

Gender bias in AI-based model according to the sex composition of the training sets: the case of a model triaging patients requiring referral to hospital emergencies among those seeking emergency or unscheduled care by phone

Laurent RIGAL

`laurent.rigal@universite-paris-saclay.fr`

Université Paris-Saclay

Camille REVOL

e-sensia

Thibault BACQUEYRISSES

e-sensia

Henri PANJO

Université Paris-Saclay

Urfan ASHRAF

SOS Médecins Grand Paris

Jean-Baptiste PERNEY

SOS Médecins Grand Paris

Research Article

Keywords: Emergency Medicine, Gender bias, Emergency Medical Dispatch, Artificial Intelligence

Posted Date: October 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10849632/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. e sensia (<https://e-sensia.com>) is a recently founded company employing several of the authors (LR, CR, TB and JBP). Its mission is to improve

emergency medical dispatch through the development of AI based models. SOS Médecins Grand Paris (of which UA is president) holds a financial stake in e-sensia. HP has no completing interest to declare.

Abstract

Background

Gender differences in emergency care are well documented and generally disadvantage women, including at the triage stage. AI-based triage tools may reproduce or amplify these differences. We previously developed a model identifying, among individuals seeking emergency or unscheduled care through a medical call centre, those requiring referral to an emergency department (ED), and observed better discrimination in men than in women. This study aimed to analyse the effect of the sex composition of the training data on this gender difference in discriminatory performance.

Methods

We used telephone calls (from 2018 to 2023) and medical records from *SOS Médecins Grand Paris* (France), which provides same-day home visits for patients. At the end of each visit, the doctor records the patient's outcome (left at home *versus* referred to EDs; 3% referrals) in the patient's medical file. Calls were transcribed using an automatic speech recognition model and then processed using a text-based model to classify whether the patient required ED attendance. From sex-balanced training, validation and test subsets ($n=601\ 560$), three subsets of identical size ($n=300\ 780$) were derived - female, male, and sex-balanced-reduced-by-half - and used to train three models. Discrimination was compared by sex using receiver operating characteristic area under the curve (ROC-AUC).

Results

Of 1 876 172 calls followed by a home visit, 1 098 748 (60%) were matched with their medical record. Each model was trained on 164 980 calls (10% referrals) and evaluated on 67 900 validation and 67 900 test calls (3% referrals). The female model achieved a ROC-AUC of 0.812 (95% confidence intervals (CI) 0.803–0.820) among women, and the male model 0.835 (95%CI 0.827–0.843) among men. For all three models, discrimination was higher among men than among women, significantly so for the male and the sex-balanced-reduced-by-half models, but not for the female model.

Conclusions

Balancing training sets by sex is neither a necessary nor a sufficient condition for fairness. A male-only model generalised poorly to women, whereas a female-only model appeared applicable to both sexes, suggesting structural differences that warrant placing fairness considerations at the design stage rather than treating them as secondary.

Trial registration

ClinicalTrials.gov Identifier: NCT07428109 (registered on 23 February 2026)

INTRODUCTION

The existence of systematic gender differences in medical care, which cannot be explained by scientific factors and which, in most cases, result in women receiving lower-quality care than men, is well documented [1]. In the field of cardiology and psychiatry, these gender biases have long been identified and analysed [2–4]. In cases of acute non-traumatic chest pain, for example, women are perceived by emergency medical staff as being in a less urgent situation than men [5]. As a result, they wait longer before receiving an initial medical evaluation and an electrocardiogram. These types of gender differences, which recur at every subsequent stage of care, ultimately lead to unfair disparities in post-heart attack mortality [6]. However, gender biases in emergency settings [7] extend beyond the cardiovascular field alone. Women also receive less appropriate care in cases of severe trauma [8]. They are prescribed fewer pain relievers and receive them later when experiencing pain [9]. Furthermore, with regard to triage, a critical initial step that can have clinical consequences due to delays in initiating appropriate emergency care, differences in Emergency Severity Index (ESI) scores [10] based on sex remain, even after adjusting for age, clinical factors, and department operating conditions [11].

Moreover, hospital emergency departments (EDs) have been experiencing increasing levels of congestion [12] for several decades. During periods of extreme pressure, such as those observed in France during summer months, access to EDs is restricted to patients who have received prior authorisation to attend [13]. New approaches are being explored within medical dispatch systems to manage requests for urgent or unscheduled care more effectively. Telephone triage could become an essential step in medical dispatch before ED access [14]. However, rapidly assessing patients without visual contact - particularly when they may be emotionally distressed or face language barriers - is notably challenging. Several triage algorithms, such as the ESI [10], are available to assist telephone operators [15], but they rely on structured clinical information that is not easily or quickly accessible during calls [16].

In recent years, artificial intelligence (AI) has emerged as a promising alternative [17] to support operators [18], enabling the processing of large volumes of unstructured data, particularly audio exchanges [18]. AI-based classification models using audio data have shown potential value in medical dispatch [19], especially for recognising conditions such as cardiac arrest [20, 21], stroke [22] or myocardial infarction.

To assist medical triage operator, we have developed an AI-based model triaging patients requiring referral to hospital emergencies among those seeking emergency or unscheduled care by phone. Beyond satisfactory performance, the analysis for gender bias revealed that the model exhibited greater discriminatory capabilities among men than among women [reference to come]. Analysing gender bias in AI systems is an important step, as it could lead to maintain or even reproduce the gender biases observed in healthcare. The literature identifies several sources of bias: those related to data, those related to algorithmic design, and those related to human use. This study examines biases that occur when the data used to train AI models are unrepresentative or incomplete. This is the case, for example, with a system trained to recognise melanomas from photographs of white individuals, which performs poorly on people with other skin tones [23]. To analyse further the gender bias observed in our triaging

model, we examine here the implications that an imbalance in the data by sex might have on the nature and extent of this bias. In other words, our study aimed to analyse the effect of the training data sex-imbalance on gender-difference in models discriminatory performances.

METHODS

Design and setting

This study used retrospective observational data from *SOS Médecins Grand Paris* to develop AI-based models using training sets with different sex ratios and analyse their discriminatory performances to identify patients requiring referral to EDs.

SOS Médecins Grand Paris is a group of approximately 150 general practitioners and emergency doctors offering same-day home visits in Paris and its surrounding areas (over 6.5 million inhabitants). Around sixty similar *SOS Médecins* groups exist in major French cities, where they constitute the primary option for non-life-threatening conditions that cannot wait until a general practitioner is available and for patients wishing to avoid attending the ED. People call a single telephone number and speak first with a medical dispatch assistant, who clarifies the reason for the call, transfers the caller to a dispatch physician or to emergency medical services (EMS) if necessary, and collects contact details so that a doctor can be sent to their home.

In this work, we focused solely callers who ultimately received a home visit. If no doctor is available, callers are advised to contact EMS for guidance on the appropriate medical response. Unlike EMS, *SOS Médecins Grand Paris* does not send ambulances or refer patients directly to ED.

Primary endpoint

At the end of each home visit, the physician recorded the patient's outcome, which defined the binary primary endpoint: left at home *versus* referred to a general or specialised cardiology or neurology emergency department (for example, in cases of acute coronary syndrome or suspected stroke).

Dataset matching

Patients' medical records and telephone exchanges were not initially matched. For each patient who contacted *SOS Médecins Grand Paris* during the period from 1 January 2018 to 31 December 2023 and received a home visit from a doctor, we attempted to match the patient's medical record with the telephone exchanges corresponding to the initial call. To do this, we first transcribed all calls. Then, we attempted to identify the call corresponding to each patient medical record. A match was validated if either (i) the telephone number was identical or (ii) the name and date of birth were the same.

Machine Learning pipeline

We employed a two-step machine learning pipeline. First, all calls were transcribed using a speech recognition model. To obtain these call transcripts we used Whisper large-v3-turbo with the Whisper-X backend [OpenAI, San Francisco]. Whisper large-v3-turbo is a fine-tuned version of the pruned Whisper large-v3 model for automatic speech recognition. Second, the CamemBERT model (a BERT-type language model pre-trained on French texts) served as the foundation for training our classification model. For this task, we used the first 512 tokens of each call transcript as input.

The three models and their training validation and test subsets

To construct sex-balanced (training validation and test) subsets (Fig. 1), every step was taken to maintain a sex ratio of one (i.e. an equal number of men and women). Of all medical records successfully matched to their corresponding calls that resulted in referral to an ED following the home visit: 80% were used for training, 10% for validation, and 10% for testing (random allocation). To address class imbalance (only 4.5% of calls with a home visit lead to a transfer to hospital), we constructed the training subset by randomly selecting 10% of ED referrals, and supplemented it with a random sample of matched calls that did not lead to ED referral so as to obtain a total of 90% non-referral cases. For the validation and test subsets (both preserving the real-life prevalence of 3% ED referrals post-call with and without home visit), we applied the same sampling strategy without rebalancing. To avoid any data leakage (situation where information normally not available is introduced into the data and can lead to bias), all calls from the same patient were assigned to the same subset (training, validation or test).

Based on these sex-balanced subsets ($n = 601\ 560$), three additional subsets of equal size ($n = 300\ 780$) were constructed: the female and male subsets, corresponding to the simple separation of women and men, and the sex-balanced reduced by half subset, corresponding to the random selection of every second person from the sex-balanced subsets.

To build the three models, called as the female, male and sex-balanced reduced by half models, from the three preceding subsets of the same name, we used focal loss, which emphasises samples that are difficult to classify. Finally, the models were calibrated.

Statistical analysis

First, we described the different subsets used according to age (0–2 years, 3–17, 18–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80 and more), sex (binary, patient-reported) and social deprivation index in tertiles (of the population served by *SOS Médecins Grand Paris*) [24].

Second, we presented ROC (receiver operating characteristic) curve [25, 26], along with their respective areas under the curve (AUC), for the three models. We computed 95% confidence intervals (CIs). To compare sex-discriminatory performances (on ROC-AUC), we performed statistical tests using

bootstrapping, a method that estimates the underlying distribution of a population from a single sample of observations, thereby enabling statistical tests to be performed. Third, to address model explainability [27, 28] and better understand the determinants of the classifier's predictions, we performed an occlusion analysis. This method consists of removing all instances of a given word from the input transcript to evaluate its impact on the model's outputs. First, we selected the words to analyse. To do this, we vectorised the transcripts and retained the top 5 000 terms with the highest overall TF-IDF (Term Frequency-Inverse Document Frequency) scores. TF-IDF quantifies the importance of a word by combining its frequency within a transcript with its rarity across the entire corpus. We then performed a Student's t-test for each of these terms to determine whether TF-IDF values differed significantly between calls that resulted in ED referral and those that did not. We retained the 200 words with the highest absolute t-statistic (all of these had a $p < 10^{-10}$). Next, for each of the selected words, we removed all its occurrences from the corpus and regenerated model predictions. For each transcript, we then compared the prediction scores before and after occlusion to quantify the influence of the removed word. We carried out these occlusion analyses for the female and male models using the sex-balanced test subsets of each sex. All analyses were realized with Python (code available on reasonable request) in Cursor (Anysphere Inc, San Francisco) development environment.

Ethical approval

The *SOS Médecins Grand Paris*' databases were declared to the National Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés*) which is responsible for ethical oversight and for protecting individuals from inappropriate or unlawful data collection practices. The institutional ethics committee (*Comité d'Ethique du Collège National des Généralistes Enseignants*, IRB00010804) approved the study (decision n°131125905) which was conducted in accordance with the Declaration of Helsinki and registered at ClinicalTrials.gov (NCT07428109) on 23 February 2026.

Results

Population characteristics

Between 1 January 2018 and 31 December 2023, we identified 1 876 172 calls that led to a home visit (Figure 1). Among these calls, 1 098 748 (almost 60%) were successfully matched with their corresponding medical files. Two successive steps were used to form three subsets called 'sex-balanced' because they contained the same number of men and women: training (comprising 10% of ED referrals), validation, and test (both comprising the real-life proportion of 3% of ED referrals,).

Table 1 presents the description of the sex-balanced subsets (training, validation and test subsets combined) depending on whether or not the call resulted in a referral to EDs. Regardless of what happened to them following the home visit, the men were significantly younger and the women older. Men who were not referred to ED after the doctor's visit were more likely than women to live in a deprived area.

Three models

We have built and evaluated three models using training, validation, and test subsets of the same size (respectively 164 980, 67 900 et 67 900), and with identical hospitalization rates (respectively 10%, 3% and 3%). These models were called female, male and sex-balanced reduced by half models.

Main results

The female model achieved a ROC-AUC among women of 0.812 (95% CI 0.803-0.820) and the male model among men of 0.835 (95% CI 0.827-0.843, Table 2). Regardless of the model considered, the ROC-AUCs were higher among men than among women. The ROC-AUC values among men were significantly higher than those among women for the male and sex-balanced reduced by half models (but not for the female model).

The ROC-AUC value among women did not differ between the female and male models ($p=0.70$), but they did among men ($p=0.01$).

Model explainability

The occlusion analysis identified words with the greatest positive influence on the classifier predictions (Table 3). The ranking of the terms most commonly associated with ED referrals differs between the female and male models. Whilst terms relating to falls or shortness of breath top the list, the next two terms are those relating to pain and the abdomen for the female model (among both sexes), and those relating to blood and dizziness for the male model (among both sexes). Furthermore, regardless of the model (female or male), the rankings are similar between the two sexes (with a maximum difference of one place). When comparing the sex-differences in the influences (differences in prediction scores before and after occlusion) between the male and female models, these differences are (with the exception of the terms relating to falls, where the differences are very similar) greater (in absolute terms) in the male model than in the female model.

Discussion

Summary

The results showed that a male model (i.e. one built solely from male data) led to the greatest gender bias against women. In contrast, a female model resulted in no (statistically significant) bias. The balanced model exhibited a bias that fell between the two previous models. Furthermore, an analysis of the explainability of the female and male models revealed that certain terms did not carry the same predictive weight for the two genders.

Interpretation

When a model built using an unbalanced dataset containing fewer women than men (the 'male model' represents the extreme situation) performs less favourably for women, two main explanations may be considered. First, this may be due to a composition effect within the dataset: if women are under-represented, it is possible that predictions will be less accurate due to increased variance [29]. This disparity decreases when female examples (the minority group) are added. Second, it may be due to a qualitative or structural difference (distribution shift): the relationships between the features and the outcome are not the same in the two sexes. Given identical presentation, the probability of using EDs differs. A single model is structurally misspecified. It strikes a compromise that automatically favours men (the majority group). Our experimental design, which alters the gender composition of the training subset, allows us to distinguish between these two explanations.

Our results confirm that a representative dataset is not a sufficient condition for ensuring a model's performance is fair (the case of the 'sex-balanced reduced by half model'). Nor is it a necessary condition; a model trained on unbalanced data does not necessarily result in bias against the minority group (the case of the 'female model').

The asymmetry between the biases observed in female and male models shows that a model developed for men cannot be generalised to women, whereas a model developed for women appears to be applicable to men. Gender-differences in chest pain could shed light to this finding. Chest pain is predominantly typical in both sexes, but is more common in men than in women, for whom the presentation is more varied [30]. Similarly, in the event of a stroke, women report more general symptoms and fewer localised signs than men [31]. These two examples illustrate that, in certain situations, the female model can be extrapolated to men, but not the other way round.

Comparison with existing literature

Despite recommendations calling for the presentation of stratified performance analyses to better assess whether models may be biased [32, 33], to the best of our knowledge, there are no articles that have analysed gender bias in AI-based telephone triage for ED referral. The literature in this field is dominated by imaging, dermatology, sepsis and the prediction of readmission. A systematic review focusing on the prediction of sepsis risk found that only 2% of studies had carried out stratified analyses by sex, and none had reported fairness metrics [34].

Strengths and limitations

Our study has several limitations. First, the gender bias could originate elsewhere in the pipeline. To begin with, transcription software performs better with male voices than female ones [35]. This could contribute to the observed bias. Second, there could be a bias in the labelled data. The outcome is not a neutral gold standard but a human decision that may already be biased. Although the decision to refer patients to EDs was made by a physician who examined the patient, it was not based on a standardised objective measure such as a prognostic score. We do not have any subsequent clinical outcomes such as mortality or emergency room outcomes, which would have been more relevant for prediction

purposes. On this point, however, it should be noted that the frequency of use of EDs is the same for both sexes. Third, AUC-ROC is not a metric for operational equity. Analysing gender disparities in terms of actual care provision would be more relevant, but this requires setting a decision threshold, which has not yet been done given the current stage of the model's development. Fourth, the study was carried out using data from a single centre and is therefore dependent on the case mix and on the use of local EDs. Its generalisability is limited. Fifth, other grounds for discrimination, such as social status, would be worth analysing, as would the presence of any potential interactions (gender inequality may be concentrated amongst disadvantaged people). Sixth, the gender differences observed are significant but small in absolute terms and are likely to be of limited clinical significance. Seventh, gender (identity, social roles) is not measured, even though it is plausibly the causal mediator of the phenomenon, rather than (biological) sex [36].

The main strength of our study is its originality. Research into gender bias in telephone triage is scarce. Furthermore, this study goes beyond purely observational studies analysing gender differences. The experimental design of our study enabled us to isolate the impact of the gender composition of the training subset on gender bias.

Implications for research and practice

First, algorithmic fairness measured using retrospective data is an indicator of the model's intrinsic properties, but does not guarantee the fairness of a socio-technical system operating under real-world conditions. Only a prospective evaluation (or a human-machine interaction study, such as randomised vignettes with and without score display, stratified by patient gender) will enable the fairness of the deployed system to be measured. Second, a training sample that is unbalanced is not necessarily disadvantageous to the majority group, although this still needs to be demonstrated. We therefore agree with the recommendations which emphasise the need to systematically compare metrics broken down by gender, rather than just overall performance figures [37]. Thirdly, to limit gender bias, there are alternatives to pre-processing (rebalancing, reweighting) or post-processing (group-specific thresholds). During training (in-processing), it is possible to use fairness-aware regularisation, which allows a fairness constraint to be incorporated directly into the optimised objective function. Fourth, in social epidemiology, all statistical analyses are stratified by gender. Researchers in this field believe that, as a matter of principle (and therefore until proven otherwise), things do not unfold in the same way for both sexes. It might be worth considering this assumption in certain situations where gender biases are known to exist, as in emergency medicine.

Conclusions

Our study of gender bias based on the sex composition of the training sets has led to two conclusions. First, balancing the training sets is neither a necessary nor a sufficient condition for a model to be fair. Second, gender bias must be investigated for ethical reasons (it is critical to ensuring that emerging technologies do not reinforce existing health inequalities), but it may also lead to a change in modelling

strategy, particularly where differences of a fundamental nature between the two sexes are identified. This final point leads to placing ethical considerations at the forefront of the design process, rather than treating them as secondary concerns.

Declarations

Ethics approval and consent to participate

SOS Médecins Grand Paris' databases have been declared to the National Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés*) which oversees ethical considerations and protects individuals from inappropriate or unlawful data collection practices. The institutional ethics committee (*Comité d'Ethique du Collège National des Généralistes Enseignants*, IRB00010804) approved the study (decision n°131125905). No consent to participate was required.

Consent for publication

No consent for publication was required.

Availability of data and materials

Anonymised datasets are available upon reasonable request.

Competing interests

e-sensia (<https://e-sensia.com>) is a recently founded company employing several of the authors (LR, CR, TB and JBP). Its mission is to improve emergency medical dispatch through the development of AI-based models. *SOS Médecins Grand Paris* (of which UA is president) holds a financial stake in e-sensia. HP has no completing interest to declare.

Funding

This study was supported by a grant from the French Ministry of Health (AAP RESPIR-23-023)

Authors' contributions

LR conceived the study. LR, HP, UA and JBP designed the study. CR and TB analysed the data. All authors interpreted the data. All authors interpreted the findings and contributed to drafting the manuscript. LR is the guarantor.

Acknowledgements

The authors would like to thank Stefan Barbary for his assistance in accessing and interpreting data from *SOS Médecins Grand Paris*.

References

1. Hamberg K. Gender bias in medicine. *Womens Health (Lond)*. 2008;4:237–43. <https://doi.org/10.2217/17455057.4.3.237>.
2. Pope JH, Aufderheide TP, Ruthazer R, Woolard RH, Feldman JA, Beshansky JR, et al. Missed diagnoses of acute cardiac ischemia in the emergency department. *N Engl J Med*. 2000;342:1163–70. <https://doi.org/10.1056/NEJM200004203421603>.
3. Healy B. The Yentl Syndrome. *New England Journal of Medicine*. 1991;325:274–6. <https://doi.org/10.1056/NEJM199107253250408>.
4. Bacigalupe A, Martín U, Triolo F, Sjöberg L, Sterner TR, Dekhtyar S, et al. Is the diagnosis and treatment of depression gender-biased? Evidence from a population-based aging cohort in Sweden. *Int J Equity Health*. 2024;23:252. <https://doi.org/10.1186/s12939-024-02320-2>.
5. Coisy F, Olivier G, Ageron F-X, Guillerrou H, Roussel M, Balen F, et al. Do emergency medicine health care workers rate triage level of chest pain differently based upon appearance in simulated patients? *Eur J Emerg Med*. 2024;31:188–94. <https://doi.org/10.1097/MEJ.0000000000001113>.
6. Vaccarino V, Parsons L, Every NR, Barron HV, Krumholz HM. Sex-based differences in early mortality after myocardial infarction. National Registry of Myocardial Infarction 2 Participants. *N Engl J Med*. 1999;341:217–25. <https://doi.org/10.1056/NEJM199907223410401>.
7. Safdar B, Ona Ayala KE, Ali SS, Seifer BJ, Hong M, Greenberg MR, et al. Inclusion of Sex and Gender in Emergency Medicine Research-A 2018 Update. *Acad Emerg Med*. 2019;26:293–302. <https://doi.org/10.1111/acem.13688>.
8. Ghika-Nanchen A, Marzorati L, Merra A, Girardello C, Truong P, Carron PN, et al. Sex and gender bias in major trauma care: a scoping review. *Scand J Trauma Resusc Emerg Med*. 2026;34:84. <https://doi.org/10.1186/s13049-026-01596-3>.
9. Guzikovits M, Gordon-Hecker T, Rekhtman D, Salameh S, Israel S, Shayo M, et al. Sex bias in pain management decisions. *Proc Natl Acad Sci U S A*. 2024;121:e2401331121. <https://doi.org/10.1073/pnas.2401331121>.
10. agency for healthcare research and quality. Emergency severity index implementation handbook. 2012.
11. Patel MD, Lin P, Cheng Q, Argon NT, Evans CS, Linthicum B, et al. Patient sex, racial and ethnic disparities in emergency department triage: A multi-site retrospective study. *Am J Emerg Med*. 2024;76:29–35. <https://doi.org/10.1016/j.ajem.2023.11.008>.
12. Kenny JF, Chang BC, Hemmert KC. Factors Affecting Emergency Department Crowding. *Emergency Medicine Clinics of North America*. 2020;38:573–87. <https://doi.org/10.1016/j.emc.2020.04.001>.
13. Yordanov Y, Penverne Y, Delelis Fanien H, Frappé P, Boissault P, Vrignaud B. Préconisations Régulation préalable à l'accès aux urgences. 2024.
14. Michel J, Manns A, Boudersa S, Jaubert C, Dupic L, Vivien B, et al. Clinical decision support system in emergency telephone triage: A scoping review of technical design, implementation and

- evaluation. *Int J Med Inform.* 2024;184:105347. <https://doi.org/10.1016/j.ijmedinf.2024.105347>.
15. FitzGerald G, Jelinek GA, Scott D, Gerdtz MF. Emergency department triage revisited. *Emerg Med J.* 2010;27:86–92. <https://doi.org/10.1136/emj.2009.077081>.
 16. Farand L, Leprohon J, Kalina M, Champagne F, Contandriopoulos AP, Preker A. The role of protocols and professional judgement in emergency medical dispatching. *Eur J Emerg Med.* 1995;2:136–48.
 17. Wachter RM, Brynjolfsson E. Will Generative Artificial Intelligence Deliver on Its Promise in Health Care? *JAMA.* 2024;331:65–9. <https://doi.org/10.1001/jama.2023.25054>.
 18. Chenais G, Lagarde E, Gil-Jardiné C. Artificial Intelligence in Emergency Medicine: Viewpoint of Current Applications and Foreseeable Opportunities and Challenges. *J Med Internet Res.* 2023;25:e40031. <https://doi.org/10.2196/40031>.
 19. Chee ML, Chee ML, Huang H, Mazzochi K, Taylor K, Wang H, et al. Artificial intelligence and machine learning in prehospital emergency care: A scoping review. *iScience.* 2023;26:107407. <https://doi.org/10.1016/j.isci.2023.107407>.
 20. Blomberg SN, Folke F, Ersbøll AK, Christensen HC, Torp-Pedersen C, Sayre MR, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation.* 2019;138:322–9. <https://doi.org/10.1016/j.resuscitation.2019.01.015>.
 21. Blomberg SN, Christensen HC, Lippert F, Ersbøll AK, Torp-Petersen C, Sayre MR, et al. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Netw Open.* 2021;4:e2032320. <https://doi.org/10.1001/jamanetworkopen.2020.32320>.
 22. Wenstrup J, Havtorn JD, Borgholt L, Blomberg SN, Maaloe L, Sayre MR, et al. A retrospective study on machine learning-assisted stroke recognition for medical helpline calls. *NPJ Digit Med.* 2023;6:235. <https://doi.org/10.1038/s41746-023-00980-y>.
 23. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv.* 2022;8:eabq6147. <https://doi.org/10.1126/sciadv.abq6147>.
 24. Rey G, Jouglu E, Fouillet A, Hemon D. Ecological association between a deprivation index and mortality in France over the period 1997 - 2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC public health.* 2009;9:33. <https://doi.org/10.1186/1471-2458-9-33>.
 25. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem.* 1993;39:561–77.
 26. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* 1988;240:1285–93. <https://doi.org/10.1126/science.3287615>.
 27. Wang Y-C, Chen T-CT, Chiu M-C. A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes. *Healthcare Analytics.* 2023;3:100183. <https://doi.org/10.1016/j.health.2023.100183>.

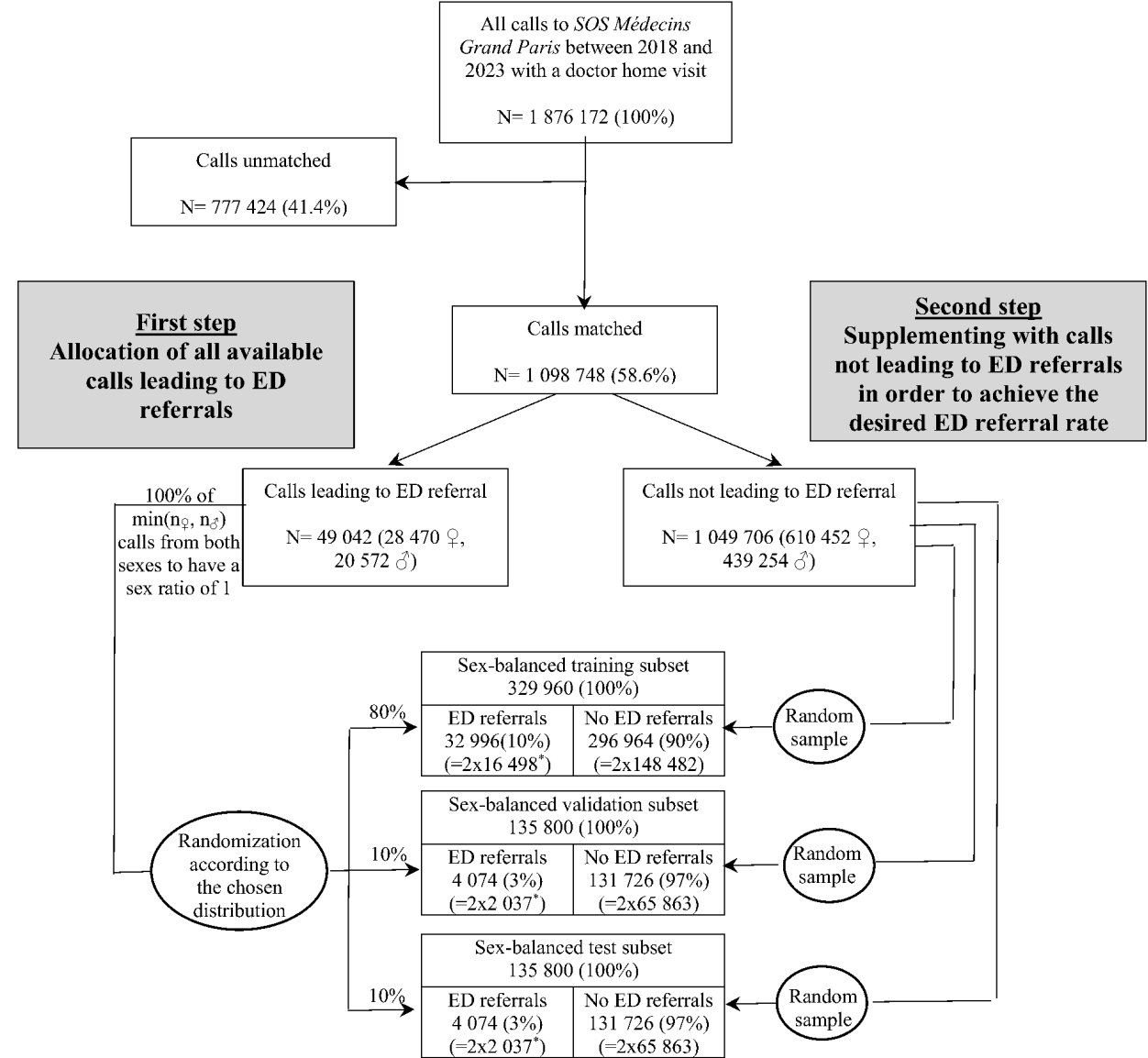
28. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3:e745–50.
[https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
29. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*. 2020;117:12592–4. <https://doi.org/10.1073/pnas.1919012117>.
30. Lapostolle F, Loyeau A, Beggaz Y, Boche T, Pires V, Le Bail G, et al. Effect of age, gender, and time of day on pain-to-call times in patients with acute ST-segment elevation myocardial infarction: the CLOC'AGE study. *Emergencias*. 2021;33:181–6.
31. Ali M, van Os HJA, van der Weerd N, Schoones JW, Heymans MW, Kruijff ND, et al. Sex Differences in Presentation of Stroke: A Systematic Review and Meta-Analysis. *Stroke*. 2022;53:345–54.
<https://doi.org/10.1161/STROKEAHA.120.034040>.
32. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.
33. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med*. 2020;3:81.
<https://doi.org/10.1038/s41746-020-0288-5>.
34. Hauschildt KE, Pan A, Bernstein T, Admon AJ, Mukherjee B, Iwashyna TJ, et al. Consideration of Sociodemographics in Machine Learning-Driven Sepsis Risk Prediction. *Crit Care Med*. 2025;53:e1815–20. <https://doi.org/10.1097/CCM.0000000000006741>.
35. Holliday NR, Reed PE. Gender and racial bias issues in a commercial “tone of voice” analysis system. *PLoS One*. 2025;20:e0314470. <https://doi.org/10.1371/journal.pone.0314470>.
36. Ballering AV, Olde Hartman TC, Rosmalen JGM. Gender scores in epidemiological research: methods, advantages and implications. *Lancet Reg Health Eur*. 2024;43:100962.
<https://doi.org/10.1016/j.lanepe.2024.100962>.
37. Badr J, Motulsky A, Denis J-L. Digital health technologies and inequalities: A scoping review of potential impacts and policy recommendations. *Health Policy*. 2024;146:105122.
<https://doi.org/10.1016/j.healthpol.2024.105122>.

Tables

Tables are available in the Supplementary Files section.

Figures

Figure 1. Flow chart explaining the constitution of the sex-balanced subsets in two steps



* As all calls from a given patient must belong to the same subset (training, validation, or test), these counts are not exactly equal to the expected values.

Figure 1

See image above for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ArticleGenderTablesV2.docx](#)