Uncovering Judgment Biases in Emergency Triage: A Public Health Approach Based on Large Language Models

Ariel Guerra-Adames, Marta Avalos, Océane Dorémus, Cédric Gil-Jardiné, Emmanuel Lagarde

▶ **To cite this version:**

Ariel Guerra-Adames, Marta Avalos, Océane Dorémus, Cédric Gil-Jardiné, Emmanuel Lagarde. Uncovering Judgment Biases in Emergency Triage: A Public Health Approach Based on Large Language Models. ML4H 2024 - Machine Learning for Health Symposium, Dec 2024, Vancoucer, Canada. . hal-04846696

HAL Id: hal-04846696
https://hal.science/hal-04846696v1

Submitted on 20 Dec 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.

Uncovering Judgment Biases in Emergency Triage:
A Public Health Approach Based on Large Language Models

Ariel Guerra-Adames[1 2 3], Marta Avalos-Fernandez[1 3], Océane Doremus[1 2], Cédric Gil-Jardiné[1 2 4], Emmanuel Lagarde[1 2]

[1] University of Bordeaux, Bordeaux Population Health Research Center (BPH), INSERM, Bordeaux, France
[2] AHeaD Team, BPH, Bordeaux, France, [3] SISTM Team, INRIA, Talence, France
[4] University Hospital of Bordeaux, Pole of Emergency Medicine, Bordeaux, France

Introduction:

- Emergency triage: **rapid assessment and categorization of patients** based on the severity of their conditions.
- Judgment biases are cognitive shortcuts in decision-making leading to **over-generalization based on stereotypes or incomplete information**.
- LLMs **match or surpass human medical decision-making accuracy.**

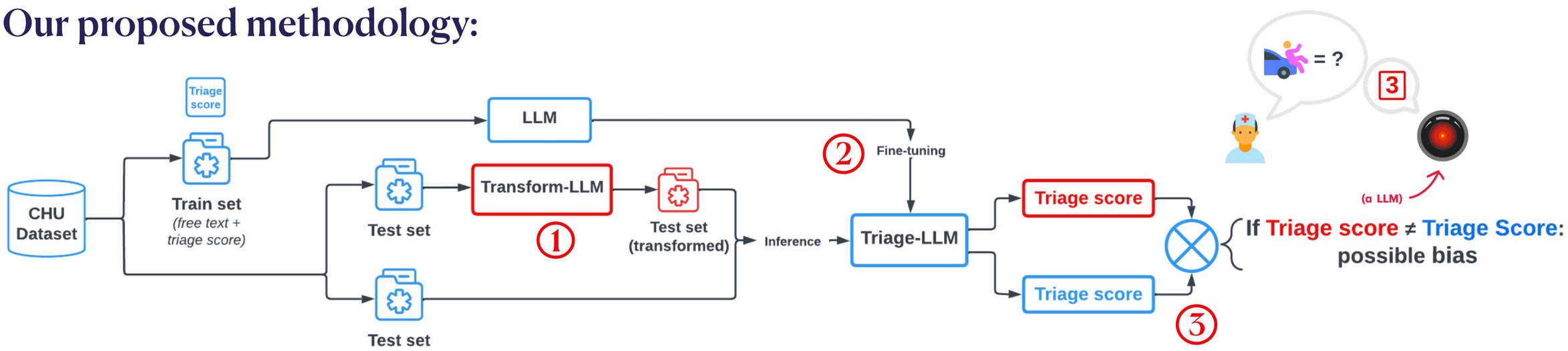
Are there variables which influence decision-making in emergency medicine, but shouldn't?
If so, how can we identify them when discrimination testing is too difficult or even unethical?

Main hypothesis: State-of-the-art LLMs can be used to answer this question.

Available data:

- **< 480,000 admissions** to the adult ED of the Bordeaux University Hospital (CHU Bordeaux), from 2013 to 2021.
- Each contains: **HPI notes, vital parameters, CC, patient context, nurse context + the associated triage score.**
- Triage is made on a 5-level scale: **1 = risk of death, 5 = not at risk of death.**

Our proposed methodology:



1. Few-shot clinical note transformation

- Using **Mixtral 8x22B v0.1, 7-shot configuration**.
- Instructed to change all references to a patient's gender towards the opposite gender (counterfactual sampling).
- Manually evaluated 100 random samples from test set.

Original synthetic extract (male)

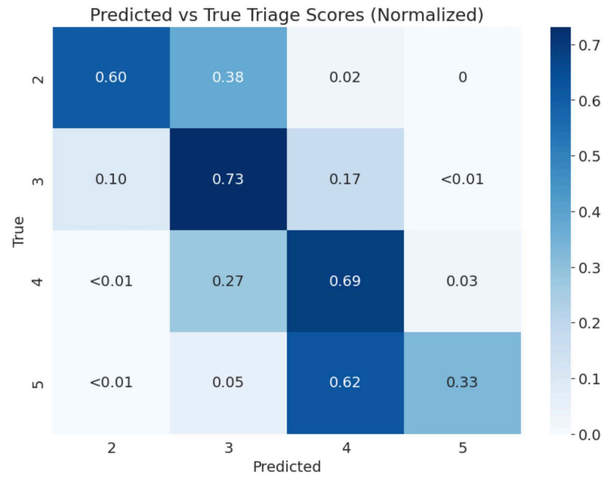
Sexe patient(e): **M**, Anamnèse patient(e) : **Patient** qui se plaint d'une d thoracique ECG fait et vu par medecin en box : GSC 14, [...] ne se dit pas **gêné** pour respirer , ps de diarrhée

Transformed synthetic extract (female)

Sexe patient(e): **F**, Anamnèse patient(e) : **Patiente** qui se plaint d'une d thoracique ECG fait et vu par medecin en box : GSC 14, [...] ne se dit pas **gênée** pour respirer , ps de diarrhée

2. LLM-based automatic emergency triage

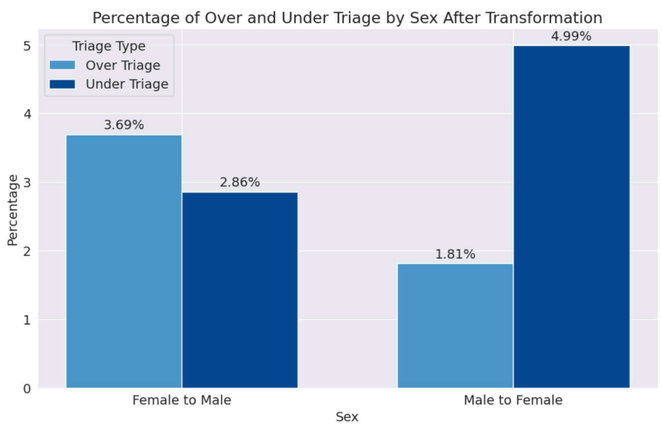
- 4 different LLMs-**Mistral 7B, BioMistral 7B, Mixtral 8x7B, Llama 3 8B-fine-tuned** on train partition to predict triage scores.
- Obtained weighted Cohen's **kappa of 0.71**: similar to human triage nurses in the literature.
- **Mistral 7B** selected for **downstream task** of triaging both versions of test set.



Model Name	Precision	Recall	Specificity	F1	κ	MAE
Mistral 7B	**0.67**	**0.67**	0.81	**0.67**	**0.71**	**0.34**
BioMistral 7B	0.66	0.66	0.81	0.65	0.69	0.35
Mixtral 8x7B	0.64	0.62	0.79	0.61	0.65	0.39
Llama 3 8B	**0.67**	**0.67**	**0.82**	0.66	**0.71**	**0.34**

3. Downstream task - bias detection

- **Compared predicted triage scores between original and transformed** samples with a paired t-test ($p<0.001$).
- On samples which originally indicated male sex, transformed towards **female**: **higher percentage of samples triaged as less critical** (≈ 5%) compared to those which were triaged as more critical (1.81%) after the transformation.
- On samples which originally indicated female sex, transformed towards male, less pronounced difference.
- We interpret this difference as **female patients being more likely to be under-triaged with respect to their male counterparts.**



Conclusion and Perspectives

- **LLMs are capable of imitating human medical decision-making**, serving as a proxy for discrimination testing.
- A **triage 'silver standard' through expert consensus** could make evaluation of the automatic triage models more robust.
- A portion of the differences may come from imitated biases, but **there may also be algorithmically-induced biases**.
- All information available during triage is not explicit: **implementation of multi-modal triage** in the near future.

Acknowledgments

We wish to thank all triage nurses who participated in the study, as well as the Digital Public Health Graduate Program of the University of Bordeaux for financing the doctoral thesis of Ariel Guerra-Adames.

Correspondance:

ariel.guerra-adames@u-bordeaux.fr