Detecting Human Bias in Emergency Triage Using LLMs: Literature Review, Preliminary Study, and Experimental Plan

Marta Avalos, Dalia Cohen, Dylan Russon, Melissa Davids, Océane Dorémus, Gabrielle Chenais, Eric Tellier, Cédric Gil-Jardiné, Emmanuel Lagarde

▶ **To cite this version:**

Marta Avalos, Dalia Cohen, Dylan Russon, Melissa Davids, Océane Dorémus, et al.. Detecting Human Bias in Emergency Triage Using LLMs: Literature Review, Preliminary Study, and Experimental Plan. FLAIRS 2024 - 37th International Florida Artificial Intelligence Research Society Conference, The Florida Artificial Intelligence Research Society, May 2024, Miramar Beach, United States. pp.6. hal-04575557

HAL Id: hal-04575557
https://inria.hal.science/hal-04575557v1

Submitted on 15 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Detecting Human Bias in Emergency Triage Using LLMs: Literature Review, Preliminary Study, and Experimental Plan

Marta Avalos-Fernandez[1,2] Dalia Cohen[1], Dylan Russon[1], Melissa Davids[1], Océane Doremus[1], Gabrielle Chenais[1], Eric Tellier[1], Cédric Gil-Jardiné[1,3], Emmanuel Lagarde[1]

[1]University of Bordeaux, Bordeaux Population Health Research Center, UMR U1219, INSERM, F-33000, Bordeaux, France
[2]SISTM team, Inria centre at the University of Bordeaux, F-33405, Talence, France
[3]University Hospital of Bordeaux, Pole of Emergency Medicine, F-33000, Bordeaux, France
{first name.last name}@etu.u-bordeaux.fr, {first name.last name}@u-bordeaux.fr

Abstract

The surge in AI-based research for emergency healthcare poses challenges such as data protection compliance and the risk of exacerbating health inequalities. Human biases in demographic data used to train AI systems may indeed be replicated. Yet, AI also offers a chance for a paradigm shift, acting as a tool to counteract human biases. Our study focuses on emergency triage, rapidly categorizing patients by severity upon arrival. Objectives include conducting a literature review to identify potential human biases in triage and presenting a preliminary study. This involves a qualitative survey to complement the review on factors influencing triage scores. Moreover, we analyze triage data descriptively and pilot AI-driven triage using an LLM with data from the local hospital. Finally, assembling these pieces, we outline an experimental plan to assess AI's effectiveness in detecting human biases in triage data.

Introduction

There is an increased interest in harnessing artificial intelligence (AI) for healthcare, especially in emergency departments (EDs). Emergency medicine requires efficient organization, coordination, and rapid decision-making for high-acuity patients, making AI a promising solution (Taylor et al. 2022; Piliuk and Tomforde 2023; Chenais, Lagarde, and Gil-Jardine 2023). Proposals for applications in various tasks show potential to improve emergency healthcare services, including prehospital settings (Rosemarin, Rosenfeld, and Kraus 2019; Lee and Lee 2020), emergency medical dispatch (Emami and K. 2023), patient flow management (Liventsev, Härmä, and Petković 2021; Arnaud et al. 2022), and emergency triage (Yu et al. 2022; Kipourgos et al. 2022; Sanchez-Salmeron et al. 2022; Cho et al. 2022; Vantu, Vasilescu, and Baicoianu 2023; Defilippo et al. 2023; Mutegeki et al. 2023; Sax et al. 2023; Gao et al. 2022), with a focus on natural language processing (NLP) applications using large language models (LLMs) (Stewart et al. 2023). While integrating AI into EDs brings benefits, challenges must be navigated. Ethical and legal considerations arise in ensuring patient privacy protection, compliance with regulations (van der Stigchel et al. 2023), and addressing potential bias in AI algorithms to prevent disparities in patient care.

A significant example of health disparity pertains to sex/gender. Biological (sex) and socio-cultural (gender) factors play a key role in chronic diseases and physiological processes, including pain sensitivity, leading to observable differences in disease epidemiology. These variations are attributable to differences in clinical conditions, onset, symptom presentation, prognosis, biomarkers, and treatment effectiveness. However, the evaluation, diagnosis, and treatment of health conditions may also be influenced by unconscious cognitive biases among health professionals and social stereotypes. Beyond human bias, the implications of bias in AI for health applications are becoming increasingly concerning. Studies show that LLMs exhibit biases aligned with stereotypical roles when trained on data with under-representation of women, such as clinical data (Kotek, Dockum, and Sun 2023; Buslon et al. 2023). Consequently, these models reinforce biases in line with societal perceptions and often overlook ambiguities in sentence structure, providing inaccurate explanations for their biased choices.

Due to the risk of AI models inheriting biases from their training data and potentially exacerbating health disparities, we conducted an extensive literature review to identify potential human biases in triage. To complement this review and gain insights into practices, we conducted a qualitative survey. Additionally, we performed a descriptive analysis of triage data using demographic factors such as patient's sex and age, along with nurse's sex (the only permissible demographic information), sourced from University Hospital of Bordeaux. We then implemented a pilot AI-driven triage using an LLM model. Finally, based on all these elements, we outlined an innovative experimental plan to assess the effectiveness of LLMs in detecting biases in triage data.

Emergency Triage

Upon arrival at the ED, a triage nurse swiftly assesses patients, documenting vital information like the reason for the visit, vital signs, and medical history in free-form text notes. The nurse relies on specific tools and protocols to enable standardized sorting. Accurate triage scoring is vital: underestimating urgency can delay care and worsen outcomes, while overestimating it may lead to excessive resource use and higher costs. Numerous studies have evaluated differ-

Copyright © 2024 by the authors.
This open access article is published under the Creative Commons Attribution-NonCommercial 4.0 International License.

ent scales (Aubrion et al. 2022), with results ranging from moderate to good validity, but no scale has emerged as the definitive gold standard.

Triage grids assisting triage nurses in categorizing patients include the widely used Emergency Severity Index (ESI) in US hospitals, the Canadian Emergency Department Triage and Acuity Scales (CTAS), the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS), the South Africa Triage Scale (SATS) for developing countries, and the FRench Emergency Nurses Classification at Hospital (FRENCH) triage scale. Though each has its specificities, all aim to objectively define case severity, care complexity, and required resources, resulting in prioritization and a maximum waiting time or number of examinations. Variables determining the triage score include common ones like vital signs and less-consensus factors like age, pain level, or medical history. Each healthcare facility selects a validated, reliable, and reproducible scale or triage grid with 4 or 5 levels, tailored to national healthcare characteristics. Recommendations on triage grids come from organizations like the French Society of Emergency Medicine[1] or the American College of Emergency Physicians[2].

Human Bias in Emergency Triage

We performed a non-systematic search in Medline/PubMed, focusing on titles or abstracts. Our search encompassed broad terms such as 'bias' and 'emergency', alongside specific terms like 'gender' and 'triage', including their synonyms and related terms. Furthermore, we employed snowballing techniques traversing through the reference lists of identified papers. We synthesize following our findings.

External factors, like ED location, time, and arrival day, can influence triage decisions and the correspondence between assigned scores and subsequent interventions (Gorick 2022; Suamchaiyaphum, Jones, and Markaki 2023). Legal concerns about triage errors, especially complications due to under-triage, which may result in the death of a patient (Hinson et al. 2019), may prompt triage nurses to assign a more urgent score when uncertain.

Studies indicate significant differences in prioritization based on patient age, ethnicity, and health insurance coverage (Portillo et al. 2023; Peitzman et al. 2023; Essa et al. 2023; Martin et al. 2023; Fekonja et al. 2023). Young individuals, African-American or Hispanic patients, and those from economically disadvantaged neighborhoods face a higher risk of triage errors (Banco et al. 2022). Unexpected behaviors are observed, with over 10% of consultations not following arrival order, prioritizing older individuals and deprioritizing racialized individuals, even within the same triage level where 'first come, first served' is the supposed principle (Lin et al. 2022).

The findings regarding sex/gender are inconsistent (Arslanian-Engoren 2000; Onal et al. 2022). Some studies suggest that elderly individuals and patients with severe symptoms or high risk are treated with equal urgency regardless of sex. However, other research indicates potential disadvantages for women at various care stages. For instance, women may experience longer visit durations, waiting times before treatment, or consultations with healthcare professionals. Additionally, studies on low-risk patients for acute coronary syndrome suggest that women are hospitalized less frequently and undergo fewer tests compared to men. Interestingly, these disparities may inadvertently benefit women by avoiding unnecessary hospitalizations or cardiac tests. In contrast, a recent study (Coisy et al. 2023) showed that altering the visualization of simulated patients with diverse characteristics impacted prioritization decisions in a standardized clinical case involving chest pain. The results showed disparities in emergency treatment, with black and female patients less likely to receive prompt care.

Interpreting gender disparities is complex due to women often presenting with "atypical" symptoms for serious conditions like strokes, heart attacks, appendicitis, or acute poisonings from substances other than opioids (Preciado et al. 2021; Mnatzaganian et al. 2020; Lopez et al. 2021). This could lead to underestimations of urgency and delays in diagnosis. Clinical studies are predominantly developed using male models, further complicating matters. Additionally, women may have more heterogeneous and less precise symptoms, along with being more prone to chronic pain. Several biopsychosocial mechanisms contribute to their greater pain sensitivity and less responsive reactions to pain treatments compared to men. However, social models related to gender differences also influence pain expression.

Research suggests that biases in triaging patients are influenced by the gender of the triage nurse rather than the patient's gender (Vigil et al. 2017). Female patients receive similar triage regardless of the nurse's gender, while male patients may receive different scores based on the nurse's gender. Female nurses tend to assign higher urgency scores to male patients with higher pain levels, whereas male nurses assign lower scores. This trend is more evident in emergency department settings with predominantly female nursing staff. The perception that men are more prone to panic and exaggeration of symptoms, while women are seen as calmer and more stoic, contributes to these differences in triage decisions.

Preliminary study

We conducted a preliminary study by interviewing triage nurses from three hospitals to understand triage practices better. Then, we analyzed triage data from the University Hospital of Bordeaux, focusing on factors influencing triage scores using insights from literature and interviews, where available and authorized. Finally, we used an LLM to predict triage scores, following the University Hospital of Bordeaux Research Ethics Committee's guidelines.

Methods

Interview-based qualitative study. The study recruited participants from the ED of the University Hospital of Bordeaux and the Saint-Antoine AP-HP and Lariboisière AP-HP hospitals in Paris. Eligible participants were nurses who

[1] www.sfmu.org/upload/referentielsSFMU/ioa2004.pdf

[2] www.acep.org/siteassets/uploads/uploaded-files/acep/clinical-and-practice-management/resources/administration/triagescaleip.pdf

had undergone triage training, possessed prior experience in triage, could provide an oral account of their experiences, and committed to participating in hour-long interview sessions. We conducted individual semi-structured interviews and recorded with a voice recorder following the nurse's consent. We utilized a comprehensive interview guide, developed through an extensive literature review, to ensure all relevant topics were covered. These included patient demographics, contextual factors around their visit, characteristics of the triage nurse, and broader institutional issues. The interviews were transcribed verbatim using transcription software, and a thematic analysis was applied to identify recurring patterns and themes across the interviews. A total of 10 triage nurses (8 females and 2 males) participated in the study, with experience ranging from 4 to 25 years. Interviews were conducted between May 17th and May 31st, 2023, with an average duration of 73 minutes per session.

Data analysis. The database comprises complete records of adult ED visits (ages 15 and older) at the University Hospital of Bordeaux from January 2013 to December 2021, totaling 480,001 visits. Variables analyzed included month/year of visit, patient's sex (M/F), patient's age, triage nurse's sex, and assigned triage score. Entries with missing patient identifiers or variables were excluded, and visits before January 2016 were filtered out due to a new triage protocol, making scores before this period incomparable. The resulting dataset comprised 273,151 visits. Potential selection bias sources were examined by comparing data with and without triage scores. Bivariate analyses assessed the association between triage score and other variables. Multivariate analyses explored the association between triage score and patient sex, considering patient age and triage nurse's sex via multinomial logistic regression. We checked age linearity on the logit, explored goodness-of-fit measures, and tested the interaction between patient sex and triage nurse's sex. Statistical significance was set at $p<0.05$.

LLMs to predict triage score. The study examined ED patient visits at the University Hospital of Bordeaux from 2013 to 2020, totaling 296,071 visits, after excluding those lacking triage level or clinical notes. All variables collected by triage nurses were analyzed, including structured data like vital signs, chief complaint, age, sex, pain, nausea, alcohol level, and visit timestamp, along with unstructured data from clinical notes. The outcome variable was the triage score. These variables were analyzed using XGBoost and LightGBM classifiers combined with TF-IDF vectorization, as well as a clinical BELGPT-2 model trained on French corpus and clinical notes. Model performance was evaluated based on precision, recall, and micro F1 score metrics.

Results

Interview-based qualitative study. Interview questions explored the impact of patient characteristics (sex/gender, ethnicity, age, accompaniment), nurse characteristics (sex/gender, experience, individual behavior, fatigue level), triage grid elements (primary complaint/symptom, vital signs, circumstances, medical history, reported/observed pain), and contextual factors (ED saturation status, Covid-19 status, arrival time and day, lack of general practitioners, ED location, mode of arrival) on triage nurse practice. Interviews were transcribed and coded using Taguette software, and tags were derived from pre-defined themes and those emerging from the interviews.

Most nurses acknowledged the stereotype that men are more "sensitive," "whiny," or "in pain" than women, but they did not necessarily agree with it. Even if men were perceived as "complaining" more or being more "dramatic," ultimately, they were thought to rate their pain similarly to women. Only one nurse noted a potential difference in empathy among female nurses towards patients with menstrual pain. Age emerged as a universal factor in triage decisions. During busy periods, nurses might over-triage elderly patients ("it is not doing them any favors to make them wait in the hallway under neon lights for 4 to 5 hours."). Elderly patients were thought to "inevitably have risk factors," have a higher risk of wandering off, have "been through wars", and "be more resistant to pain."

Cultural influences were also recognized as playing a role in triage. Nurses perceived certain populations as more predisposed to specific health conditions, and also reported paying closer attention to some populations seen as not "coming for nothing. Often, when they arrive at the ED, it's because they are serious." Cultural differences in the expression of pain were noted, particularly the "Mediterranean syndrome" and individuals who "are theatrical," or some communities whose members are "much more explosive and numerous and therefore very present." Nurses emphasized their efforts to address these factors without bias.

Experience level was the primary determinant of decisions. Novices followed guidelines more closely, while experienced nurses integrated intuition and "gut feelings" into their decision-making.

Data analysis. Among the 273,151 visits analyzed, 53% were males and 47% were females. The average age was 48 years, and the mean varied across triage scores, showing a significant linear trend toward decreasing from 62 (level 1) to 38 years (level 5). Female nurses managed 80% of visits, and recent abdominal pain was the most common chief complaint (9% of visits). In the multinomial model, both the patient's sex and age, and the nurse's sex, were found to have significant associations with the triage score. The age effect, adjusted for patient and nurse sex, remained consistent across successive score comparisons: older age increased the likelihood of receiving a more urgent score.

Adjusted for age and nurse's sex, the probability of being triaged g+1 (less urgent) rather than g (more urgent) was higher for a female compared to a male patient for more urgent triage scores (OR = 0.85, 95%CI $[0.76, 0.97]$ for triage 2 Vs. 1 and OR = 0.81, 95%CI $[0.79, 0.83]$ for triage 3 Vs. 2). Conversely, the probability of being triaged g+1 rather than g was higher for a male compared to a female patient for less urgent triage scores (OR = 1.15, 95%CI $[1.13, 1.17]$ for triage 4 versus 3 and OR = 1.20, 95%CI $[1.16, 1.24]$ for triage 5 versus 4). Adjusted for age and patient's sex, a male nurse had a lower probability of assigning a score of 3 (the

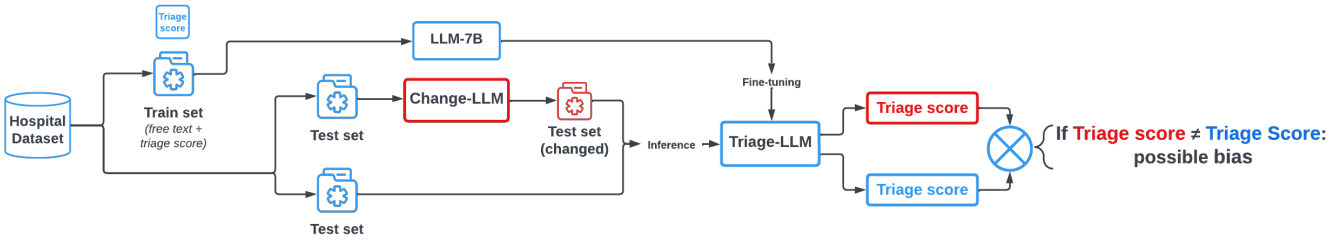

Figure 1: Using LLMs to assess gender bias in ED triage scores.

least specific) than a female nurse (4 vs 3 OR = 1.18, 95%CI $[1.15, 1.21]$, 3 vs 2 OR = 0.97, 95%CI $[0.94, 0.99]$).

LLMs to predict triage score. Clinical BELGPT-2 outperformed all methods, achieving a precision of 0.63, a recall of 0.63, and a micro F1-score of 0.62. XGBoost/TF-IDF and LightGBM/TF-IDF achieved a precision of 0.59 and 0.61, a recall of 0.57 and 0.59, and a micro F1-score of 0.55 and 0.57, respectively. Potential biases and errors in triage nurse decision-making, implicit in the training and test data, may impact the model's performance. Further analysis of the F1-score in each class reveals the common issue of imbalance: all methods perform better in the majority classes (i.e., triage levels 2, 3, and 4) than in the minority classes (i.e., levels 1 and 5). LightGBM/TF-IDF and Clinical BELGPT-2 achieved F1-scores of 0.26 and 0.21 (triage level 1), 0.48 and 0.64 (level 2), 0.64 and 0.62 (level 3), 0.61 and 0.66 (level 4), and 0.31 and 0.40 (level 5), respectively.

Discussion

Promising AI applications hold potential to improve emergency healthcare services, particularly in triage (Yu et al. 2022; Kipourgos et al. 2022; Sanchez-Salmeron et al. 2022; Cho et al. 2022; Vantu, Vasilescu, and Baicoianu 2023; Defilippo et al. 2023; Mutegeki et al. 2023; Sax et al. 2023; Gao et al. 2022; Stewart et al. 2023). However, these models may inherit biases, potentially worsening health disparities. Our review highlights disparities in prioritization based on patient age, ethnicity, and socioeconomic status (Portillo et al. 2023; Peitzman et al. 2023; Essa et al. 2023; Martin et al. 2023; Fekonja et al. 2023). The role of patient and nurse gender/sex is inconsistent, likely due to nuances (favoring women in some situations and disadvantaging them in others), and influenced by various factors.

Nurse interviews reveal the complexity of the triage process, with experienced nurses navigating these complexities more effectively. Descriptive analysis shows a linear association between increasing age and more urgent scores. Interestingly, men tend to receive slightly more urgent scores, while women may receive somewhat more urgent scores for consultation visits. Nurse gender also affects evaluation, with male nurses rating urgency levels slightly lower than female counterparts. Misstriage cannot be inferred from these differences. Additional variables like pain scores and diagnoses could provide further insights.

Finally, our pilot study uses an LLM to predict triage scores, with clinical BELGPT-2 yielding better results than machine learning methods. However, the outcomes remain modest. Hosting data locally limits options, but new open-source Transformer-based LLMs offer competitiveness and in-house deployment capabilities, opening avenues for improvement. The triage score prediction model serves as a proof of concept, ready for further application.

Perspectives for an Experimental Plan

Traditional experiments to assess gender bias in educational or human resources fields often involve manipulating male/female names (Doornkamp et al. 2022). On the other hand, literature presents strategies applying NLP models to evaluate potential language biases towards gender in letters of recommendation (Fu et al. 2023), as well as gender disparities in scientific review processes (Verharen 2023) and bias in job descriptions (Frissen, Adebayo, and Nanda 2022). Studies have proposed identifying and removing gendered language from clinical-note datasets (Minot et al. 2022) and legal datasets (Bozdag, Sevim, and Koç 2023).

In this context, we propose a strategy inspired by these approaches to assess biased behaviors. We aim to evaluate the predictive performance of an LLM for triage scoring while also using an LLM to manipulate male/female language in clinical notes (switching between masculine and feminine gender in the texts). Figure (1) illustrates the concept. Our hypothesis suggests that the predictive model trained using an extensive corpus of patient records can faithfully replicate the decision-making process of healthcare professionals, as demonstrated in our pilot study, incorporating underlying cognitive biases. To evaluate these biases, we will compare the triage levels assigned to an original patient file with those of its manipulated version, where gender (or any other variable of interest corresponding to the studied bias) is altered by our LLM. To our knowledge, this approach is innovative and has not yet been utilized to detect human bias.

Acknowledgments

The idea for this work arose from reflections following MD and DC internships in the Public Health Master program in 2022 and 2023, respectively, both under the BPH Inserm's TARPON project, in collaboration with Bordeaux University Hospital. Ariel Guerra-Adames joined the team and is currently putting the idea into practice as part of his master's internship in 2024. We thank Ariel for his contributions to refining the "Perspectives" section.

References

Arnaud, E.; Elbattah, M.; Ammirati, C.; Dequen, G.; and Ghazali, D. A. 2022. Use of artificial intelligence to manage patient flow in emergency department during the Covid-19 pandemic: A prospective, single-center study. *Int J Environ Res Public Health* 19(15):9667.

Arslanian-Engoren, C. 2000. Gender and age bias in triage decisions. *J Emerg Nurs* 26(2):117–24.

Aubrion, A.; Clanet, R.; Jourdan, J.; Creveuil, C.; Roupie, E.; and Macrez, R. 2022. FRENCH versus ESI: Comparison between two nurse triage emergency scales with referent scenarios. *BMC Emerg Med* 22:201.

Banco, D.; Chang, J.; Talmor, N.; Wadhera, P.; Mukhopadhyay, A.; and Lu, X. 2022. Sex and race differences in the evaluation and treatment of young adults presenting to the emergency department with chest pain. *J Am Heart Assoc* 11(10):e024199.

Bozdag, M.; Sevim, N.; and Koç, A. 2023. Measuring and mitigating gender bias in legal contextualized language models. *ACM Trans Knowl Discov Data*.

Buslon, N.; Cortes, A.; Catuara-Solarz, S.; and Cirillo, D, R. M. 2023. Raising awareness of sex and gender bias in artificial intelligence and health. *Front Glob Womens Health* 4:970312.

Chenais, G.; Lagarde, E.; and Gil-Jardine, C. 2023. Artificial intelligence in emergency medicine: Viewpoint of current applications and foreseeable opportunities and challenges. *Med Internet Res* 25:e40031.

Cho, A.; Min, I. K.; Hong, S.; Chung, H. S.; Lee, H. S.; and Kim, J. H. 2022. Effect of applying a real-time medical record input assistance system with voice artificial intelligence on triage task performance in the ED: Prospective interventional study. *JMIR Med Inform* 10(8):e39892.

Coisy, F.; Olivier, G.; Ageron, F.-X.; Guillermou, H.; Roussel, M.; Balen, F.; Grau-Mercier, L.; and Bobbia, X. 2023. Do emergency medicine health care workers rate triage level of chest pain differently based upon appearance in simulated patients? *Eur J Emerg Med*.

Defilippo, A.; Bertucci, G.; Zurzolo, C.; Veltri, P.; and Guzzi, P. 2023. On the computational approaches for supporting triage systems. *Interdiscip Med* 1(3):e20230015.

Doornkamp, L.; Van der Pol, L.; Groeneveld, S.; Mesman, J.; Endendijk, J.; and Groeneveld, M. 2022. Understanding gender bias in teachers' grading: The role of gender stereotypical beliefs. *Teach Teach Educ* 118:103826.

Emami, P., and K., J. 2023. Enhancing emergency response through artificial intelligence in emergency medical services dispatching; a letter to editor. *Arch Acad Emerg Med.* 11(1):e60.

Essa, C.; Victor, G.; Khan, S.; Ally, H.; and Khan, A. 2023. Cognitive biases regarding utilization of emergency severity index among emergency nurses. *Am J Emerg Med* 73:63–8.

Fekonja, Z.; Kmetec, S.; Fekonja, U.; Mlinar Reljić, N.; Pajnkihar, M.; and Strnad, M. 2023. Factors contributing to patient safety during triage process in the emergency department: A systematic review. *J Clin Nurs* 32:5461–77.

Frissen, R.; Adebayo, K.; and Nanda, R. 2022. A machine learning approach to recognize bias and discrimination in job advertisements. *AI Soc* 38:1–14.

Fu, S.; Calley, D.; Rasmussen, V.; Hamilton, M.; Lee, C.; Kalla, A.; and Liu, H. 2023. Gender-based language differences in letters of recommendation. *AMIA Jt Summits Transl Sci Proc* 196–205.

Gao, Z.; Qi, X.; Zhang, X.; Gao, X.; He, X.; Guo, S.; and Li, P. 2022. Developing and validating an emergency triage model using machine learning algorithms with medical big data. *Risk Manag Healthc Policy* 15:1545–51.

Gorick, H. 2022. Factors that affect nurses' triage decisions in the emergency department: A literature review. *Emerg Nurse* 30(3):14–9.

Hinson, J. S.; Martinez, D. A.; Cabral, S.; George, K.; Whalen, M.; Hansoti, B.; and Levin, S. 2019. Triage performance in emergency medicine: A systematic review. *Ann Emerg Med* 74(1):140–52.

Kipourgos, G.; Tzenalis, A.; Koutsojannis, C.; and Hatzilygeroudis, I. 2022. An artificial intelligence based application for triage nurses in emergency department, using the emergency severity index protocol. *Int J Caring Sci* 15:1764.

Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI'23, 12–24. New York, NY, USA: Association for Computing Machinery.

Lee, S., and Lee, Y. 2020. Improving emergency department efficiency by patient scheduling using deep reinforcement learning. *Healthc* 8(2):77.

Lin, P.; Argon, N.; Cheng, Q.; Evans, C.; Linthicum, B.; and Liu, Y. 2022. Disparities in emergency department prioritization and rooming of patients with similar triage acuity score. *Acad Emerg Med* 29(11):1320–8.

Liventsev, V.; Härmä, A.; and Petković, M. 2021. Towards effective patient simulators. *Front Artif Intell Appl.* 4:798659.

Lopez, R.; Snair, M.; Arrigain, S.; Schold, J.; Hustey, F.; and Walker, L. 2021. Sex-based differences in timely emergency department evaluations for patients with drug poisoning. *Public Health* 199:57–64.

Martin, S.; Heyming, T.; Kain, A.; Krauss, B.; and Campos, B. 2023. Eliminating pain disparities for children in the emergency department. *Acad Emerg Med* 30(10):1075–7.

Minot, J. R.; Cheney, N.; Maier, M.; Elbers, D. C.; Danforth, C. M.; and Dodds, P. S. 2022. Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. *ACM Trans Comput Healthcare* 3(4).

Mnatzaganian, G.; Hiller, J.; Braitberg, G.; Kingsley, M.; Putland, M.; and Bish, M. 2020. Sex disparities in the assessment and outcomes of chest pain presentations in emergency departments. *Heart* 106(2):111–8.

Mutegeki, H.; Nahabwe, A.; Nakatumba-Nabende, J.; and Marvin, G. 2023. Interpretable machine learning-based triage for decision support in emergency care. In *7th International Conference on Trends in Electronics and Informatics*, 983–90.

Onal, E.; Knier, K.; Hunt, A.; Knudsen, J.; Nestler, D.; and Campbell, R. 2022. Comparison of emergency department throughput and process times between male and female patients: A retrospective cohort investigation by the reducing disparities increasing equity in emergency medicine study group. *J Am Coll Emerg Physicians Open* 3(5):e12792.

Peitzman, C.; Carreras Tartak, J.; Samuels-Kalow, M.; Raja, A.; and Macias-Konstantopoulos, W. 2023. Racial differences in

triage for emergency department patients with subjective chief complaints. *West J Emerg Med* 24(5):888–93.

Piliuk, K., and Tomforde, S. 2023. Artificial intelligence in emergency medicine. a systematic literature review. *Int J Med Inform* 180:105274.

Portillo, E.; Rees, C.; Hartford, E.; Foughty, Z.; Pickett, M.; Gutman, C.; Shihabuddin, B.; Fleegler, E.; Chumpitazi, C.; Johnson, T.; Schnadower, D.; and Shaw, K. 2023. Research priorities for pediatric emergency care to address disparities by race, ethnicity, and language. *JAMA Netw Open* 6(11):e2343791.

Preciado, S.; Sharp, A.; Sun, B.; Baecker, A.; Wu, Y.; and Lee, M. 2021. Evaluating sex disparities in the emergency department management of patients with suspected acute coronary syndrome. *Ann Emerg Med* 77(4):416–24.

Rosemarin, H.; Rosenfeld, A.; and Kraus, S. 2019. Emergency department online patient-caregiver scheduling. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:10013–4.

Sanchez-Salmeron, R.; Gomez-Urquiza, J.; Albendin-Garcia, L.; Correa-Rodriguez, M.; Martos-Cabrera, M.; Velando-Soriano, A.; and Suleiman-Martos, N. 2022. Machine learning methods applied to triage in emergency services: A systematic review. *Int Emerg Nurs* 60:101109.

Sax, D.; Warton, E.; Sofrygin, O.; Mark, D.; Ballard, D.; Kene, M.; Vinson, D.; and ME, R. 2023. Automated analysis of unstructured clinical assessments improves emergency department triage performance: A retrospective deep learning analysis. *J Am Coll Emerg Physicians Open* 4(4):e13003.

Stewart, J.; Lu, J.; Goudie, A.; Arendts, G.; Meka, S.; Freeman, S.; Walker, K.; Sprivulis, P.; Sanfilippo, F.; Bennamoun, M.; and Dwivedi, G. 2023. Applications of natural language processing at emergency department triage: A narrative review. *PLOS One* 18:e0279953.

Suamchaiyaphum, K.; Jones, A.; and Markaki, A. 2023. Triage accuracy of emergency nurses: An evidence-based review. *J Emerg Nurs* 158(1767):00251–9.

Taylor, A.; Murakami, M.; Kim, S.; Chu, R.; and Riek, L. D. 2022. Hospitals of the future: Designing interactive robotic systems for resilient emergency departments. *Proc ACM Hum-Comput Interact* 6.

van der Stigchel, B.; van den Bosch, K.; van Diggelen, J.; and P, H. 2023. Intelligent decision support in medical triage: Are people robust to biased advice? *J Public Health* 45(3):689–96.

Vantu, A.; Vasilescu, A.; and Baicoianu, A. 2023. Medical emergency department triage data processing using a machine-learning solution. *Heliyon* 9(8):e18402.

Verharen, J. P. H. 2023. ChatGPT identifies gender disparities in scientific peer review. *eLife* 12:RP90230.

Vigil, J.; Coulombe, P.; Alcock, J.; Stith, S.; Kruger, E.; and Cichowski, S. 2017. How nurse gender influences patient priority assignments in US emergency departments. *Pain* 158(3):377–82.

Yu, J. Y.; Xie, F.; Nan, L.; Yoon, S.; Ong, M. E. H.; Ng, Y. Y.; and Cha, W. C. 2022. An external validation study of the score for emergency risk prediction (SERP), an interpretable machine learning-based triage score for the emergency department. *Sci Rep* 12(1):17466.