited increases in service time and does not offset the reductions achieved by more delay-sensitive profiles. In this sense, intermediate values of α align sequencing decisions more closely with the joint structure of arrival composition and delay sensitivity. Whether a particular value of α yields maximal benefits remains setting-specific and depends on both the arrival mix and the structure of waiting time sensitivity. More generally, α functions as a policy lever that governs how second-stage prioritization trades off concentration of delay across patient profiles against overall congestion reduction, subject to clinical and organizational constraints.

Intensity I	Complexity C		
	1	2	3
1	41	5	4
2	22	3	3
3	9	2	1

Table 5. Average number of patients by intensity–complexity (IC) profile.

Because delay-induced service time inflation becomes operationally relevant primarily under congestion, the magnitude of the benefits generated by Priority-IC also depends on service capacity. A sensitivity analysis varying the number of available

ARTICLE IN PRESS

resources, illustrating how congestion moderates the benefits of Priority-IC across values of α, is reported in Appendix D.

5 Discussion

This study shows how routinely collected triage information can be transformed into operational signals that support second-stage prioritization without altering existing triage systems. The analysis formalizes within-class prioritization as a distinct operational problem and demonstrates that explicitly accounting for patient heterogeneity along two clinically interpretable dimensions—clinical intensity and clinical complexity—can improve system performance under congestion. The framework operates after triage assignment and is explicitly designed to complement it. Clinical intensity and clinical complexity do not replace acuity-based classification; instead, they characterize differences among patients within the same urgency class, which are routinely treated as operationally interchangeable. However, differences in acute physiological instability, multimorbidity, functional vulnerability, and anticipated management burden imply that delays have heterogeneous operational consequences across patients within the same class. Making these heterogeneity explicit allows prioritization decisions to move beyond arrival order while preserving the foundational role of triage in risk stratification.

Within this framework, the role of the LLM is to infer representations of clinical intensity and clinical complexity from routine triage documentation. The LLM does not introduce new clinical information and does not automate triage decisions. Instead, it synthesizes information already available at first assessment——such as vital signs, presenting complaints, brief anamneses——into ordinal measures that capture complementary aspects of patient heterogeneity. Intensity captures the immediacy of physiological instability, while complexity reflects longitudinal vulnerability and anticipated management burden. The contribution is not to replace triage or clinician judgment, but to examine whether routinely documented clinical information can be transformed into relative prioritization signals after triage assignment.

The validation results should be interpreted in this context. The validation of LLM inferred clinical intensity and clinical complexity establishes note-based concordance with physician annotations, not clinical outcome validation. Both the LLM and the annotating physicians assess the same triage documentation, and the reported agreement measures the extent to which the LLM reproduces physician interpretation of those notes. The results indicate meaningful ordinal concordance, particularly for clinical complexity, with disagreements concentrated between adjacent categories and rare extreme misclassification. This supports the use of the inferred labels as documentation-based operational signals for second-stage prioritization. At the same time, this validation does not establish that intensity or complexity predict subsequent deterioration, unplanned escalation of care, return visits, adverse events, or adverse outcomes attributable to waiting. Such outcome-linked validation would be required before the framework could be considered for clinical deployment. Accordingly, the present study should be interpreted as an operational evaluation of structured within-class sequencing based on documented clinical heterogeneity, rather than as a clinical-risk prediction study.

The simulation analysis then illustrates the operational mechanism through which second-stage prioritization, Priority-IC, can improve ED performance under congestion. Under a neutral sequencing rule such as first-come–first-served, patients within the same triage category are treated as operationally interchangeable. Waiting time is therefore distributed uniformly within the class, independently of whether a patient is more sensitive to delay because of higher intensity, higher complexity, or both. However, when service duration increases with waiting time, delaying patients with greater waiting-time sensitivity can increase downstream workload and contribute to congestion. Second-stage prioritization based on clinical intensity and clinical complexity differentiates patients within the same urgency class. By structuring sequencing decisions around one or both dimensions, the prioritization rule reallocates waiting time away from patients for which delay has larger operational consequences and toward profiles with lower expected delay sensitivity. This redistribution explains why prioritization based on intensity or complexity alone improves performance relative to arrival-order sequencing, and why intermediate values of the prioritization parameter α perform more consistently: jointly accounting for both dimensions reduces the risk of delaying patients who are vulnerable along both dimensions.

The magnitude of these improvements depends on how the arrival-order benchmark is interpreted. FCFS is useful as a transparent structural baseline because it represents the case in which patients within the same triage class are treated as operationally equivalent. However, in real emergency departments, within-class sequencing is unlikely to be purely arrival-based. Clinicians may informally reprioritize patients on the basis of clinical gestalt, perceived vulnerability, symptom evolution, nursing reassessment, and local operational pressure. Such informal prioritization can attenuate the observed gap between FCFS and an explicit second-stage prioritization rule, because part of the benefit of moving beyond arrival order may already be embedded in routine practice but typically in a diffuse, context-specific, and non-transparent way.

The empirical dynamic-rank comparator helps quantify this attenuation. The historical data show that clinicians already deviate from strict arrival order, particularly when UL3 queues are longer. Consistently, the calibrated empirical comparator improves upon FCFS and recovers approximately 41% of the total FCFS-to-Priority-IC reduction in the combined UL3+UL4 waiting burden. However, the remaining improvement is still substantial: Priority-IC further reduces the combined waiting

ARTICLE IN PRESS

burden relative to the empirical dynamic-rank comparator, suggesting that observed informal prioritization captures only part of the operational value of structured within-class sequencing.

This result also has implications for implementation and equity. If informal within-class prioritization already occurs in practice, then an explicit rule such as Priority-IC should not be viewed only as a mechanism for improving efficiency, but also as a way to make sequencing criteria more transparent and auditable. At the same time, making these criteria explicit raises the question of which patient characteristics are systematically affected by the reallocation of waiting time. For this reason, we examined whether the clinical intensity and clinical complexity labels used by Priority-IC have demographic structure.

The age and sex-stratified validation analyses suggest that the equity implications of Priority-IC are more likely to arise through the complexity component than through the intensity component. Physician-labelled intensity shows no meaningful association with age (Spearman's $\rho = -0.019$, $p = 0.403$), whereas physician-labelled complexity has a moderate positive association with age (Spearman's $\rho = 0.293$, $p < 0.001$). This pattern is clinically interpretable because complexity captures multimorbidity, chronic disease burden, functional dependence, medication burden, loss of autonomy, and anticipated care-coordination needs, which are more frequent among older and clinically vulnerable patients. The sex-stratified analyses do not show meaningful separation between male and female patients.

These findings indicate that accounting for complexity may help avoid treating clinically heterogeneous patients within the same triage class as operationally interchangeable. At the same time, they also show why implementation should include equity monitoring. Complexity-related features may overlap with disability, socioeconomic vulnerability, limited social support, or barriers to accessing care. Therefore, future deployment-oriented evaluations should assess subgroup-specific waiting times, delay redistribution, and clinical outcomes in addition to aggregate congestion metrics. In this sense, the goal of Priority-IC is not to replace clinician judgment, but to make within-class prioritization more explicit, reproducible, clinically interpretable, and auditable.

These considerations clarify the scope of the present study and motivate the following limitations. First, the LLM-based pipeline depends on the quality and completeness of triage documentation. Although the pipeline is designed to handle heterogeneous language, abbreviations, and informal phrasing, systematic under-documentation of symptoms or longitudinal factors may affect the accuracy of inferred intensity and complexity in some settings. At the same time, the use of free-text information represents a key strength relative to approaches based solely on structured variables, as it enables the scalable incorporation of clinical nuance that is difficult to encode through predefined scores.

Second, the validation analysis assesses concordance with physician annotations based on the same triage notes. This is appropriate for evaluating whether the LLM reproduces physician interpretation of documented information, but it does not establish outcome-linked predictive validity. Future work should evaluate whether the inferred intensity and complexity measures are associated with clinically relevant outcomes, including deterioration, escalation of care, return visits, adverse events, and outcomes attributable to prolonged waiting. Such validation should ideally be conducted across multiple institutions and patient cohorts before any deployment-oriented use.

Third, the prioritization policies are evaluated within a stylized simulation framework intended to isolate the effects of within-class sequencing. The model abstracts from staff heterogeneity, diagnostic pathways, and institution-specific workflows. This limits direct generalizability to any single emergency department, but allows the underlying mechanisms—through which post-triage prioritization interacts with congestion, waiting times, and service dynamics—to be examined transparently. The relationships linking waiting time, service duration, clinical intensity, and clinical complexity are estimated from historical data and reflect average effects. The analysis does not account for adaptive clinician reasoning and behavior, dynamic reprioritization beyond the modeled rules, or learning effects that could arise under real-world deployment. Evaluating these dimensions, as well as broader outcomes such as equity, extreme delays, and patient-centered measures, remains an important direction for future work.

Overall, this study demonstrates the operational value of making patient heterogeneity within the same urgency class explicit after triage through measures of clinical intensity and clinical complexity, and of using these measures to support second-stage prioritization. By leveraging the potential of domain-adapted LLM and information already available at triage, the proposed framework points toward a direction for digital medicine in acute care that emphasizes operational enrichment rather than clinical replacement.

Declarations

Author Contributions

M.C., D.N.M. and C.M. jointly conceptualized and planned the study, developed the LLM-based pipeline and the simulation framework, conducted all computational analyses, including implementation and analysis of the simulation experiments, and drafted the manuscript. S.P. and F.C. acquired funding and supervised the project. F.M. contributed to conceptualization, supervised the clinical reasoning aspects, and supported prompt calibration. F.P., S.F. and P.S. performed clinician expert validation of LLM-derived outcomes. F.L. oversaw data acquisition and clinical data governance.

ARTICLE IN PRESS

Competing Interests

All authors declare no financial or non-financial competing interests.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available because they contain sensitive information and are subject to institutional restrictions imposed by the host hospital, but can be available from the authors on reasonable request.

Funding

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work was also supported by Regione Puglia under the PR FESR–FSE 2021–2027 Program (ASSI PRORITARI I e II – Azioni 1.1.3 e 1.10.8 – PIA), project code PIA001633. This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

ARTICLE IN PRESS

ARTICLE IN PRESS

References

1. Kuriyama, A., Urushidani, S. & Nakayama, T. Five-level emergency triage systems: variation in assessment of validity. *Emerg. Medicine J.* **34**, 703–710 (2017).
2. Zachariasse, J. M. *et al.* Performance of triage systems in emergency care: a systematic review and meta-analysis. *BMJ open* **9**, e026471 (2019).
3. Alkhatib, A., Aylin, P., Klaber, R. & Woodcock, T. Shifting patterns in emergency department attendance: a time series analysis. *Emerg. Medicine J.* (2025).
4. Bele, S. *et al.* Medical complexity in emergency and urgent care settings: a scoping review protocol. *BMJ open* **15**, e086984 (2025).
5. Ouyang, L., Yu, S., Hu, Z., Lin, Y. & Liu, D. Enhancing emergency department triage for older patients: a prospective study on the integration of the identification of seniors at risk. *BMC Emerg. Medicine* **25**, 91 (2025).
6. Mowbray, F. *et al.* Examining the relationship between triage acuity and frailty to inform the care of older emergency department patients: findings from a large canadian multisite cohort study. *Can. J. Emerg. Medicine* **22**, 74–81 (2020).
7. Blomaard, L. C. *et al.* Geriatric screening, triage urgency, and 30-day mortality in older emergency department patients. *J. Am. Geriatr. Soc.* **68**, 1755–1762 (2020).
8. Ding, Y., Park, E., Nagarajan, M. & Grafstein, E. Patient prioritization in emergency department triage systems: An empirical study of the canadian triage and acuity scale (ctas). *Manuf. & Serv. Oper. Manag.* **21**, 723–741 (2019).
9. Li, W., Sun, Z. & Hong, L. J. Who is next: Patient prioritization under emergency department blocking. *Oper. Res.* **71**, 821–842 (2023).
10. Shamout, F. E. *et al.* An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *NPJ digital medicine* **4**, 80 (2021).
11. King, Z. *et al.* Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digit. Medicine* **5**, 104 (2022).
12. Mitchell, R. *et al.* Multimodal learning for emergency department triage implementation: experiences from papua new guinea during the covid-19 pandemic. *The Lancet Reg. Heal. Pac.* **33** (2023).
13. Liu, T. *et al.* A foundational triage system for improving accuracy in moderate acuity level emergency classifications. *Commun. Medicine* **5**, 322 (2025).
14. Kurz, C. F., Merzhevich, T., Eskofier, B. M., Kather, J. N. & Gmeiner, B. Benchmarking vision-language models for diagnostics in emergency and critical care settings. *npj Digit. Medicine* **8**, 423 (2025).
15. Wallace, W. *et al.* The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ digital medicine* **5**, 118 (2022).
16. Kopka, M., von Kalckreuth, N. & Feufel, M. A. Accuracy of online symptom assessment applications, large language models, and laypeople for self–triage decisions. *npj Digit. Medicine* **8**, 178 (2025).
17. Brasen, C. L. *et al.* Machine learning in diagnostic support in medical emergency departments. *Sci. Reports* **14**, 17889 (2024).
18. Morlotti, C., Cattaneo, M., Paleari, S., Manelli, F. & Locati, F. The digitalization of emergency department triage: the perspectives of health professionals and patients. *BMC Heal. Serv. Res.* **24**, 1406 (2024).
19. Lee, J.-T., Hsieh, C.-C., Lin, C.-H., Lin, Y.-J. & Kao, C.-Y. Prediction of hospitalization using artificial intelligence for urgent patients in the emergency department. *Sci. reports* **11**, 19472 (2021).
20. Morlotti, C., Cattaneo, M., Paleari, F., Manelli, F. & Locati, F. Predicting emergency department revisits with hospitalization in a multihospital network: a large-scale assessment. *Eur. J. Emerg. Medicine* **32**, 141–143 (2025).
21. Sitthiprawiat, P., Wittayachamnankul, B., Sirikul, W. & Laohavisudhi, K. Development and internal validation of an ai-based emergency triage model for predicting critical outcomes in emergency department. *Sci. Reports* **15**, 31212 (2025).
22. Gisev, N., Bell, J. S. & Chen, T. F. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res. Soc. Adm. Pharm.* **9**, 330–338 (2013).
23. Gaber, F. *et al.* Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digit. Medicine* **8**, 263 (2025).

ARTICLE IN PRESS

ARTICLE IN PRESS

24. Guan, G., Lee, C. M. Y., Begg, S., Crombie, A. & Mnatzaganian, G. The use of early warning system scores in prehospital and emergency department settings to predict clinical deterioration: A systematic review and meta-analysis. *PLOS ONE* **17**, 1–16, DOI: 10.1371/journal.pone.0265559 (2022).

25. Gaughan, J., Kasteridis, P., Mason, A. & Street, A. Why are there long waits at english emergency departments? j. gaughan et al. *The Eur. J. Heal. Econ.* **21**, 209–218 (2020).

26. Marsilio, M., Roldan, E. T., Salmasi, L. & Villa, S. Operations management solutions to improve ed patient flows: evidence from the italian nhs. *BMC Heal. Serv. Res.* **22**, 974 (2022).

27. Guttmann, A., Schull, M. J., Vermeulen, M. J. & Stukel, T. A. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from ontario, canada. *Bmj* **342** (2011).

28. Richardson, D. B. Increase in patient mortality at 10 days associated with emergency department overcrowding. *Med. journal Aust.* **184**, 213–216 (2006).

29. Jones, S. *et al.* Association between delays to patient admission from the emergency department and all-cause 30-day mortality. *Emerg. Medicine J.* **39**, 168–173 (2022).

30. Webb, E., Yu, Q. & Bretthauer, K. Linking delay announcements, abandonment, and service time. *SSRN Electron. J.* (2019).

31. Wu, C. a., Bassamboo, A. & Perry, O. When service times depend on customers' delays: A relationship between two models of dependence. *Oper. Res.* **70**, 3345–3354 (2022).

32. Chan, C. W., Farias, V. F. & Escobar, G. J. The impact of delays on service times in the intensive care unit. *Manag. Sci.* **63**, 2049–2072 (2017).

ARTICLE IN PRESS

ARTICLE IN PRESS

A Prompt Specification

The box below reports the prompt used to guide the large language models toward structured, context-aware discrete clinical intensity and complexity assignments.

Role and Task

You are an expert clinical assistant trained in emergency department triage.
Your task is to assess a patient presenting to the emergency department by integrating:

– *structured clinical data (age, sex, vital signs, trauma, and chronic conditions, if reported);*

– *available information on functional or pathological status (e.g., chronic illness, functional dependence, pregnancy);*

– *structured, context-aware clinical inference, as performed by an experienced triage physician.*

Your objective is to assign:

– a ***Clinical Intensity Index*** (range 1–3);

– a ***Clinical Complexity Index*** (range 1–3).

Assignments must be based on realistic clinical judgment rather than numerical scoring rules.

Input Variables

The model receives the following structured fields:

– *Age, Sex*

– *Heart rate, Respiratory rate, Systolic blood pressure, Oxygen saturation, Body temperature*

– *Functional autonomy (Yes/No)*

– *Pregnancy status (Yes/No)*

– *Trauma (Yes/No)*

– *Reported symptoms (list of)*

– *Chronic conditions (list of)*

– *Acute exacerbation of chronic disease (Yes/No)*

– *Medications in use (list of)*

ARTICLE IN PRESS

Calibrated Definitions (1–3 Scale)

Clinical Intensity *reflects the perceived urgency and need for immediate medical attention.*

1: *Stable, non-urgent, compensated condition.*

2: *Relevant symptoms or mild-to-moderate abnormalities, overall compensated.*

3: *True medical urgency, acute instability, or immediate life-threatening risk.*

Clinical Complexity *reflects the overall difficulty of case management, independently of urgency.*

1: *Autonomous patient, low management burden.*

2: *Relevant comorbidities or partial dependence requiring careful management.*

3: *Severe disability, active multimorbidity, total dependence, or major cognitive impairment.*

Critical Inferential Rules

1. Do not react to keywords in isolation; always interpret symptoms in context.
2. Weight symptoms by contextual comorbidity-related risk.
3. Do not conflate the number of chronic conditions with acute severity.
4. In the absence of alarm symptoms, do not infer acute physiological instability from missing vital signs.[a]

ARTICLE IN PRESS

Output Format
Return the following JSON object:

```
{
  "Intensity": 1-3,
  "Complexity": 1-3
}
```

[a]This is consistent with our specific ED-context and local clinical practice.

B The relationship between service time and waiting time

As discussed in subsection 2.3 the simulation entails endogenous service times for UL3 patients which depend on realized waiting time and the intensity-complexity profile associated to the patient. To this end, we investigate the relationship between service time (ST) and waiting time (WT) by estimating the following econometric specification:

$$\ln(ST_i) = \alpha + \sum_{c=1}^{3}\sum_{k=1}^{3} \beta_{c,k} \left(\ln(WT_i) \times C_{c,i} \times I_{k,i}\right) + \varepsilon_i. \tag{1}$$

For each observation i, the coefficient $\beta_{c,k}$ captures the marginal effect of waiting time conditional on complexity level c and intensity level k, with $C, I \in \{1,2,3\}$ denoting increasing levels from low (1) to high (3).

Although waiting time is endogenously affected by sequencing policies in the simulation, the estimated relationship is treated as policy-invariant within the range of waiting times observed in the data. The purpose of this specification is not causal identification, but to represent how delays translate into downstream workload under realistic operational conditions, enabling the simulation to capture feedback effects between prioritization, waiting time exposure, and service-time inflation.

The regression is estimated on the full population of emergency department visits classified as urgency level 3 (UL3) over a two-year observation period. A logarithmic specification is adopted to allow for an elasticity-based interpretation of the estimated coefficients and to ensure that predicted values remain strictly positive. To mitigate the influence of outliers, both the service time and waiting time distributions are trimmed at the 95th percentile, and minimum service time is set at 10 minutes.

Table 6 reports the estimated relationship between service time and waiting time. The coefficients represent the conditional elasticity of service time with respect to waiting time and indicate that the effect of waiting time on service time increases monotonically with both complexity and intensity. Specifically, holding complexity constant, the elasticity of service time with respect to waiting time increases with intensity. A 10% increase in waiting time is associated with an increase in service time ranging from 2.2% to 8.9% for patients with complexity level 1 as intensity increases from level 1 to level 3. Similarly, for patients with complexity level 2 (3), a 10% increase in waiting time corresponds to an increase in service time ranging from 4.5% (9.7%) to 10.7% (14.7%) as intensity increases from level 1 to level 3.

As an illustrative example, Figure 4 shows the predicted service time across different combinations of intensity and complexity for a fixed waiting time of $WT = 120$ minutes. When both intensity and complexity are equal to 1, the predicted service time is approximately 62 minutes. This value increases along both dimensions, reaching approximately 112 minutes when both intensity and complexity are equal to 3.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 6. Relationship between service time and waiting time conditional on complexity and intensity

Variables	$\ln(ST)$
$C_1 \times I_1 \times \ln(WT)$	0.0233***
	(0.0047)
$C_1 \times I_2 \times \ln(WT)$	0.0675***
	(0.0053)
$C_1 \times I_3 \times \ln(WT)$	0.0889***
	(0.0066)
$C_2 \times I_1 \times \ln(WT)$	0.0452***
	(0.0084)
$C_2 \times I_2 \times \ln(WT)$	0.0795***
	(0.0084)
$C_2 \times I_3 \times \ln(WT)$	0.1066***
	(0.0109)
$C_3 \times I_1 \times \ln(WT)$	0.0971***
	(0.0118)
$C_3 \times I_2 \times \ln(WT)$	0.1269***
	(0.0086)
$C_3 \times I_3 \times \ln(WT)$	0.1469***
	(0.0111)
Constant	4.0170***
	(0.0153)
Observations	19,697

Standard errors in parentheses.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

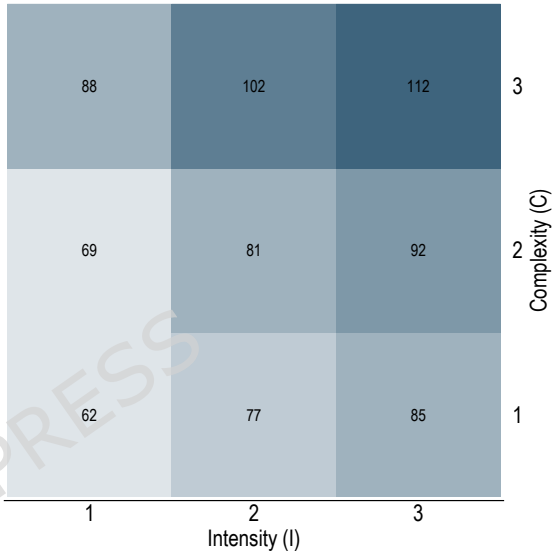


Figure 4. Predicted service time as a function of complexity and intensity for $WT = 120$ minutes

ARTICLE IN PRESS

ARTICLE IN PRESS

C Simulation Setup and procedure

The effectiveness of second-stage prioritization within UL3 urgency class is evaluated using a discrete-event simulation model, with results reported in subsection 2.3. The model represents patient flows in a single ED serving four urgency levels (UL1–UL4), capturing arrivals, queueing, service, and discharge processes under alternative within-class prioritization rules. The analysis compares a neutral within-class discipline (FCFS) with a second-stage prioritization rule (Priority-IC) while holding demand and service capacity constant. The simulation horizon corresponds to one operational day. To capture stochastic variability in arrivals we perform one-year of independent replications, which can be interpreted as daily realizations of ED activity. Each replication is preceded by a warm-up period of 3 days to mitigate bias from empty initial conditions and to ensure realistic queue states at the start of data collection.

Patient arrivals are modeled as non-homogeneous Poisson processes with hourly arrival rates varying by hour of the day and urgency level. For patients with urgency levels not subject to Priority-IC rule (UL1,UL2,UL4), arrivals are characterized solely by urgency and arrival time. In contrast, UL3 arrivals are additionally assigned an intensity-complexity profile, sampled from the empirical probability of LLM-derived intensity and complexity measures estimated on the two-year ED dataset.

Service capacity is represented by a fixed number of identical servers, interpreted as an aggregate abstraction of available clinical personnel and treatment space. The number of servers can also vary to evaluate sensitivity to capacity constraints (see Appendix D for further details). For urgency levels UL1, UL2, and UL4, service times are exogenously sampled and calibrated from empirical, hour-specific service-time distributions. For UL3 patients, service times are endogenously determined within the simulation as a function of the realized waiting time and the associated intensity-complexity profile. Specifically, service times are computed using the estimated regression model described in Appendix B, which captures the interaction between waiting time, the intensity-complexity patient profile, and their effect on service duration. This formulation allows waiting-induced service-time amplification to vary across patient profiles, with greater sensitivity for patients with higher intensity and higher complexity.

Patient sequencing is modeled through alternative within-class service disciplines. Under FCFS, patients are served in order of arrival within each urgency class, whereas under Priority-IC—applied only to UL3 patients—sequencing is determined by a prioritization index defined as a linear combination of clinical intensity and clinical complexity. The relative importance of the two dimensions is controlled by a parameter $\alpha \in [0,1]$, with lower values placing greater weight on complexity and higher values placing greater weight on intensity. For each replication, policies are evaluated under identical arrival streams, so that differences in outcomes reflect the sequencing rule rather than differences in realized demand.

Because FCFS is intended as a transparent structural benchmark rather than as a literal representation of observed ED practice, we also implement an empirical dynamic-rank comparator for UL3 sequencing. This comparator is designed to reproduce the degree of non-FCFS behavior observed in the historical data without imposing an explicit clinical priority rule. It therefore provides a context-specific baseline between strict arrival-order sequencing and the proposed Priority-IC rule. The empirical dynamik-rank comparator is calibrated from the historical within-UL3 service order. For each observed UL3 service start, we reconstruct the set of UL3 patients who had already arrived and had not yet been taken in charge, and compute the FCFS rank of the patient actually selected for service. A rank of 1 corresponds to strict FCFS, whereas ranks greater than 1 indicate that one or more earlier UL3 arrivals remain waiting. We then estimate the empirical distribution of this rank conditional on queue-length bins.

In the simulation, whenever service capacity becomes available and a UL3 patient is to be selected under the empirical dynamic rank comparator, the current UL3 queue is first ordered by arrival time. A FCFS rank is then sampled from the empirical distribution corresponding to the current queue length bin, and the patient at that rank is selected for service. The stochasticity of this rule reflects the modelling approach used to reproduce the observed distribution of service ranks.

All other model components are kept unchanged under the empirical dynamic rank comparator. UL3 patients still receive intensity-complexity profiles at arrival, and their service times remain endogenous functions of the true intensity-complexity profile and realized waiting time. Thus, the comparator isolates the effect of replacing strict FCFS with a data-driven approximation of observed local sequencing behavior, while preserving the same demand, capacity, profile assignment, and service-time mechanisms used in the main simulation.

Beyond the UL3 sequencing rules, the simulation also incorporates two additional mechanisms to better represent ED flow. First, UL1 patients (i.e., the most urgent) are allowed to preempt ongoing service. To account for operational frictions, preemption is not instantaneous, but a short delay of 10-15 minutes is introduced, during which service completion may occur naturally and direct preemption is avoided. When preemption occurs, the interrupted patient is assigned the highest non-preemptive priority upon re-entry into the queue. Second, we incorporate a fast-track mechanism for UL4 patients. With a calibrated probability estimated from ED data, UL4 patients bypass the ED queue entirely and are sent directly to the appropriate wards to receive care. The complete simulation procedure is summarized in the following algorithm:

ARTICLE IN PRESS

Algorithm 1 Discrete-event simulation with FCFS and Priority-IC prioritization

Require: Hourly arrival rates by urgency; service-time distributions; intensity–complexity distributions; regression coefficients; set of server levels N; prioritization weights $A = \{0, 0.4, 0.7, 1\}$

for replication $r = 1, \ldots, 365$ **do**
  Set random seed s_r
  for number of servers $n \in N$ **do**
    for sequencing policy $p \in \{\text{FCFS}, \text{Empirical Dynamic Rank}, \text{Priority-IC}\}$ **do**
      for $\alpha \in A$ **do**
        if ($p =$ FCFS or $p =$ Empirical Dynamic Rank) **and** $\alpha \neq 0$ **then**
          continue
        end if
        Run warm-up period (3 simulated days)
        while simulation time $< T$ **do**
          Generate patient arrivals by urgency
          if arrival urgency = UL 3 **then**
            Sample intensity–complexity profile
          end if
          if arrival urgency = UL4 and fast-track eligible **then**
            Bypass queue
          end if
          if arrival urgency = UL1 **then**
            Handle preemption
          end if
          Assign service according to policy p
          if patient urgency = UL 3 **then**
            Compute service time endogenously using regression model
          else
            Sample service time exogenously
          end if
        end while
        Record waiting times, service times, and queue statistics
      end for
    end for
  end for
end for

ARTICLE IN PRESS

ARTICLE IN PRESS

D Changes in resource availability

To examine the impact of service capacity—captured by the number of resources—on post-triage prioritization within the UL3 urgency class, we vary the number of available resources (n) while holding arrival processes, service parameters, and prioritization rules fixed, and compare system-wide average waiting times under FCFS and Priority-IC across values of the prioritization parameter α.

Table 7 reports average waiting times in minutes aggregated across all urgency levels and simulation replications, together with absolute and relative differences between Priority-IC and FCFS highlighting the interaction between congestion level, resource availability, and the magnitude of benefits achievable through within-class prioritization.

Resources n	α	Priority-IC wait [min]	FCFS wait [min]	Δ [min]	$\Delta_{\%}$ [%]
6	0.0	874.87	1178.4	-303.53	-25.76
6	0.4	579.92	1178.4	-598.48	-50.79
6	0.7	577.67	1178.4	-600.73	-50.98
6	1.0	667.28	1178.4	-511.12	-43.37
7	0.0	298.51	405	-106.48	-26.29
7	0.4	216.22	405	-188.77	-46.61
7	0.7	216.23	405	-188.76	-46.61
7	1.0	240.84	405	-164.15	-40.53
8	0.0	94.33	113.78	-19.45	-17.1
8	0.4	81.14	113.78	-32.63	-28.68
8	0.7	81.5	113.78	-32.28	-28.37
8	1.0	85.31	113.78	-28.46	-25.02
9	0.0	36.26	40.39	-4.14	-10.25
9	0.4	33.56	40.39	-6.83	-16.92
9	0.7	33.58	40.39	-6.82	-16.88
9	1.0	34.35	40.39	-6.04	-14.96
10	0.0	14.72	15.88	-1.16	-7.28
10	0.4	13.76	15.88	-2.12	-13.37
10	0.7	13.79	15.88	-2.09	-13.18
10	1.0	14.09	15.88	-1.79	-11.3

Table 7. System-wide average waiting time with absolute and relative difference of Priority-IC compared with FCFS for different n (resources) and prioritization parameter α.

When the number of available resources is low ($n = 6$ or $n = 7$) the system operates in a highly congested regime. Under this conditions, FCFS exhibits very large average waiting times, reflecting sustained queue accumulation and limited service capacity relative to demand. In this regime, Priority-IC yields substantial reductions in waiting times across all values of α, with relative improvements ranging from approximately 25% to nearly 50%, confirming that structured within-class prioritization provides substantial benefits when congestion is severe. As resource availability increases ($n \geq 8$), the system progressively transitions toward a less congested regime. In this case, average waiting time decreases under both FCFS and Priority-IC, and the absolute magnitude of improvement shrinks, depending on the value of α, ranging between 19-to-32 minutes at $n = 8$ and 2 minutes at $n = 10$. Nonetheless, Priority-IC continues to outperform FCFS across all values of α, with improvements that persist but become smaller, reflecting the reduced scope for delay-induced service-time amplification as congestion eases.

ARTICLE IN PRESS

ARTICLE IN PRESS

E Sensitivity to LLM Misclassification

The main simulation analysis assumes that the intensity and complexity labels used in Priority-IC are observed without error. Because these labels are inferred by the LLM pipeline, we conduct an additional sensitivity analysis to quantify how realistic LLM misclassification affects the operational results of post-triage prioritization.

Let $G = (I, C)$ denote the reference intensity-complexity profile of a UL3 patient. This reference profile determines the patient's service time sensitivity to waiting, denoted by β_G (see Appendix B). Let $H = (\tilde{I}, \tilde{C})$ denote the profile observed by the prioritization rule after LLM classification, which determines the priority score for sequencing. We calibrate a label-noise model from the physician-vs-LLM confusion matrices in Figure 1, which provide the marginal validation performance for intensity and complexity. Let $P_I(j \mid i)$ be the probability that a patient with physician-assigned intensity $I = i$ is assigned LLM intensity $\tilde{I} = j$, and let $P_C(k \mid c)$ be the analogous probability for complexity. For each reference profile $g = (i, c)$ and observed priority profile $h = (j, k)$ the noise model is expressed through the transition probability: $\Pr(\tilde{I} = j, \tilde{C} = k \mid I = i, C = c) = P_I(j \mid i)P_C(k \mid c) = \Pr(H = h \mid G = g) = Q_{g,h}$. The matrix $Q_{g,h}$ therefore maps reference profiles into the observed profiles used by Priority-IC.

In the noise-label experiments, each simulated UL3 patient retains the reference profile $G = (I, C)$ to determine the service-time response to waiting. Only the profile observed by the sequencing rule is perturbed. Therefore, Priority-IC computes the sequencing score using $S_\alpha(H) = \alpha\tilde{I} + (1 - \alpha)\tilde{C}$.

Table 8. Effect of label noise on priority-profile mass and service-time sensitivity

Profile (i,c)	$\pi^G_{i,c}$	$\pi^H_{i,c}$	$\pi^H_{i,c} - \pi^G_{i,c}$	$\beta_{i,c}$	$\mathbb{E}[\beta_G \mid H = (i,c)]$	$\Delta\beta_{i,c}$
(1,1)	0.456	0.210	-0.246	0.023	0.032	0.009
(1,2)	0.073	0.049	-0.024	0.045	0.051	0.005
(1,3)	0.044	0.014	-0.030	0.097	0.100	0.003
(2,1)	0.226	0.445	0.219	0.068	0.046	-0.021
(2,2)	0.036	0.105	0.069	0.080	0.063	-0.016
(2,3)	0.022	0.029	0.008	0.127	0.110	-0.016
(3,1)	0.114	0.113	-0.001	0.089	0.074	-0.015
(3,2)	0.018	0.027	0.008	0.107	0.089	-0.017
(3,3)	0.011	0.007	-0.003	0.147	0.132	-0.015

Let $\pi_g = \Pr(G = g)$ denote the reference case-mix distribution of intensity-complexity profiles. Applying the noise model, the distribution of observed priority profiles is $\tilde{\pi}_h = \Pr(H = h) = \sum_g \pi_g Q_{g,h}$. This quantity defines the fraction of patients observed by the priority rule as belonging to each intensity-complexity profile after LLM misclassification. As shown in Table 8 the reference distribution is concentrated in low-intensity and low-complexity patients. In particular, the reference profile $I = 1, C = 1$ accounts for approximately 45.6% of UL3 patients. After applying the validation-calibrated LLM noise model, the mass observed as $I = 1, C = 1$ decreases to approximately 21%, while the mass observed as $I = 2, C = 1$ increases to 44.5%. This shift is driven by the intensity confusion matrix in Figure 1a where 55.1% of physician-assigned $I = 1$ cases are classified as $I = 2$ by the LLM. The resulting effect is a compression of the observed priority-profile distribution toward intermediate labels. Low-priority reference profiles are often shifted upward, while some high-priority profiles are shifted downward or toward intermediate classes.

This comparison, however, does not by itself indicate whether the observed profiles remain clinically meaningful for prioritization. To assess this, we compute the posterior mean service-time sensitivity among patients observed in each profile h. For an observed profile h, the posterior probability that a patient belongs to the reference profile g is:$\Pr(G = g \mid H = h) = \frac{\pi_g Q_{g,h}}{\sum_{g'} \pi_{g'} Q_{g',h}}$. Therefore, the posterior mean service-time sensitivity of observed profile h is $\mathbb{E}[\beta_G \mid H = h] = \frac{\sum_g \pi_g Q_{g,h} \beta_g}{\sum_g \pi_g Q_{g,h}}$.

This quantity measures the average true service time sensitivity of the mixture of patients who are observed by the LLM as belonging to profile h. It should be interpreted as the average β_G of the reference profiles contained in that observed class.

For example, Table 8 shows that the observed class $\tilde{I} = 2, \tilde{C} = 1$ represents approximately 44.5% of patients after noise, but its posterior mean β is 0.046, which is lower than the reference value of $\beta_{2,1} = 0.0675$. This occurs because the observed class $(2, 1)$ contains many patients whose reference profile is $(1, 1)$, a low-intensity and low-complexity group with weaker service time sensitivity to waiting. Thus, label noise does not eliminate the clinical information contained in the LLM-derived labels but it attenuates the separation between observed priority classes and their true service time sensitivity, reducing the clinical targeting precision of Priority-IC.

We then run the simulation experiments under the noisy-label model. In these experiments, service time continues to depend on the reference profile $g = (I, C)$, while Priority-IC computes the sequencing score using the observed profile $h = (\tilde{I}, \tilde{C})$. The results, reported in Table 9, show that the aggregate headline waiting-time gains do not degrade under the validation-calibrated

ARTICLE IN PRESS

label-noise model. For UL3 patients, mean waiting time decreases from 136.34 minutes under FCFS to 96.85 minutes under clean Priority-IC and 94.03 minutes under noisy-label Priority-IC for $\alpha = 0.4$. For $\alpha = 0.7$, the corresponding values are 136.34, 97.31, and 93.48 minutes. Similarly, at the all-class level, noisy-label Priority-IC retains aggregate gains relative to FCFS.

Table 9. Average waiting-time performance under validation-calibrated label noise. Waiting times are reported in minutes

Outcome	α	FCFS	Priority-IC clean	Priority-IC noisy
UL3	0.4	136.34	96.85	94.03
UL3	0.7	136.34	97.31	93.48
All-class	0.4	113.78	81.14	79.04
All-class	0.7	113.78	81.50	78.71

The slightly lower aggregate waiting time under noisy labels is not an evidence that misclassification improves prioritization. Rather, this aggregate result reflects the interaction between the empirical case mix, nonlinear queueing dynamics, and the structure of the validation-calibrated noise. Since the UL3 population is concentrated in the low-intensity, low-complexity profile $(1,1)$, and since many of these patients are shifted upward to intermediate observed priority classes, this large group can experience shorter waits under noisy-label prioritization. This can preserve or slightly improve the aggregate mean even while the targeting of higher-sensitivity profiles worsens. For this reason, results should be interpreted together with the profile-level analysis. The relevant question is how label noise redistributes waiting time across reference intensity-complexity profiles.

To examine this redistribution, Figure 5 reports the profile-specific difference in mean waiting time between noisy-label and clean-label Priority-IC. Negative values indicate that patients with reference profile g wait less under noisy-label prioritization, whereas positive values indicate that they wait more.

The effect of label noise is heterogeneous across reference profiles. The low-intensity, low-complexity profile $(I = 1, C = 1)$ experiences a reduction in waiting time of approximately 42.7 minutes for $\alpha = 0.4$ and 47.3 minutes for $\alpha = 0.7$. This is consistent with the analytical results above: many reference $(1,1)$ patients are observed as intermediate-priority cases and are served earlier than they would be under clean label Priority-IC. Because $(1,1)$ is also the largest profile in the UL3 case mix, this reduction contributes largely to the aggregate mean waiting-time results.

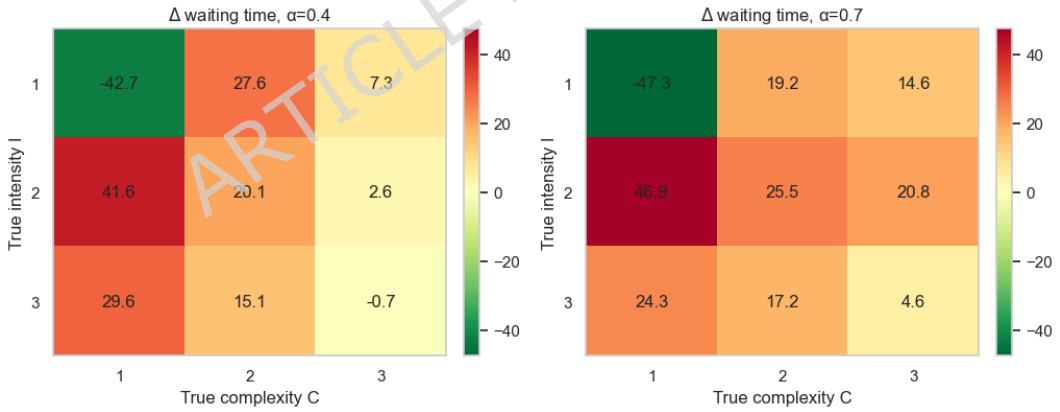


Figure 5. Difference in mean waiting time between noisy-label and clean-label Priority-IC by reference intensity-complexity profile.

At the same time, several profiles with higher service-time sensitivity wait longer under noisy-label prioritization. For $\alpha = 0.4$, the waiting time of $(I = 2, C = 1)$ increases by about 41.6 minutes, while that of $(I = 3, C = 1)$ increases by about 29.6 minutes. For $\alpha = 0.7$, the corresponding increases are about 46.9 and 24.3 minutes. In practice, noisy labels reduce the precision with which the sequencing rule identifies patients whose service time is more sensitive to waiting. Some higher-sensitivity patients are no longer consistently separated from lower-sensitivity patients in the priority ordering.

Overall, the noisy-label simulation results distinguish two effects. At the aggregate level, the headline waiting-time gains of Priority-IC relative to FCFS are retained. At the profile level, noisy labels reduce the precision with which waiting-time reductions are directed toward profiles with higher service-time sensitivity. In other words, the prioritization rule remains operationally robust in terms of mean waiting time, but misclassification attenuates the clinical precision of the priority signal.

ARTICLE IN PRESS

F Interpretable Estimation of Clinical Intensity and Complexity

LLMs represent the frontier of recent methodological innovation in clinical decision support. However, their deployment is often constrained by substantial computational requirements and does not offer clear visibility on how explanatory variables relate to core components such as, in our setting, clinical intensity and clinical complexity. Therefore, we introduce an interpretable, regression-based framework to estimate clinical intensity and clinical complexity, using the independent variables considered in the prompt (see Appendix A) as regressors. The purpose of this framework is interpretative rather than operational: all downstream analyses and simulations presented in the paper rely exclusively on the LLM-derived intensity and complexity measures.

In line with the definitions reported in subsection 2.1 and with medical decision making, clinical intensity reflects the current severity of the patient's condition and the urgency of medical intervention. It is modeled as a function of acute physiological instability, captured by out-of-range values of heart rate, respiratory rate, systolic blood pressure, oxygen saturation, and body temperature. In addition, the presence of trauma, symptom burden, altered mental status, and acute exacerbations of chronic disease may further increase intensity.

In contrast, clinical complexity captures the overall management burden independently of immediate acuity, reflecting underlying vulnerability rather than acute instability. Complexity is therefore modeled as a function of chronic disease burden, polypharmacy, and functional dependence. Demographics (sex and age) are considered for both components as control variables, while pregnancy status is included only in the clinical complexity model.[4]

To account for the ordinal nature of both outcomes, we employed generalized ordered logistic regression models on the same analytically restricted sample used for the estimation of the waiting time–service time relationship (see Appendix B). This approach relaxes the proportional odds assumption by allowing predictors to exert threshold-specific effects when supported by the data, thereby accommodating the non-linear and stage-dependent structure of clinical categorization decisions.[5]

The generalized ordered logit estimates two cumulative logit equations. The first equation models the log-odds of transitioning from level 1 to at least level 2, while the second equation models the log-odds of attaining the highest level (level 3). In other words, being $Y \in \{1,2,3\}$ an ordinal clinical outcome, where Y represents either clinical intensity (I) or clinical complexity (C), the following cumulative logit equations are specified:

$$\log \frac{P(Y \geq 2)}{P(Y < 2)} = \kappa_1 + \beta_1^\top \mathbf{X}_Y, \tag{2}$$

$$\log \frac{P(Y = 3)}{P(Y < 3)} = \kappa_2 + \beta_2^\top \mathbf{X}_Y, \tag{3}$$

where coefficients are allowed to vary across thresholds when the proportional odds assumption is violated.

For clinical intensity ($Y = I$), the covariate vector $\mathbf{X}_I$ includes acute physiological and symptomatic features, namely vital sign abnormalities (heart rate, systolic blood pressure, body temperature, and oxygen saturation), altered mental status, acute trauma, symptom burden, and demographic covariates. The definition of 'out of range' measures was derived from the NEWS2 (National Early Warning Score 2)[6] scales. For clinical complexity ($Y = C$), the covariate vector $\mathbf{X}_C$ includes variables reflecting chronic disease burden and management difficulty, including multimorbidity, polypharmacy, functional dependence, acute exacerbation of chronic conditions, pregnancy status, and demographic covariates.

Table 10 reports the generalized ordered logit estimates for both components.

For clinical intensity, acute physiological and symptomatic variables are the primary drivers of escalation across thresholds. Altered mental status is strongly associated with exceeding low intensity (Threshold 1), but does not exhibit a statistically significant independent association with reaching the highest intensity level (Threshold 2), indicating that confusion alone is sufficient to prompt escalation from low intensity but not to justify high levels of acuity.

Trauma exhibits a strong positive association with the transition from intensity level 1 to higher levels, coupled with a negative association with the transition to level 3. This is consistent with the differentiation between minor and major trauma presentations.

Vital signs that are out-of-range are strong predictors of intensity escalation. Heart rate and systolic blood pressure out-of-range values exert particularly large effects at Threshold 1, highlighting their role in triggering escalation from low intensity, while oxygen saturation and body temperature show stable and sizable effects across both thresholds. The number of reported symptoms is positively associated with escalation across both thresholds, indicating that symptom burden contributes to intensity assessment without acting as a sole determinant of maximal urgency.

[4] In the sample, pregnancy was rare (3% of occurrences) and observed exclusively among patients classified in the highest intensity category, resulting in quasi-complete separation and unstable estimates. Accordingly, it was not included as a predictor in the clinical intensity model.

[5] We automatically tested the parallel-lines assumption, and proportionality constraints were imposed only when statistically justified. Predictors exhibiting differential effects across outcome thresholds were therefore allowed to vary freely.

[6] https://www.rcp.ac.uk/improving-care/resources/national-early-warning-score-news-2/

ARTICLE IN PRESS

For clinical complexity, each additional chronic condition substantially increases the odds of both exceeding low complexity (Threshold 1) and reaching the highest complexity level (Threshold 2). Functional dependence exerts a strong and uniform effect across both thresholds, indicating that the loss of autonomy robustly defines higher clinical complexity.

Polypharmacy independently increases clinical complexity, with positive and significant effects at both thresholds. Acute exacerbation displays a threshold-dependent pattern, reducing the likelihood of transitioning from low to moderate complexity (Threshold 1) while substantially increasing the probability of reaching the highest complexity level (Threshold 2), suggesting that exacerbations contribute to complexity primarily when superimposed on already complex clinical profiles.

Demographic and contextual factors show modest and threshold-dependent associations across both outcomes. For clinical complexity, age contributes limited additional information once multimorbidity and functional dependence are accounted for, while female sex is associated with lower odds of higher complexity. Pregnancy increases the likelihood of transitioning from low to moderate complexity but does not substantially affect the probability of reaching maximal complexity. For clinical intensity, age is positively associated with escalation across thresholds, whereas female sex is associated with higher odds of intensity escalation, conditional on observed clinical features.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 10. Generalized ordered logit estimates for clinical intensity and clinical complexity

	Clinical Intensity		**Clinical Complexity**	
	Threshold 1	Threshold 2	Threshold 1	Threshold 2
Sex (Female)	0.1315***	0.2316***	-0.1657***	-0.1657***
	(0.0322)	(0.0408)	(0.0560)	(0.0560)
Age	0.0317***	0.0080***	-0.0358***	0.0055*
	(0.0009)	(0.0010)	(0.0021)	(0.0030)
Pregnancy			10.5487***	0.5622
			(1.0033)	(0.4360)
Altered mental status	2.1418***	-0.1855		
	(0.2946)	(0.2466)		
No. of symptoms	0.3659***	0.2440***		
	(0.0119)	(0.0141)		
Trauma	3.0609***	-0.6828***		
	(0.0735)	(0.0870)		
Heart rate out-of-range	1.5415***	1.5415***		
	(0.1945)	(0.1945)		
Systolic blood pressure out-of-range	3.6115***	0.8463***		
	(0.7247)	(0.2890)		
Oxygen saturation out-of-range	2.2937***	2.2937***		
	(0.2639)	(0.2639)		
Body temperature out-of-range	1.3408***	1.3408***		
	(0.2612)	(0.2612)		
No. of chronic diseases			3.1594***	2.4796***
			(0.0566)	(0.0593)
No. of medications in use			0.3195***	0.1864***
			(0.0259)	(0.0246)
Functional dependence			3.7830***	3.7830***
			(0.1857)	(0.1857)
Acute exacerbation			-0.2172**	0.4531***
			(0.0891)	(0.1008)
Constant	-2.8143***	-2.7914***	-3.0427***	-7.0492***
	(0.0548)	(0.0623)	(0.0906)	(0.2047)
Observations	19,608		19,697	
LL	-15,693		-14,252	

Notes: Threshold 1 models the log-odds of exceeding level 1 (i.e., outcome ≥ 2), while Threshold 2 models the log-odds of attaining the highest level (level 3). Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

ARTICLE IN PRESS