Réflexions pour la conception d'un protocole expérimental de détection des biais dans le triage d'urgence hospitalière à l'aide de modèles de langage

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

Réflexions pour la conception d'un protocole expérimental de détection des biais dans le triage d'urgence hospitalière à l'aide de modèles de langage

Ariel Guerra-Adames[1] 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) Université de Bordeaux, Bordeaux Population Health Research Center, UMR U1219, INSERM, F-33000, Bordeaux, France

(2) Équipe SISTM, Centre INRIA de l'Université de Bordeaux, F-33405, Talence, France

(3) CHU de Bordeaux, Pôle urgences adultes, F-33000, Bordeaux, France

prenom.nom1-nom2@u-bordeaux.fr

RÉSUMÉ

L'essor de la recherche basée sur l'IA dans les soins de santé d'urgence soulève des défis tels que la conformité à la protection des données et le risque d'accentuer les inégalités en reproduisant les biais présents dans les données utilisées pour entraîner les systèmes d'IA. Toutefois, l'IA offre également la possibilité de corriger ces biais. Notre étude porte en particulier sur le triage d'urgence, visant à classer rapidement les patients selon la gravité de leur état à leur arrivée. Nous avons réalisé une revue de la littérature pour identifier les biais potentiels dans le triage et mené une étude préliminaire impliquant une enquête qualitative et une analyse descriptive des données de triage. A partir de ces éléments, nous décrivons un protocole expérimental destiné à évaluer l'efficacité de l'IA dans la détection des biais au sein des données de triage.

ABSTRACT

Reflections for the design of an experimental protocol for Bias Detection in Hospital Emergency Triage using Language Models

The rise of AI-based research in emergency healthcare raises challenges such as data protection compliance and the risk of accentuating inequalities by reproducing the biases present in the data used to train AI systems. However, AI also offers the possibility of correcting these biases. Our study focuses in particular on emergency triage, aimed at rapidly classifying patients according to the severity of their condition on arrival. We carried out a literature review to identify potential biases in triage, and conducted a preliminary study involving a qualitative survey and descriptive analysis of triage data. Based on this, we describe an experimental protocol designed to assess the effectiveness of AI in detecting bias within triage data.

MOTS-CLÉS : Biais Humain, Grands Modèles de Langue, Traitement Automatique des Langues, Service d'Urgences, Triage.

KEYWORDS: Human Bias, Large Language Models, Natural Language Processing, Emergency Department, Triage.

1 Introduction

L'intérêt croissant pour l'utilisation de l'intelligence artificielle (IA) dans le domaine de la santé, en particulier dans les services d'urgences, témoigne de la recherche constante d'amélioration de l'efficacité et de la précision des soins. La médecine d'urgence exige une organisation rigoureuse, une coordination sans faille et une prise de décision rapide pour les patients présentant des pathologies aiguës, faisant de l'IA une solution particulièrement prometteuse (Taylor *et al.*, 2022; Piliuk & Tomforde, 2023; Chenais *et al.*, 2023). Les propositions d'application de l'IA dans divers domaines montrent un potentiel significatif pour l'amélioration des services de médecine d'urgence, y compris dans les contextes préhospitaliers (Rosemarin *et al.*, 2019; Lee & Lee, 2020), les services de régulation médicale (Emami & K., 2023), la gestion des flux de patients (Liventsev *et al.*, 2021; Arnaud *et al.*, 2022) et le triage d'urgence (Yu *et al.*, 2022; Kipourgos *et al.*, 2022; Sanchez-Salmeron *et al.*, 2022; Cho *et al.*, 2022; Vantu *et al.*, 2023; Defilippo *et al.*, 2023; Mutegeki *et al.*, 2023; Sax *et al.*, 2023; Gao *et al.*, 2022), avec un accent particulier sur les applications de traitement automatique des langues (TAL) utilisant des grands modèles de langue (LLM) (Stewart *et al.*, 2023). Toutefois, l'intégration de l'IA dans les services d'urgences soulève également des défis considérables. Il est impératif de prendre en compte les considérations éthiques et légales pour assurer la protection de la vie privée des patients, la conformité aux régulations en vigueur (van der Stigchel *et al.*, 2023), et pour aborder les biais potentiels dans les algorithmes d'IA afin de prévenir les disparités dans les soins aux patients.

Un exemple frappant de disparité en matière de santé concerne le sexe/genre. Les facteurs biologiques (sexe) et socioculturels (genre) jouent un rôle déterminant dans les maladies chroniques et les processus physiologiques, y compris la sensibilité à la douleur, conduisant à des différences notables dans l'épidémiologie des maladies. Ces variations peuvent être attribuées à des différences dans les conditions cliniques, le début, la présentation des symptômes, le pronostic, les biomarqueurs et l'efficacité des traitements. Néanmoins, l'évaluation, le diagnostic et le traitement des conditions de santé peuvent également être influencés par des biais cognitifs inconscients chez les professionnels de santé et par les stéréotypes sociaux. Au-delà des biais humains, les implications des biais dans les algorithmes d'IA pour les applications de santé sont de plus en plus préoccupantes. Les études montrent que les LLM présentent des biais alignés avec les rôles stéréotypés lorsqu'ils sont entraînés sur des données où les femmes sont sous-représentées, comme les données cliniques (Kotek *et al.*, 2023; Buslon *et al.*, 2023). En conséquence, ces modèles renforcent les biais conformes aux perceptions sociétales et ignorent souvent les ambiguïtés dans la structure des phrases, fournissant des explications inexactes pour leurs choix biaisés.

Face au risque que les modèles d'IA héritent des biais de leurs données d'entraînement et potentiellement exacerbent les disparités en matière de santé, nous avons entrepris une revue exhaustive de la littérature pour identifier les biais humains potentiels dans le triage. Pour compléter cette revue et obtenir des informations pratiques, nous avons mené une enquête qualitative. De plus, nous avons effectué une analyse descriptive des données de triage en utilisant des facteurs démographiques tels que le sexe et l'âge du patient, ainsi que le sexe de l'infirmier ou de l'infirmière organisateur(trice) de l'accueil (IOA), provenant du CHU de Bordeaux. Nous avons ensuite mis en œuvre un triage piloté par l'IA en utilisant un modèle LLM. Enfin, en combinant ces éléments, nous avons décrit un plan expérimental novateur pour évaluer l'efficacité des LLM dans la détection des biais dans les données de triage d'urgence.

2 Triage d'Urgence

À leur arrivée au service des urgences, une infirmière ou un infirmier organisateur(trice) de l'accueil (IOA) procède rapidement à l'évaluation des patients, consignant des informations essentielles telles que le motif de la visite, les signes vitaux et les antécédents médicaux sous forme de notes textuelles libres. L'IOA utilise des outils et des protocoles spécifiques pour assurer un triage standardisé. Une évaluation précise de la gravité des cas est cruciale : une sous-estimation de l'urgence peut retarder les soins et aggraver les résultats cliniques, tandis qu'une surestimation peut entraîner une utilisation excessive des ressources et des coûts plus élevés. De nombreuses études ont évalué différentes échelles (Aubrion *et al.*, 2022), avec des résultats allant d'une validité modérée à bonne, sans qu'aucune échelle n'émerge comme la norme de référence absolue.

Les grilles de triage, utilisées par les IOA pour catégoriser les patients, incluent l'Indice de Gravité des Urgences (Emergency Severity Index, ESI) largement adopté dans les hôpitaux américains, les Échelles Canadiennes de Triage et de Gravité aux Urgences (CTAS), l'Échelle de Triage Australasienne (ATS), l'Échelle de Triage de Manchester (MTS), l'Échelle de Triage Sud-Africaine (SATS) pour les pays en développement, et la *FRench Emergency Nurses Classification in-Hospital triage* (FRENCH) en France. Bien que chacune possède ses spécificités, elles visent toutes à définir objectivement la gravité des cas, la complexité des soins requis et les ressources nécessaires, aboutissant ainsi à une priorisation et à un temps d'attente maximum ou à un nombre d'examens définis. Les variables déterminant le score de triage comprennent des facteurs communs tels que les signes vitaux et des facteurs moins consensuels tels que l'âge, le niveau de douleur ou les antécédents médicaux. Chaque établissement de santé sélectionne une échelle ou une grille de triage validée, fiable et reproductible, comportant 4 ou 5 niveaux, et adaptée aux caractéristiques nationales de la santé. Les recommandations concernant les grilles de triage proviennent d'organisations telles que la Société Française de Médecine d'Urgence [1] ou l'American College of Emergency Physicians [2].

3 Biais Humains dans le Triage d'Urgence

Nous avons conduit une recherche non systématique dans la bases de données bibliographiques Medline/PubMed, en nous focalisant sur les titres et résumés des publications. Notre recherche incluait des termes généraux tels que 'biais' et 'urgence', ainsi que des termes spécifiques comme 'genre' et 'triage', comprenant leurs synonymes et termes associés. En outre, nous avons employé des techniques de recherche en boule de neige en parcourant les listes de références des articles identifiés. Nous présentons ici une synthèse de nos résultats.

Des facteurs externes, tels que l'emplacement du service des urgences, l'heure et le jour d'arrivée, peuvent influencer les décisions de triage et la correspondance entre les scores attribués et les interventions ultérieures (Gorick, 2022; Suamchaiyaphum *et al.*, 2023). Les préoccupations juridiques relatives aux erreurs de triage, notamment les complications dues à un sous-triage pouvant conduire au décès d'un patient (Hinson *et al.*, 2019), peuvent inciter les IOA à attribuer un score de gravité plus élevé en cas d'incertitude.

Les études révèlent des différences significatives dans la priorisation en fonction de l'âge du patient,

1. www.sfmu.org/upload/referentielsSFMU/ioa2004.pdf
2. www.acep.org/siteassets/uploads/uploaded-files/acep/clinical-and-practice-managem
resources/administration/triagescaleip.pdf

de son appartenance ethnique et de sa couverture d'assurance maladie (Portillo *et al.*, 2023; Peitzman *et al.*, 2023; Essa *et al.*, 2023; Martin *et al.*, 2023; Fekonja *et al.*, 2023). Les jeunes, les patients afro-américains ou hispaniques, ainsi que ceux provenant de quartiers économiquement défavorisés, sont plus susceptibles d'être victimes d'erreurs de triage (Banco *et al.*, 2022). Des comportements inattendus sont observés, avec plus de 10% des consultations ne respectant pas l'ordre d'arrivée, favorisant les personnes âgées et défavorisant les personnes racisées, même au sein du même niveau de triage où le principe supposé est 'premier arrivé, premier servi' (Lin *et al.*, 2022).

Les résultats relatifs au sexe/genre sont tantôt contradictoires, tantôt non concluants (Arslanian-Engoren, 2000; Onal *et al.*, 2022). Certaines études suggèrent que les personnes âgées et les patients présentant des symptômes graves ou à haut risque sont traités avec la même urgence, quel que soit leur sexe. Cependant, d'autres recherches indiquent des désavantages potentiels pour les femmes à divers stades de soins. Par exemple, les femmes peuvent connaître des durées de visite plus longues, des temps d'attente avant traitement plus élevés, ou des consultations avec des professionnels de santé moins fréquentes. De plus, les études portant sur les patients à faible risque de syndrome coronarien aigu montrent que les femmes sont moins souvent hospitalisées et subissent moins d'examens que les hommes. Fait intéressant, ces disparités peuvent, de manière involontaire, bénéficier aux femmes en leur évitant des hospitalisations ou des tests cardiaques inutiles. À l'inverse, une étude récente portant sur la douleur thoracique (Coisy *et al.*, 2023) a montré que la visualisation de patients simulés avec différentes caractéristiques modifiait la décision de priorité. Comparés aux patients blancs, les patients noirs étaient moins susceptibles de recevoir un traitement d'urgence. Il en était de même pour les femmes par rapport aux hommes. Les résultats ont révélé des disparités dans le traitement d'urgence, les patients noirs et les femmes étant moins susceptibles de recevoir des soins rapides.

L'interprétation des disparités entre les sexes est complexe en raison des symptômes souvent "atypiques" présentés par les femmes pour des conditions graves telles que les AVC, les crises cardiaques, l'appendicite ou les intoxications aiguës par des substances autres que les opioïdes (Preciado *et al.*, 2021; Mnatzaganian *et al.*, 2020; Lopez *et al.*, 2021). Cela pourrait mener à une sous-estimation de l'urgence et à des retards de diagnostic. Les études cliniques sont principalement développées en utilisant des modèles masculins, compliquant ainsi davantage la situation. De plus, les femmes peuvent présenter des symptômes plus hétérogènes et moins précis, et être plus sujettes à la douleur chronique. Plusieurs mécanismes biopsychosociaux contribuent à leur plus grande sensibilité à la douleur et à des réactions moins efficaces aux traitements de la douleur comparativement aux hommes. Cependant, les modèles sociaux liés aux différences de genre influencent également l'expression de la douleur.

Des travaux suggèrent que les biais dans le triage des patients sont davantage influencés par le genre de l'IOA que par celui du patient (Vigil *et al.*, 2017). Les patientes reçoivent un triage similaire quel que soit le genre de l'IOA, tandis que les patients masculins peuvent recevoir des scores différents en fonction du genre de l'IOA. Les infirmières attribuent des scores de gravité plus élevés aux patients masculins présentant des niveaux de douleur plus élevés, tandis que les infirmiers attribuent des scores plus bas. Cette tendance est plus marquée dans les services d'urgence avec un personnel infirmier majoritairement féminin. La perception selon laquelle les hommes sont plus enclins à paniquer et à exagérer leurs symptômes, tandis que les femmes sont perçues comme plus calmes et stoïques, contribue à ces différences dans les décisions de triage.

4 Étude Préliminaire

Afin de mieux comprendre les pratiques de triage et de confronter notre revue de la littérature, nous avons mené une étude comportant des entretiens avec des IOA de trois hôpitaux. Par ailleurs, nous avons réalisé une analyse descriptive des données de triage du CHU de Bordeaux, en nous concentrant sur les facteurs influençant le plus les scores de triage selon la littérature ou les entretiens, lorsque disponibles et leur utilisation était autorisée. Enfin, nous avons utilisé un LLM pour prédire les scores de triage. Cette étude est conforme aux directives du Comité d'Éthique de la Recherche du CHU de Bordeaux.

4.1 Méthodes

Étude qualitative basée sur des entretiens. L'étude a recruté des participants dans les services d'urgences du CHU de Bordeaux ainsi que des hôpitaux Saint-Antoine AP-HP et Lariboisière AP-HP à Paris. Les participants éligibles étaient des IOA ayant suivi une formation au triage, possédant une expérience préalable en triage, capables de fournir un compte rendu oral de leurs expériences et s'engageant à participer à des entretiens d'une heure. L'ensemble des entretiens semi-structurés individuels ont été menés par une seule enquêtrice qui a enregistré les conversations à l'aide d'un enregistreur vocal après consentement des participants. Un guide d'entretien complet, élaboré à partir de la revue de la littérature, a été utilisé pour couvrir tous les sujets pertinents, incluant les données démographiques des patients, les facteurs contextuels entourant leur visite, les caractéristiques des IOA, ainsi que des questions institutionnelles plus larges. Les entretiens ont été transcrits textuellement à l'aide d'un logiciel de transcription, et une analyse thématique a été appliquée pour identifier les motifs et thèmes récurrents à travers les entretiens. Dix IOA (8 femmes et 2 hommes) ont participé à l'étude, avec une expérience variant de 4 à 25 ans. Les entretiens se sont déroulés entre le 17 mai et le 31 mai 2023, avec une durée moyenne de 73 minutes par session et ont eu lieu dans les hôpitaux respectifs dans lesquels travaille chaque IOA.

Analyse des données. La base de données comprend des enregistrements complets des visites aux urgences des adultes (âgés de 15 ans et plus) au CHU de Bordeaux, de janvier 2013 à décembre 2021, totalisant 480.001 visites. Les variables analysées incluaient le mois/année de la visite, le sexe du patient (M/F), l'âge du patient, le sexe de l'IOA, et le score de triage attribué. Les entrées avec l'identifiant du patient ou des variables manquants ont été exclues, et les visites avant janvier 2016 ont été filtrées en raison d'un nouveau protocole de triage, rendant les scores avant cette période incomparables. L'ensemble de données résultant comprenait 273.151 visites. Les sources potentielles de biais de sélection ont été examinées en comparant les données avec et sans scores de triage. Des analyses bivariées ont été effectuées pour évaluer l'association entre le score de triage et les autres variables. Des analyses multivariées ont permis d'explorer l'association entre le score de triage et le sexe du patient, en tenant compte de l'âge du patient et du sexe de l'IOA via une régression logistique multinomiale. Nous avons vérifié la linéarité entre l'âge et le logit, exploré les mesures d'adéquation, et testé l'interaction entre le sexe du patient et le sexe de l'IOA. La signification statistique a été fixée à $p < 0.05$.

Utilisation des LLM pour prédire le score de triage. Les visites des patients aux urgences du CHU de Bordeaux de 2013 à 2020 ont été incluses. Après exclusion des visites présentant des valeurs manquantes du triage ou des notes cliniques, un total de 296.071 visites a été retenu. Toutes les variables collectées par les IOA ont été analysées, y compris les données structurées telles que

les signes vitaux, le motif principal de la consultation, l'âge, le sexe, la douleur, les nausées, le niveau d'alcool, et le moment de la visite, ainsi que les données non structurées des notes cliniques. L'outcome était le score de triage. Ces variables ont été analysées en utilisant les classifieurs XGBoost et LightGBM combinés à la vectorisation TF-IDF, ainsi que le modèle de langues BELGPT-2 (Louis, 2020). Ce modèle, basé sur l'architecture GPT-2, avait déjà été pré-entraîné sur un large corpus français hétérogène, et nous l'avons en outre pré-entraîné sur notre corpus du domaine clinique pendant quatre époques. Nous avons ensuite affiné le modèle sur la tâche en aval de classification des scores de triage en fonction de toutes les informations disponibles pour les IOA ainsi que de leurs notes de triage pendant deux époques. Les performances du modèle ont été évaluées en fonction des métriques de précision, de rappel et du micro F1 score.

4.2 Résultats

Étude qualitative basée sur des entretiens. Lors des entretiens, les questions ont exploré l'impact des caractéristiques des patients (sexe/genre, ethnicité, âge, accompagnants), des caractéristiques des IOA (sexe/genre, expérience, comportement individuel, niveau de fatigue), des éléments de la grille de triage (symptômes principaux, signes vitaux, circonstances, antécédents médicaux, douleur rapportée/observée), et des facteurs contextuels (état de saturation des urgences, statut Covid-19, heure et jour d'arrivée, manque de médecins généralistes, emplacement des urgences, mode d'arrivée) sur la pratique des IOA. Les entretiens ont été transcrits et codés à l'aide du logiciel Taguette, et les tags ont été déclinés en thèmes prédéfinis ainsi qu'en thèmes émergents des entretiens.

Presque tous les IOA ont mentionné le stéréotype selon lequel les hommes sont un peu plus "douillets", "chochottes" ou "douloureux" que les femmes, mais ils ne partageaient pas nécessairement cette opinion. Bien que les hommes soient perçus comme plus "plaintifs" ou "dramatiques", en fin de compte, ils évaluent leur douleur de manière similaire à celle des femme. Une seule IOA a mentionné une différence potentielle d'empathie chez les infirmières envers les patientes souffrant de douleurs menstruelles. L'âge est apparu comme un facteur universel dans les décisions de triage. Durant les périodes de forte affluence, les IOA pouvaient surévaluer les patients âgés ("ce n'est pas leur rendre service de les faire attendre dans le couloir sous des néons pendant 4 à 5 heures"). Les patients âgés étaient perçus comme ayant "inévitablement des facteurs de risque", présentant un risque plus élevé de déambulation, ayant "traversé des guerres", et "plus résistants à la douleur".

Les influences culturelles ont également été reconnues comme jouant un rôle dans le triage. Les IOA percevaient certaines populations comme plus prédisposées à certaines conditions de santé et rapportaient accorder une attention particulière à certaines populations considérées comme ne venant pas "pour rien. Souvent, quand ils arrivent aux urgences, c'est qu'ils sont graves". Les différences culturelles se traduisent aussi parfois dans l'expression de la douleur, notamment avec le "syndrome méditerranéen" et les personnes qui "sont dans la dramaturgie", ou des communautés dont les membres sont "beaucoup plus explosifs et très nombreux et donc très très présents".

Les IOA ont souligné leurs efforts pour aborder ces facteurs sans biais, considèrant qu'il est essentiel de maintenir une approche factuelle, en évitant d'entrer dans des considérations sociologiques ou émotionnelles.

Le niveau d'expérience était le principal déterminant des décisions. Les novices vont "avoir tendance par sécurité à [se] référer au texte et uniquement au texte", tandis que celles avec plus d'ancienneté vont s'appuyer davantage sur leur expérience et leur intuition, le ressenti. Les IOA qui débutent auront

plus tendance à surcoter les patients, surtout en cas de doute, par “peur de passer à côté de quelque chose, de [se] tromper”.

De tous les facteurs discutés avec les IOA, l’état des urgences revient comme le facteur primordial, car “c’est là où [le triage] prend toute sa signification, c’est quand c’est le bordel, c’est quand il y a de l’attente. . .On ne fait pas le même tri en fonction de l’état des urgences. Ça c’est sûr et certain”. Le triage étant une réponse au déséquilibre entre besoins et ressources dans un contexte où les moyens sont dépassés, la question du niveau de tri se pose beaucoup moins quand on sait que le patient sera pris en charge rapidement peu importe son score. La saturation du service des urgences, aggravée par le manque de personnel médical (à l’hôpital ainsi qu’en médecine de ville), la fermeture de services, les réorientations de patients qui pourraient pourtant être pris en charge, la population vieillissante, et le manque de compréhension ou de littératie en santé des “gens [qui] n’ont pas saisi le sens du mot urgence”, accentue l’importance du choix du tri.

Analyse des données. Parmi les 273.151 visites analysées, 53% des patients étaient des hommes et 47% étaient des femmes. L’âge moyen était de 48 ans, et la moyenne variait en fonction des scores de triage, montrant une tendance linéaire significative à la baisse de 62 ans (niveau 1) à 38 ans (niveau 5). Les infirmières géraient 80% des visites, et la douleur abdominale récente était la plainte principale la plus courante (9% des visites). Dans le modèle multinomial, à la fois le sexe et l’âge du patient, ainsi que le sexe de l’IOA, ont été trouvés significativement associés au score de triage. L’effet de l’âge, ajusté pour le sexe du patient et celui de l’IOA, restait constant à travers les comparaisons de scores successives : un âge plus avancé augmentait la probabilité de recevoir un score plus urgent.

Ajusté pour l’âge et le sexe de l’IOA, la probabilité d’être trié g+1 (moins urgent) plutôt que g (plus urgent) était plus élevée pour une femme comparée à un homme pour les scores de triage plus urgents (OR = 0.85, IC95% [0.76, 0.97] pour triage 2 vs 1 et OR = 0.81, IC95% [0.79, 0.83] pour triage 3 vs 2). Inversement, la probabilité d’être trié g+1 plutôt que g était plus élevée pour un homme comparé à une femme pour les scores de triage moins urgents (OR = 1.15, IC95% [1.13, 1.17] pour triage 4 vs 3 et OR = 1.20, IC95% [1.16, 1.24] pour triage 5 vs 4). Ajusté pour l’âge et le sexe du patient, un infirmier avait une probabilité plus faible d’attribuer un score de 3 (le moins spécifique) qu’une infirmière (4 vs 3 OR = 1.18, IC95% [1.15, 1.21], 3 vs 2 OR = 0.97, IC95% [0.94, 0.99]).

Utilisation des LLM pour prédire le score de triage. Le modèle clinique BELGPT-2 a surpassé toutes les méthodes, atteignant une précision de 0.63, un rappel de 0.63, et un micro F1-score de 0.62. Les combinaisons XGBoost/TF-IDF et LightGBM/TF-IDF ont atteint une précision de 0.59 et 0.61, un rappel de 0.57 et 0.59, et un micro F1-score de 0.55 et 0.57, respectivement. Une analyse plus approfondie du F1-score dans chaque classe révèle le problème commun des classes déséquilibrées : toutes les méthodes performent mieux dans les classes majoritaires (c’est-à-dire les niveaux de triage 2, 3 et 4) que dans les classes minoritaires (c’est-à-dire les niveaux 1 et 5). LightGBM/TF-IDF et BELGPT-2 clinique ont atteint des F1-scores de 0.26 et 0.21 (niveau de triage 1), 0.48 et 0.64 (niveau 2), 0.64 et 0.62 (niveau 3), 0.61 et 0.66 (niveau 4), et 0.31 et 0.40 (niveau 5), respectivement.

5 Discussion

Les applications prometteuses de l’IA offrent un potentiel considérable pour l’amélioration des services des urgences hospitalières, notamment en matière de triage (Yu *et al.*, 2022; Kipourgos *et al.*, 2022; Sanchez-Salmeron *et al.*, 2022; Cho *et al.*, 2022; Vantu *et al.*, 2023; Defilippo *et al.*,

2023; Mutegeki *et al.*, 2023; Sax *et al.*, 2023; Gao *et al.*, 2022; Stewart *et al.*, 2023; Zaboli *et al.*, 2024; Meral *et al.*, 2024). Cependant, ces modèles peuvent hériter de biais inhérents, exacerbant potentiellement les disparités en matière de santé. Notre revue met en lumière des disparités dans la priorisation en fonction de l'âge des patients, de l'ethnicité et du statut socio-économique (Portillo *et al.*, 2023; Peitzman *et al.*, 2023; Essa *et al.*, 2023; Martin *et al.*, 2023; Fekonja *et al.*, 2023). Le rôle du sexe/genre des patients et des IOA reste ambigu, probablement en raison de nuances qui peuvent favoriser les femmes dans certaines situations et les désavantager dans d'autres, et influencé par divers facteurs.

Les entretiens avec les IOA révèlent la complexité inhérente au processus de triage, où les IOA expérimentés naviguent avec une efficacité accrue au sein de ces intrications. L'analyse descriptive montre que la probabilité de se voir attribuer des scores de triage plus urgents augmente avec l'âge. De manière intéressante, il apparaît que les hommes ont tendance à recevoir des scores légèrement plus élevés dans les situations où les scores de triage les plus urgents sont considérés, tandis que les femmes obtiennent généralement des scores légèrement plus élevés dans les situations où les scores de triage les moins urgents sont en jeu. Par ailleurs, le genre de l'IOA semble également être associé à la fréquence de chaque niveau de triage, les infirmiers masculins accordant des niveaux d'urgence légèrement plus bas que leurs homologues féminins. Aucune erreur de triage ne peut être déduite à partir de ces disparités, qui pourraient parfaitement correspondre à la réponse attendue. Des variables supplémentaires telles que les scores de douleur et les diagnostics (ultérieurs au processus de triage) pourraient fournir des informations complémentaires.

Enfin, notre étude pilote utilise un LLM pour prédire les scores de triage, le modèle clinique BELGPT-2 obtenant de meilleurs résultats que les méthodes d'apprentissage automatique plus classiques. Cependant, les résultats demeurent modestes. Les biais et les erreurs potentiels dans la prise de décision des IOA, implicites dans les données d'entraînement et de test, peuvent impacter les performances du modèle. L'hébergement des données en local limite les options, mais les récents LLM basés sur des Transformers en open-source offrent compétitivité et capacités de déploiement interne, ouvrant ainsi des perspectives d'amélioration. Le modèle de prédiction des scores de triage sert de preuve de concept pour une application future.

6 Vers un Protocole Expérimental de Détection des Biais

Les méthodes traditionnelles visant à évaluer les biais dans des domaines tels que l'éducation, ou à mesurer les discriminations dans des disciplines comme la sociologie ou le management, impliquent souvent la manipulation des noms (masculins/féminins, dénotant une origine, etc.) (Doornkamp *et al.*, 2022). Selon l'observatoire des inégalités [3], la méthode privilégié pour détecter les discriminations dans des situations réelles, comme une embauche, par l'expérience pratique est le test de situation (ou *testing*), qui consiste à comparer les résultats de deux groupes de candidats identiques en tous points, sauf pour une caractéristique spécifique. L'organisation du *testing* est lourde et coûteuse, et les résultats spécifiques à un moment et à un endroit donnés, rendant difficile la reproductibilité ainsi que la généralisation.

Des méthodes proches reposant sur une automatisation des tâches à l'aide de modèles de TAL ont été adoptées afin de palier à ces problèmes. La littérature présente ainsi des stratégies pour évaluer les biais linguistiques liés au genre dans les lettres de recommandation (Fu *et al.*, 2023), les disparités de

3. https://inegalites.fr/Comment-mesurer-les-discriminations

genre dans les processus de *reviewing* scientifique (Verharen, 2023) ou encore dans les descriptions de postes (Frissen *et al.*, 2022). En réaction à ces biais linguistiques, certaines études proposent d'identifier, puis supprimer le langage genré de l'ensemble des notes cliniques (Minot *et al.*, 2022) et des données juridiques (Bozdag *et al.*, 2023).

Dans le contexte du triage d'urgence, nous proposons une stratégie inspirée de ces approches pour détecter des comportements biaisés. Une esquisse initiale a été avancée dans (Avalos *et al.*, 2024). Nous envisageons d'évaluer la performance prédictive d'un LLM pour le score de triage tout en utilisant également un LLM pour manipuler les caractéristiques individuelles dans les notes cliniques qui peuvent avoir un impact sur le score de triage, telles que le sexe du patient, le sexe de l'IOA, ou même les années d'expérience de l'IOA. Dans le cas du sexe du patient, nous avons élaboré une méthodologie illustrée par la Figure 1. À notre connaissance, cette approche est innovante et n'a pas encore été utilisée pour détecter les biais humains.

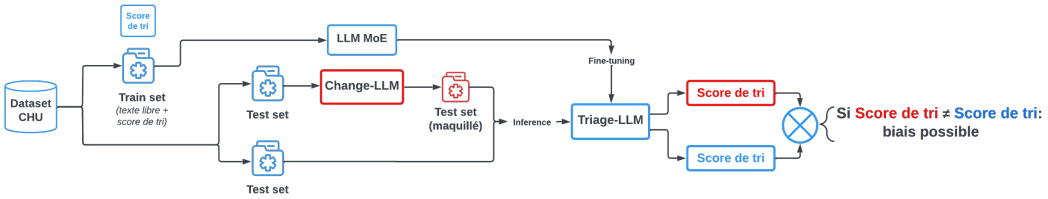

FIGURE 1 – Utilisation des LLM pour évaluer les biais de genre dans les scores de triage aux urgences.

Nous diviserons nos données disponibles en deux parties : l'une servira à l'entraînement du LLM, qui imitera le jugement des IOA en effectuant le triage à partir des notes cliniques et du contexte de l'admission, tandis que l'autre servira à tester la performance du LLM ainsi qu'à réaliser une analyse statistique sur l'impact du sexe du patient sur son score de triage. Cette démarche repose sur notre hypothèse selon laquelle un modèle prédictif, entraîné sur un corpus constitué des notes cliniques d'admission des patients et du contexte dans lequel l'IOA effectue le triage, peut assez fidèlement reproduire le processus de prise de décision des professionnels de santé, comme montré dans notre étude pilote, tout en intégrant les biais cognitifs sous-jacents.

Pour évaluer ces biais, nous comparerons les scores de triage attribués par le modèle entraîné aux notes cliniques originales avec ceux de la version modifiée, où le sexe (ou toute autre variable d'intérêt pouvant refléter l'existence de biais cognitifs dans ce processus décisionnel) est altéré par un LLM. Après avoir prédit les scores de triage pour les deux versions de chaque échantillon, nous pourrons ensuite stratifier en fonction d'autres facteurs décrits dans la littérature comme potentiellement influents pour le triage, tels que l'âge, le sexe de l'IOA, la douleur exprimée, le motif d'admission, voire même l'heure de la journée. Afin de vérifier que ces biais sont intrinsèques aux données apprises et non au choix du modèle, nous pourrions utiliser la même méthode pour transformer la variable modifiée, comme le sexe d'un patient, à son état original, et prédire à nouveau le score de triage qui devrait être attribué à chaque échantillon. Si ces biais sont effectivement liés à la variable d'intérêt, nous devrions observer les mêmes différences dans les scores de triage lors de la transformation de 'retour' que nous avons observées lors de la transformation 'aller' par rapport à la variable modifiée.

Un modèle open-source de 7 milliards de paramètres, tel que Mistral 7B, devrait être suffisamment complexe pour capturer le processus décisionnel nuancé derrière le triage d'urgence. Le choix du modèle chargé de modifier les caractéristiques, telles que le sexe, dans les notes cliniques, peut

quant à lui entraîner des changements dans le flux de travail, comme illustré dans la Figure 1. Les développements récents dans l'évaluation des modèles open-source à mélange d'experts (MoE) suggèrent que des modèles tels que Mixtral 8x22B seront capables d'effectuer les changements suggérés, comme les références au sexe d'un patient en *zero-* ou *few-shot*, éliminant ainsi le besoin d'annoter des paires de notes cliniques (avec et sans modification du sexe des patients) pour ajuster un modèle de triage.

Des limites peuvent néanmoins se manifester. On peut concevoir que des informations omises dans les notes cliniques (portant sur le patient, l'environnement ou la situation de l'IOA) puissent s'avérer cruciales pour nuancer tel ou tel choix qui pourrait, à tort, être identifié comme un biais ou une erreur. Une bonne appréhension des pratiques s'avère indispensable, et cela peut être réalisé, tel que déjà entrepris précédemment, au moyen d'une étude qualitative basée sur des entretiens avec des IOA ou par l'analyse des discours présents dans les textes cliniques. Cependant, après avoir mis en œuvre le flux de travail susmentionné, consistant en un modèle de triage de 7 milliards de paramètres et un modèle MoE pour la transformation du sexe des patients, les résultats préliminaires indiquent que cette méthode est viable pour l'identification des biais liés au sexe des patients. Une évaluation rigoureuse des résultats est actuellement en cours.

Remerciements

Ce travail a été mené sous l'égide de la Phase I du projet TARPON (*Traitement Automatique des Résumés des Passages aux Urgences Pour un Observatoire National*) du Centre de recherche INSERM Bordeaux Population Health, en collaboration avec le Centre Hospitalier Universitaire de Bordeaux.

Références

ARNAUD E., ELBATTAH M., AMMIRATI C., DEQUEN G. & 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. DOI : 10.3390/ijerph19159667.

ARSLANIAN-ENGOREN C. (2000). Gender and age bias in triage decisions. *J Emerg Nurs*, **26**(2), 117–124. DOI : 10.1016/S0099-1767(00)90053-9.

AUBRION A., CLANET R., JOURDAN J., CREVEUIL C., ROUPIE E. & MACREZ R. (2022). FRENCH versus ESI : Comparison between two nurse triage emergency scales with referent scenarios. *BMC Emerg Med*, **22**, 201. DOI : 10.1186/s12873-022-00752-z.

AVALOS M., COHEN D., RUSSON D., DAVIDS M., DOREMUS O., CHENAIS G., TELLIER E., GIL-JARDINÉ C. & LAGARDE E. (2024). Detecting human bias in emergency triage using LLMs : Literature review, preliminary study, and experimental plan. *The International FLAIRS Conference Proceedings*, **37**(1).

BANCO D., CHANG J., TALMOR N., WADHERA P., MUKHOPADHYAY A. & 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. DOI : 10.1161/JAHA.121.024199.

BOZDAG M., SEVIM N. & KOÇ A. (2023). Measuring and mitigating gender bias in legal contextualized language models. *ACM Trans Knowl Discov Data*. DOI : 10.1145/3628602.

BUSLON N., CORTES A., CATUARA-SOLARZ S. & CIRILLO, D R. M. (2023). Raising awareness of sex and gender bias in artificial intelligence and health. *Front Glob Womens Health*, **4**, 970312. DOI : 10.3389/fgwh.2023.970312.

CHENAIS G., LAGARDE E. & GIL-JARDINE C. (2023). Artificial intelligence in emergency medicine : Viewpoint of current applications and foreseeable opportunities and challenges. *Med Internet Res*, **25**, e40031. DOI : 10.2196/40031.

CHO A., MIN I. K., HONG S., CHUNG H. S., LEE H. S. & 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. DOI : 10.2196/39892.

COISY F., OLIVIER G., AGERON F.-X., GUILLERMOU H., ROUSSEL M., BALEN F., GRAU-MERCIER L. & 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*. DOI : 10.1097/MEJ.0000000000001113.

DEFILIPPO A., BERTUCCI G., ZURZOLO C., VELTRI P. & GUZZI P. (2023). On the computational approaches for supporting triage systems. *Interdiscip Med*, **1**(3), e20230015. DOI : 10.1002/INMD.20230015.

DOORNKAMP L., VAN DER POL L., GROENEVELD S., MESMAN J., ENDENDIJK J. & GROENEVELD M. (2022). Understanding gender bias in teachers' grading : The role of gender stereotypical beliefs. In *Teach Teach Educ*, volume 118, p. 103826. DOI : 10.1016/j.tate.2022.103826.

EMAMI P. & 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. DOI : 10.22037/aaem.v11i1.2097.

ESSA C., VICTOR G., KHAN S., ALLY H. & KHAN A. (2023). Cognitive biases regarding utilization of emergency severity index among emergency nurses. *Am J Emerg Med*, **73**, 63–68. DOI : 10.1016/j.ajem.2023.08.021.

FEKONJA Z., KMETEC S., FEKONJA U., MLINAR RELJIĆ N., PAJNKIHAR M. & STRNAD M. (2023). Factors contributing to patient safety during triage process in the emergency department : A systematic review. *J Clin Nurs*, **32**, 5461–5477. DOI : 10.1111/jocn.16622.

FRISSEN R., ADEBAYO K. & NANDA R. (2022). A machine learning approach to recognize bias and discrimination in job advertisements. *AI Soc*, **38**, 1–14. DOI : 10.1007/s00146-022-01574-0.

FU S., CALLEY D., RASMUSSEN V., HAMILTON M., LEE C., KALLA A. & LIU H. (2023). Gender-based language differences in letters of recommendation. In *AMIA Jt Summits Transl Sci Proc*, p. 196–205.

GAO Z., QI X., ZHANG X., GAO X., HE X., GUO S. & LI P. (2022). Developing and validating an emergency triage model using machine learning algorithms with medical big data. *Risk Manag Healthc Policy*, **15**, 1545–1551. DOI : 10.2147/RMHP.S355176.

GORICK H. (2022). Factors that affect nurses' triage decisions in the emergency department : A literature review. *Emerg Nurse*, **30**(3), 14–19. DOI : 10.7748/en.2022.e2123.

HINSON J. S., MARTINEZ D. A., CABRAL S., GEORGE K., WHALEN M., HANSOTI B. & LEVIN S. (2019). Triage performance in emergency medicine : A systematic review. *Ann Emerg Med*, **74**(1), 140–152. DOI : 10.1016/j.annemergmed.2018.09.022.

KIPOURGOS G., TZENALIS A., KOUTSOJANNIS C. & HATZILYGEROUDIS I. (2022). An artificial intelligence based application for triage nurses in emergency department, using the emergency severity index protocol. In *Int J Caring Sci*, volume 15, p. 1764.

KOTEK H., DOCKUM R. & SUN D. (2023). Gender bias and stereotypes in large language models. p. 12–24. DOI : 10.1145/3582269.3615599.

LEE S. & LEE Y. (2020). Improving emergency department efficiency by patient scheduling using deep reinforcement learning. *Healthc*, **8**(2), 77. DOI : 10.3390/healthcare8020077.

LIN P., ARGON N., CHENG Q., EVANS C., LINTHICUM B. & LIU Y. (2022). Disparities in emergency department prioritization and rooming of patients with similar triage acuity score. *Acad Emerg Med*, **29**(11), 1320–1328. DOI : 10.1111/acem.14598.

LIVENTSEV V., HÄRMÄ A. & PETKOVIĆ M. (2021). Towards effective patient simulators. *Front Artif Intell Appl*, **4**, 798659. DOI : 10.3389/frai.2021.798659.

LOPEZ R., SNAIR M., ARRIGAIN S., SCHOLD J., HUSTEY F. & WALKER L. (2021). Sex-based differences in timely emergency department evaluations for patients with drug poisoning. *Public Health*, **199**, 57–64. DOI : 10.1016/j.puhe.2021.08.011.

LOUIS A. (2020). BelGPT-2 : a GPT-2 model pre-trained on french corpora.

MARTIN S., HEYMING T., KAIN A., KRAUSS B. & CAMPOS B. (2023). Eliminating pain disparities for children in the emergency department. *Acad Emerg Med*, **30**(10), 1075–1077. DOI : 10.1111/acem.14723.

MERAL G., ATES S., GUNAY S., OZTURK A. & KUSDOGAN M. (2024). Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. *The American Journal of Emergency Medicine*, **81**, 146–150.

MINOT J. R., CHENEY N., MAIER M., ELBERS D. C., DANFORTH C. M. & DODDS P. S. (2022). Interpretable bias mitigation for textual data : Reducing genderization in patient notes while maintaining classification performance. In *ACM Trans Comput Healthcare*, volume 3. DOI : 10.1145/3524887.

MNATZAGANIAN G., HILLER J., BRAITBERG G., KINGSLEY M., PUTLAND M. & BISH M. (2020). Sex disparities in the assessment and outcomes of chest pain presentations in emergency departments. *Heart*, **106**(2), 111–118. DOI : 10.1136/heartjnl-2019-315667.

MUTEGEKI H., NAHABWE A., NAKATUMBA-NABENDE J. & MARVIN G. (2023). Interpretable machine learning-based triage for decision support in emergency care. In *7th International Conference on Trends in Electronics and Informatics*, p. 983–990. DOI : 10.1109/ICOEI56765.2023.10125918.

ONAL E., KNIER K., HUNT A., KNUDSEN J., NESTLER D. & 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. DOI : 10.1002/emp2.12792.

PEITZMAN C., CARRERAS TARTAK J., SAMUELS-KALOW M., RAJA A. & MACIAS-KONSTANTOPOULOS W. (2023). Racial differences in triage for emergency department patients with subjective chief complaints. *West J Emerg Med*, **24**(5), 888–893. DOI : 10.5811/westjem.59044.

PILIUK K. & TOMFORDE S. (2023). Artificial intelligence in emergency medicine. a systematic literature review. *Int J Med Inform*, **180**, 105274. DOI : 10.1016/j.ijmedinf.2023.105274.

PORTILLO E., REES C., HARTFORD E., FOUGHTY Z., PICKETT M., GUTMAN C., SHIHABUDDIN B., FLEEGLER E., CHUMPITAZI C., JOHNSON T., SCHNADOWER D. & SHAW K. (2023). Research priorities for pediatric emergency care to address disparities by race, ethnicity, and language. *JAMA Netw Open*, **6**(11), e2343791. DOI : 10.1001/jamanetworkopen.2023.43791.

PRECIADO S., SHARP A., SUN B., BAECKER A., WU Y. & LEE M. (2021). Evaluating sex disparities in the emergency department management of patients with suspected acute coronary syndrome. *Ann Emerg Med*, **77**(4), 416–424. DOI : 10.1016/j.annemergmed.2020.10.022.

ROSEMARIN H., ROSENFELD A. & KRAUS S. (2019). Emergency department online patient-caregiver scheduling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 10013–10014. DOI : 10.1609/aaai.v33i01.3301695.

SANCHEZ-SALMERON R., GOMEZ-URQUIZA J., ALBENDIN-GARCIA L., CORREA-RODRIGUEZ M., MARTOS-CABRERA M., VELANDO-SORIANO A. & SULEIMAN-MARTOS N. (2022). Machine learning methods applied to triage in emergency services : A systematic review. *Int Emerg Nurs*, **60**, 101109. DOI : 10.1016/j.ienj.2021.101109.

SAX D., WARTON E., SOFRYGIN O., MARK D., BALLARD D., KENE M., VINSON D. & 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. DOI : 10.1002/emp2.13003.

STEWART J., LU J., GOUDIE A., ARENDTS G., MEKA S., FREEMAN S., WALKER K., SPRIVULIS P., SANFILIPPO F., BENNAMOUN M. & DWIVEDI G. (2023). Applications of natural language processing at emergency department triage : A narrative review. *PLOS One*, **18**, e0279953. DOI : 10.1371/journal.pone.0279953.

SUAMCHAIYAPHUM K., JONES A. & MARKAKI A. (2023). Triage accuracy of emergency nurses : An evidence-based review. *J Emerg Nurs*, **158**(1767), 00251–9. DOI : 10.1016/j.jen.2023.10.001.

TAYLOR A., MURAKAMI M., KIM S., CHU R. & RIEK L. D. (2022). Hospitals of the future : Designing interactive robotic systems for resilient emergency departments. *Proc ACM Hum-Comput Interact*, **6**. DOI : 10.1145/3555543.

VAN DER STIGCHEL B., VAN DEN BOSCH K., VAN DIGGELEN J. & P H. (2023). Intelligent decision support in medical triage : Are people robust to biased advice ? *J Public Health*, **45**(3), 689–696. DOI : 10.1093/pubmed/fdad005.

VANTU A., VASILESCU A. & BAICOIANU A. (2023). Medical emergency department triage data processing using a machine-learning solution. *Heliyon*, **9**(8), e18402. DOI : 10.1016/j.heliyon.2023.e18402.

VERHAREN J. P. H. (2023). ChatGPT identifies gender disparities in scientific peer review. *eLife*, **12**, RP90230. DOI : 10.7554/eLife.90230.3.

VIGIL J., COULOMBE P., ALCOCK J., STITH S., KRUGER E. & CICHOWSKI S. (2017). How nurse gender influences patient priority assignments in US emergency departments. *Pain*, **158**(3), 377–382. DOI : 10.1097/j.pain.0000000000000725.

YU J. Y., XIE F., NAN L., YOON S., ONG M. E. H., NG Y. Y. & 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. DOI : 10.1038/s41598-022-22233-w.

ZABOLI A., BRIGO F., SIBILIO S., MIAN M. & TURCATO G. (2024). Human intelligence versus Chat-GPT : who performs better in correctly classifying patients in triage ? *Am J Emerg Med*, **79**, 44–47.