Fairness in AI-Based Mental Health: Clinician Perspectives and Bias Mitigation

Gizem Sogancioglu[1], Pablo Mosteiro[2], Albert Ali Salah[1], Floortje Scheepers[3], Heysem Kaya[1]*

[1] Department of Information and Computing Sciences, Utrecht University, The Netherlands
[2] Department of Methodology and Statistics, Utrecht University, The Netherlands
[3] Department of Psychiatry, University Medical Center Utrecht, The Netherlands
g.sogancioglu@uu.nl, p.j.mosteiroromero@uu.nl, a.a.salah@uu.nl, f.e.scheepers-2@umcutrecht.nl, h.kaya@uu.nl

Abstract

There is limited research on fairness in automated decision-making systems in the clinical domain, particularly in the mental health domain. Our study explores clinicians' perceptions of AI fairness through two distinct scenarios: violence risk assessment and depression phenotype recognition using textual clinical notes. We engage with clinicians through semi-structured interviews to understand their fairness perceptions and to identify appropriate quantitative fairness objectives for these scenarios. Then, we compare a set of bias mitigation strategies developed to improve at least one of the four selected fairness objectives. Our findings underscore the importance of carefully selecting fairness measures, as prioritizing less relevant measures can have a detrimental rather than a beneficial effect on model behavior in real-world clinical use.

Introduction

In the evolving landscape of mental healthcare, the integration of automated decision-making systems presents both opportunities and challenges. Among these challenges, ensuring fairness in algorithmic decisions is critical, given the potential for these systems to perpetuate or even amplify existing biases. Despite its importance, there is a notable scarcity in the empirical exploration of fairness, specifically in psychiatric predictive models (Şahin et al. 2024).

This paper seeks to bridge this gap by focusing on the concept of (binary) gender fairness within Natural Language Processing (NLP) models in mental health. We analyze two important Electronic Health Records (EHR) applications: the assessment of violence risk and the recognition of depression phenotypes, both through the analysis of textual clinical notes. These scenarios are important in mental health care, where accurate and fair assessments can significantly influence patient treatment paths and outcomes.

Fairness assessment is typically driven by machine learning researchers to ensure their algorithms are fair, and several mathematical or statistical approaches are used to test the fairness of algorithms. While this is surely useful, it misses a crucial aspect. It is important to involve different groups in such assessments, such as domain experts, ethics experts, and lay people (e.g., patients) (World Health Organization 2024). Our exploration begins with an investigation into clinicians' perceptions of fairness in these contexts. Understanding the clinical perspective is essential, as it directly influences the acceptance and effectiveness of automated systems in real-world settings. We conducted semi-structured interviews to gather insights from clinicians, aiming to identify and prioritize quantitative fairness objectives that align with clinical expectations and ethical standards.

Building upon these insights, we compare and evaluate a range of bias mitigation strategies. Each method is designed to address one or more of the quantitative fairness objectives. This comparative analysis is crucial for understanding how different approaches to fairness can influence the behavior and effectiveness of NLP systems in mental health applications. Our results underscore the significance of choosing suitable fairness metrics. We show that emphasizing less pertinent measures can lead to detrimental effects on the model's behavior and, consequently, on fairness in patient care, rather than yielding positive outcomes. This highlights the complex interplay between technical objectives and clinical outcomes in developing fair NLP systems.

We summarize the main contributions of this study as follows:

- We conduct semi-structured interviews with clinicians to gain insights into their perceptions of fairness using two case studies in mental healthcare. We summarize and discuss seven main themes that emerged from our qualitative analysis of the interviews.
- We evaluate various automatic bias mitigation techniques for a set of fairness objectives. This evaluation highlights the importance of involving domain experts in carefully selecting fairness measures as an essential component of the ML development process in the clinical domain. We show that in some cases, even though the methods seem to improve a few fairness dimensions, the most critical

[*] This is the uncorrected author proof. Please cite as: G. Sogancioglu, P. Mosteiro, A.A. Salah, F. Scheepers, H. Kaya, "Fairness in AI-Based Mental Health: Clinician Perspectives and Bias Mitigation", Proc. 7th AAAI/ACM Conference on AI, Ethics, and Society (AIES), San Jose, USA, 2024.

fairness measure chosen by the clinicians is not satisfied.

Fairness of NLP models in mental health

Fairness studies in natural language processing (NLP) can be broadly categorized into two areas: fairness in downstream models and fairness in word embeddings, respectively. Research in these categories has addressed a variety of domain-independent problems (Mathew et al. 2021; Bolukbasi et al. 2016). Although more research is needed, there is a growing recognition of the importance of fairness in clinical NLP tasks as well. Studies have addressed issues in mortality risk prediction models (Singh et al. 2021; Zhang et al. 2020), depression recognition models using social media (Aguirre, Harrigian, and Dredze 2021; Cheong et al. 2023), and gender bias in word embeddings for the depression domain (Sogancioglu, Kaya, and Salah 2023).

In this study, we focus on the fairness of clinical NLP models in the mental health domain. These applications are based on the analysis of clinical experts' case notes, or patient transcripts.

Biases can be introduced at various stages of the pipeline of such systems, including sampling bias (e.g., healthcare access disparities), design choices for AI models, or language variations in patient case notes (Bagheri et al. 2023). Naturally, we want our systems to be fair with regards to patient demographics, such as biological sex[1] and age. However, these are clinically relevant features for many medical issues. Omitting these features in an attempt to be more fair can lead to lower accuracies for some tasks, which is particularly problematic in medicine. Understanding what fairness means for a given clinical problem is crucial for addressing it appropriately.

In the mental health domain, various modalities have been explored beyond fairness analysis of NLP models (Aguirre, Harrigian, and Dredze 2021). These include motor activity data (Cheong et al. 2023), electronic health record (EHR) data (Mosteiro et al. 2022), audio-only (Bailey and Plumbley 2021), and audio-video models (Wei et al. 2023). While some studies report only performance score differences between gender groups (Yoon et al. 2022; Rejaibi et al. 2022), many others examine multiple fairness measures to analyze fairness and compare bias mitigation methods (Mosteiro et al. 2022; Cheong et al. 2023). These measures include *Equal Accuracy*, *Disparate Impact*, *Statistical Parity*, *Equal Opportunity*, and *Equalized Odds*. However, to the best of our knowledge, none of these approaches involve domain experts in evaluating the importance of these measures. On the other hand, the importance of collaborations between AI researchers and clinicians is highlighted by several studies to ensure fairness and usability (Banja et al. 2023; Liu et al. 2023).

This paper aims to address this gap by collaborating with clinicians to assess their needs with regard to fairness through two case studies. In the next section, we describe the fairness measures considered in this paper. This will be followed by our survey design, which was developed to gather clinicians' input.

Quantitative fairness objectives

Setting the appropriate fairness objectives for a given task is the first decision for designing a fair AI model. A recent review paper lists ten definitions of fairness and its measures (Mehrabi et al. 2021). However, in practice, it is not possible to satisfy multiple fairness measures at the same time (Chouldechova 2017). The appropriate fairness measure for a model depends on its use case and the definition of fairness for such an application (Czarnowska, Vyas, and Shah 2021).

The fairness notions can be mainly categorized into "group" and "individual" fairness, based on the definition of bias. For our study, considering the clinical use cases, we carefully selected four statistical fairness measures from the available quantitative measures in the literature: 1. Counterfactual Token Fairness, 2. Equal Opportunity, 3. Predictive Equality, and 4. Equal Accuracy measures. Before providing more detail, we note here that while performance parity measures are often used in the literature, individual fairness is less studied, and we chose counterfactual token fairness to evaluate models from this perspective.

Counterfactual Token Fairness (CTF)

Counterfactual fairness (Kusner et al. 2017) aims to satisfy individual fairness, and a decision is considered counterfactually fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belongs to a different demographic group. We measure *counterfactual token fairness* (Garg et al. 2019), which is based on the idea of creating perturbations by substituting tokens associated with identity groups.

In our use cases, this measure expects the model to behave the same given two clinical notes that differ only in gender pronouns. To quantify the fairness of trained machine learning (ML) models in terms of counterfactual token fairness, we first create counterfactual pairs for each note in the test set by substituting gender pronouns (e.g., he→she) in the original datasets. Then, we use mismatch ratio (R_m), namely the number of pairs with mismatched predictions (d) divided by the total number of counterfactual pairs (N_{CP}), as given by the following equation:

$$\mathrm{R}_m = \frac{d}{N_{CP}}. \tag{1}$$

Equal Opportunity ($EqOpp$)

EqOpp states that each group of given sensitive attributes should have equal true positive rates (Hardt, Price, and Srebro 2016). In other words, it ensures that people with a positive actual outcome have the same chances for positive predictions, regardless of their sensitive group.

Let $G \in \{g_0, g_1\}$ be the sensitive attribute and let Y and Z denote the actual label and predicted label, respectively. For

[1] We use *gender* and *sex* interchangeably here, as it is frequently done in the fairness literature, but note that gender is a complex socio-cultural construct.

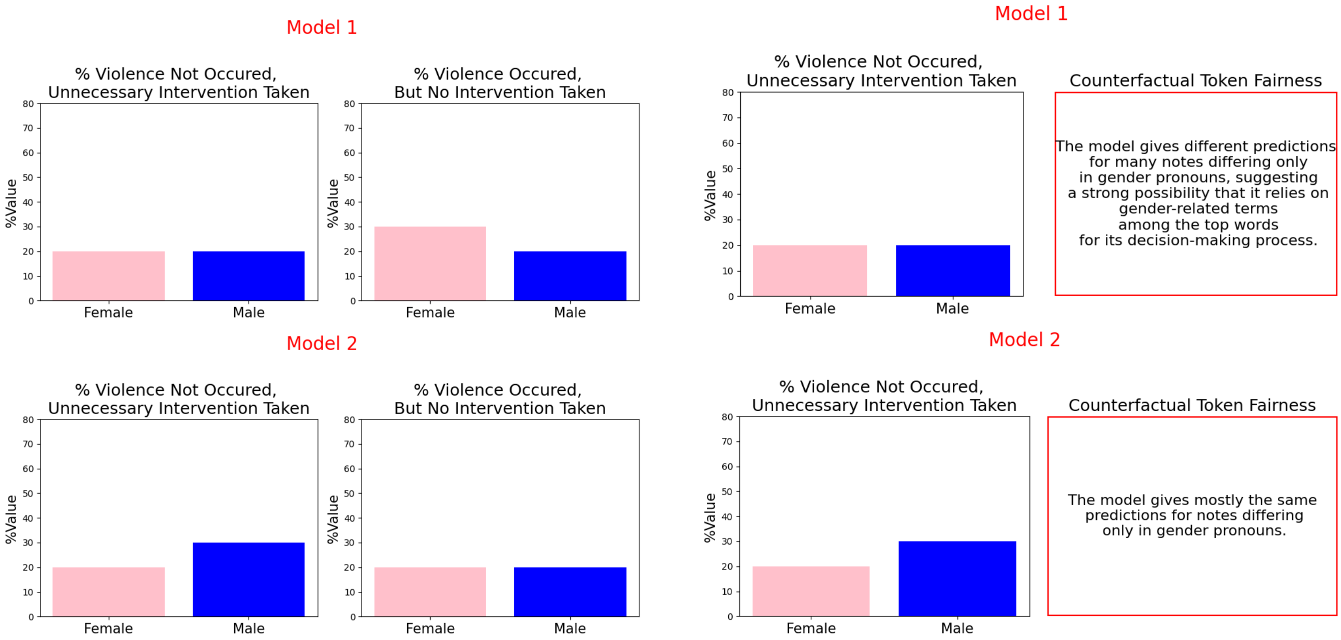

Figure 1: Example pairwise comparisons of fairness metrics from structured interview questions of Violence Risk Assessment use case: [Left] Comparison between Equal Opportunity and Predictive Equality; [Right] Comparison between Predictive Equality and Counterfactual Token Fairness.

a classifier to be considered fair according to Equal Opportunity, it must satisfy the condition $P[Z = 1|Y = 1, G = g_0] = P[Z = 1|Y = 1, G = g_1]$. Thus, the EqOpp fairness measure is defined as the ratio of the true positive rates of groups g_0 and g_1, as given below:

$$EqOpp = \frac{P[Z = 1|Y = 1, G = g_0]}{P[Z = 1|Y = 1, G = g_1]}. \tag{2}$$

Predictive Equality ($PredEq$)

PredEq requires false positive rates to be equal for given two sensitive groups g_0, g_1 (Corbett-Davies et al. 2017).

$$PredEq = \frac{P[Z = 1|Y = 0, G = g_0]}{P[Z = 1|Y = 0, G = g_1]}. \tag{3}$$

Equal Accuracy ($EAcc$)

EAcc ensures that both subgroups, g_0 and g_1 should have equal rates of accuracy. Since for both classification models, macro-averaged F1 is used as a performance measure, we formulate the $EAcc$ accordingly.

$$EAcc = \frac{F1_{g_0}}{F1_{g_1}}. \tag{4}$$

For all performance parity measures ($EqOpp$, $PredEq$, and $EAcc$), the disadvantaged group's performance is normalized by the advantaged group's performance to ensure the score is in the [0, 1] range. The higher the ratios, the fairer the model's predictions. An ideal score of 1 indicates a perfectly fair system.

Deciding on fairness objectives

Since theoretically it is impossible to satisfy all fairness measures simultaneously (Chouldechova 2017), we will aim to understand and determine the ranking of these measures by their importance to the domain experts. Involving domain experts is especially important for tasks that demand extensive domain knowledge. However, effectively communicating potential fairness measures and objectives to a non-technical audience poses some challenges (Saha et al. 2020). Inspired by the previous literature on designing surveys on fairness measures (Harrison et al. 2020), we prepared a survey and tested it for two use cases: violence risk assessment and depression phenotype recognition, respectively. We have two goals in these surveys: gaining a better understanding of the fairness perceptions of clinicians, and ranking the quantitative fairness measures for a given use case.

The survey consists of four main sections:

- *Project information and consent form*: We provide an overview of the project, detailing our aims and objectives. We emphasize that there will be no recordings, survey outputs are kept strictly confidential, and that results are only reported in an aggregated form. The survey starts only after participants have reviewed this information and provided consent.
- *Demographic information and knowledge level questions*: Participants provide (optional) personal details and

respond to queries assessing their understanding of the subject matter.

- *Explanation of bias measures and assessment of their significance*: Besides open-ended questions, bias measures are explained, and participants are asked to rate their importance within given scenarios using a 4-point scale (1: not a relevant measure; 2: maybe important, 3: should be satisfied, 4: must be satisfied).
- *Pairwise model comparisons*: Participants are presented with pairs of models and are asked to choose the one they perceive as being fairer. Example pairwise choices are illustrated in Figure 1. In the example presented on the left, the aim is to identify the significance of Equal Opportunity versus Predictive Equality. In the other example (right block), participants are told to select between a model emphasizing Predictive Equality and one prioritizing Counterfactual Token Fairness. We selected the values of 20% and 30% to represent a noticeable, yet realistic disparity. Using hard-coded values focused the user study on relative perceptions of fairness measures, avoiding the complexity of varied model outputs. Consistent use across all experiments maintained simplicity and consistency, aiding reader comprehension and comparison.

To ensure clarity of (technical) questions for non-technical participants, the survey was improved iteratively based on feedback from four AI researchers and one senior psychiatrist.

We identified three challenges for the survey setting. First, it was particularly challenging for participants with limited AI experience, and some questions required further clarification (such as the definition of counterfactual token fairness). It should be noted that the use cases we worked with were not necessarily familiar to the clinicians. The second challenge was the terminology differences between computer science and psychiatry. For instance, clinicians more frequently used the term ‘sensitivity’ rather than ‘true positive rates’ or ‘recall’, indicating that the concept of ‘equal opportunity’ is more easily understood when explained through the measure of sensitivity. Third, as fairness in mental health is a highly complex topic that requires consideration of numerous issues, it was not suitable for a shorter survey setting. Consequently, we decided to conduct longer interviews, in which participants completed the survey with assistance. This approach also enabled us to ask more open-ended questions based on their responses.

We restricted our pool of potential participants to individuals with clinical experience, specifically in psychiatry or clinical psychology, and reached out to them through departmental email lists and personal networks. The survey can be found in the Github repository[1].

Case studies

In this section, we describe our two case studies, both NLP based problems in mental health: violence risk assessment and depression phenotype recognition, respectively.

Violence risk assessment

For this task, we have used a dataset for predicting violence incidents, procured from the Psychiatry Department of the University Medical Center Utrecht (UMCU) in the Netherlands.

The study was approved by the ethics board of the hospital; the data access was limited, and personal data was not taken out of the servers. Violent incidents are reported by healthcare professionals, typically involving verbal and physical aggression from patients directed at staff or other patients. The objective is to predict violence incidents between the first and 28th day of admission, based on clinical texts written up to and including the first day of admission.

Most of the clinically relevant information was entered into the Electronic Health Record (EHR) in free text format by psychiatrists and nurses, with entries typically containing between 100 and 500 words in Dutch. These are respectively referred to as ‘doctor notes’ and ‘nurse notes’. Doctor notes mainly contain information about patient history, current treatment details (e.g., types of medication and therapy), and changes therein. Nurse notes typically contain details about the patient’s current well-being and activities. All notes were de-identified using the De-identification Method for Dutch Medical Text (DEDUCE) before any other processing took place (Menger, Scheepers, and Spruit 2018; Mosteiro et al. 2021). All notes collected between 28 days before and 1 day after the beginning of the admission period are concatenated and considered a single period note for each admission period. The outcome variable is determined based on the occurrence of a violence incident within 1 to 28 days after the start of the admission period. If such an incident occurs, the outcome is recorded as violent (positive), otherwise, it is noted as non-violent (negative).

	Negative	Positive	Total
UMCU-Violence			
Male	1822	259	2081
Female	2033	166	2199
Total	3855	425	4280
MIMIC-Depression			
Male	278	134	412
Female	184	119	303
Total	462	180	715

Table 1: The number of examples per gender for Violence and MIMIC-III datasets, respectively.

Depression phenotype recognition

Our second case study is about the recognition of depression phenotypes from clinical notes. We have used the Multiparameter Intelligence Monitoring in Intensive Care (MIMIC-III) (Johnson et al. 2016) Clinical Database, which consists of clinical note events in the English language that describe the diagnosis and treatment of more than 40.000 adult patients at the Intensive Care Unit of Beth Israel Deaconess Medical Center (located in Boston, Massachusetts, USA),

[1] https://github.com/gizemsogancioglu/gender-bias-mental-health

between 2001 and 2012. Access to the MIMIC-III dataset is possible after completing a recognized course in protecting human research participants and signing a data use agreement [2].

The clinical notes are of varying types, such as discharge summaries, nursing notes, and radiology reports. Moreover, patient demographic information and ICD-9 (International Classification of Diseases, 9th Revision) codes are stored in electronic health records. ICD-9 codes, assigned by the billing department, represent the conditions or treatments that patients receive. However, much of the information is only detailed in the clinical notes. For this reason, the subset of the dataset was annotated by two groups consisting of both physicians and clinical researchers (Moseley et al. 2020a) for 13 clinical patient binary phenotypes, including mental disorders and physical diseases, such as heart disease.

Since our focus is depression phenotype recognition, we used the examples annotated as 'depression' phenotype as positive examples. The *depression phenotype* description used for annotation includes the diagnosis of depression, prescription of anti-depressant medication, or any description of intentional drug overdose, suicide, or self-harm attempts (Moseley et al. 2020a). The annotated database has 844 examples. If a phenotype occurs in the clinical note, this was annotated as 1, otherwise 0. A single clinical note can contain multiple phenotypes. For each instance in the dataset, if at least one of the annotators labeled it as a positive example, it was subsequently used as a positive instance. From the remaining notes, we disregard the ones that are annotated for psychiatric disorders phenotype, as they might share several symptoms (comorbidity) with depression examples. Therefore, the remaining notes were used as negative examples.

Classification models and experimental setup

Predictive models

Following earlier studies (Moseley et al. 2020b; Menger, Scheepers, and Spruit 2018), we train binary classification models for both depression phenotype recognition and violence risk assessment problems. We apply the same preprocessing steps for both datasets to clean them before feature extraction; including the removal of digits, special symbols, punctuation, and one-character terms.

We use pre-trained non-contextual and biomedical domain-specific FastText (Bojanowski et al. 2017) embeddings, namely Bio-W2Vec, to extract features. The UMCU dataset, used for training the violence risk assessment models, is in Dutch, while the MIMIC-III dataset, used for training the depression phenotype classifier, is in English. Therefore, we utilize language-specific pre-trained word embeddings that are similar in methodology and domain. Zhang et al. (2019) provide 200-dimensional pre-trained English biomedical word embeddings, trained on Pubmed. Unfortunately, there were no publicly available Dutch non-contextual clinical word embeddings, so we followed Menger, Scheepers, and Spruit (2018) and trained the FastText model on the entire UMCU Violence dataset, yielding 100-dimensional word vectors.

To extract features for a clinical note, we first obtain each word vector from the pre-trained embeddings, and then use the average vector representation as the feature representation for the text. The vectors are z-normalized using parameters estimated from the respective training set, and Support Vector Machine (SVM) models are trained for each binary classification problem.

Evaluation of performance. In clinical models, macro-averaged F1 score and Area Under Curve (AUC) are among the most common measures due to the imbalanced nature of the problems (Gehrmann et al. 2018). In this study, we use a macro-averaged F1 score as the main performance measure for tuning all models. We chose it not only to provide consistency across the models, but also to focus on binary predictions rather than on continuous scores. Fairness measures for binary predictions are also easier to interpret, and consequently, binary predictions simplify the evaluation of fairness measures.

Training procedure. We employ a consistent experimental framework to train binary models for both *violence risk assessment* and *depression phenotype recognition*. To ensure the generalizability of our findings, we implement five iterations (random seed = $\{0,1,2,3,4\}$) of 10-fold stratified cross-validation, resulting in the training, validating and testing of 50 models per task. Each train, validation and test set has an equal number of examples for each gender-class group and they reflect distributions similar to those observed in the original dataset.

Hyperparameter tuning is performed on the validation set for each model, tuning the C parameter (ranging from 0.01 to 10 with exponential increments) and the selected kernel ($\in \{rbf, linear\}$). These models, trained on the original dataset, are referred to as *original* models in our experiments.

Bias mitigation methods

Bias mitigation methods often come with trade-offs, and as such, are carefully selected and evaluated to find a balance between predictive power and fairness. We use and evaluate five different approaches that are used at different stages of the machine learning (ML) pipeline.

Bias mitigation algorithms can be categorized into three main groups (Mehrabi et al. 2021): 1. *Pre-processing* modifies the training dataset or features to prevent learning biased relationships. 2. *In-processing* adjusts learning algorithms during training to reduce discrimination. 3. *Post-processing* transforms predictions into fairer outcomes after training.

Clinical notes for both datasets are written by the nurse or the practitioner. Thus, gender pronouns are explicitly used in the datasets to refer to the patient. Although these are not explicitly given as a feature of the machine learning model, they are extensively available in textual notes. Moreover, as shown earlier in the Case Studies section, class distributions vary across gender groups for both tasks, which can impact model performance for less represented groups. Given this

[2] https://physionet.org/content/mimiciii/1.4/

Male terms	Female terms	Neutralized terms
he	she	patient
man	woman	patient
his	her	its
him	her	its
male	female	<SPACE CHAR>
Sex: M	Sex: F	Sex: <SPACE CHAR>

Table 2: The English list of female, male, and neutralized terms that are used for gender swapping and gender neutralization. Swapping: female terms ⟷ male terms, Neutralization: (female or male) terms → neutralized terms.

disparity, pre-processing approaches are promising in enhancing fairness. We evaluate three common pre-processing methods, along with one post-processing method. Each of these approaches is explained in detail below.

Pre-processing: Data Augmentation. The training dataset is augmented by substituting gender pronouns with those of another gender group. Subsequently, a binary classifier is trained using both the original and the modified datasets combined. This approach is assessed as a bias mitigation method, specifically targeting counterfactual token fairness. It aims to instruct the model not to exhibit any explicit differences between counterfactual pairs (e.g., swapping 'she' with 'he', or 'woman' with 'man') as can be seen in Table 2.

Pre-processing: Data Neutralization. In addition to the training data, the entire dataset is rendered gender-neutral by either eliminating or substituting all gender-specific terms used to refer the patient with gender-neutral alternatives (see Table 2). This straightforward approach has also been employed in prior studies (Garg et al. 2019).

The data augmentation and data neutralization approaches are both central to the counterfactual token fairness assumption, which expects two identical notes with different gender terms to have the same outcome/prediction. These approaches were previously shown to be very successful for counterfactual token fairness (Garg et al. 2019) in different problems such as hate speech detection (Zhao et al. 2018) and patient phenotype recognition (Sogancioglu, Kaya, and Salah 2023). The list of Dutch terms that are neutralized/swapped can be found in the Github repository[1].

Pre-processing: Reweighing Kamiran and Calders (2012) proposed a pre-processing method that assigns weights to each training example based on its sensitive attribute-class class combination. This method is particularly useful when dealing with data imbalance, as it aims to enhance fairness by assigning higher weights to examples from minority groups, ensuring these underrepresented groups and outcomes receive more emphasis.

Post-processing: Reject Option Classification (ROC) One widely used post-processing-based bias mitigation technique is the Reject Option Classification (ROC), introduced by Kamiran, Karim, and Zhang (2012). This method targets the low-confidence region of a classifier's predictions, utilizing the validation set to identify an optimal threshold that balances fairness. In our approach, we apply the ROC method to adjust decision thresholds within this low-confidence region, aiming to optimize both true and false positive rates.

Gender-specific models We train separate models for each demographic group, a common practice in clinical tasks that may show differences due to gender. During prediction, examples from female individuals are predicted using the model trained specifically on the female subset, while examples from male individuals are classified by the model trained exclusively on the male subset.

Findings and experimental results

In this section, we first present the results of the interviews and discuss the themes that emerged from them. Then, we elaborate on the experimental results of the NLP models and bias mitigation methods in terms of fairness and performance.

The results of semi-structured interviews: clinicians' perceptions on AI fairness

A total of 8 participants (2 clinical psychologists and 6 psychiatrists) were interviewed between March-April 2024. The duration of the interviews was 20-30 minutes per use-case. Three participants were interviewed for two use-cases, while the remaining participants were randomly shown one of the two use-cases. This resulted in five interviews for the depression phenotype classification use-case and six for the violence risk assessment use-case, respectively. Overview of participants and demographics are given in Table 3. The participants have varying levels of knowledge and experience with AI models and EHR data.

The thematic analysis was conducted by the first author. Initial codes were developed manually, without the use of qualitative analysis software. Codes were organized into potential themes, which were then reviewed and refined through an iterative process. The refinement process focused on ensuring clarity and distinctiveness of each theme, resulting in seven final themes which we explain in detail below.

- *Use-Case/Goal/Intervention dependent fairness:* All participants recognize the importance of fairness in predictive models. Fairness is context-dependent, with the desired balance of performance measures differing between scenarios such as violence and depression. This reflects a broader principle that fairness must be adaptable to the specifics of each clinical application, recognizing the unique challenges and implications of different conditions. It has been noted that in medicine, the main principle is "first, do no harm", leading to a greater emphasis on avoiding false positives. However, perceptions of fairness may change depending on the intervention and potential outcomes.
- *No sacrifice from accuracy:* The importance of overall prediction performance (e.g., F1) is highlighted by all

Background	Experience (years)	Age Group	Gender	Use-case	Country	AI Level	EHR Level
Psychiatry	5-10	36-45	F	D	NL	4	4
Psychiatry	1-5	25-35	F	D	NL	1	4
Psychiatry	1-5	25-35	F	V	NL	3	3
Psychiatry	1-5	25-35	F	V	NL	4	3
Psychiatry	> 10	56-65	M	V	NL	3	3
Psychiatry	> 10	46-55	F	V + D	NL	4	4
Clinical Psychology	5-10	25-35	F	V + D	NL	3	3
Clinical Psychology	> 10	46-55	F	V + D	TR	2	2

Table 3: Overview of Participants. V: Violence risk assessment use-case, D: Depression phenotype recognition use-case, F: Female, M: Male. AI and EHR knowledge levels are on a 5-point scale (5: extremely, 4: very, 3: moderately, 2: slightly knowledgeable, 1: not knowledgeable at all).

	$EqOpp$	$EAcc$	$PredEq$	CTF
Violence				
must be satisfied	2	0	1	0
should be satisfied	2	3	5	0
maybe important	2	3	0	3
not relevant	0	0	0	3
important	4/6	3/6	6/6	0/6
Depression				
must be satisfied	2	1	1	2
should be satisfied	3	4	4	3
maybe important	0	0	0	0
not relevant	0	0	0	0
important	5/5	5/5	5/5	5/5

Table 4: Importance distributions of fairness measures.

participants (n=8), and sacrificing it to equalize performance measures across gender groups is often not acceptable for models implemented for the mental health domain. The primary goal is to provide the best possible care for each patient, group, and subgroup, taking into account their unique characteristics. It was suggested that when performance disparities exist between groups, efforts should focus on understanding and enhancing the outcomes for disadvantaged groups rather than diminishing the accuracy for advantaged groups.

- *No fairness through unawareness:* Fairness through unawareness, as measured in fairness literature, is defined as follows: “An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process” (Mehrabi et al. 2021). On the other hand, sex and age are clinically relevant features for most clinical problems, including depression diagnosis and violence risk assessment. Thus, excluding it when relevant is perceived as unfair by clinicians as it can worsen the performance and fairness of the models. Clinicians emphasize the importance of being aware of clinical differences (e.g., symptoms) among certain groups and considering these differences in the decision-making process. While they indicate that they would not use gender for identifying depression phenotype if they were decision-makers, the majority (n=4 over 5 participants) do not view its use in NLP models as unfair.
- *Variable importance of performance measures by gender:* The importance of different performance measures (e.g., false positive and negative rates) is seen to vary by gender, especially in the context of violence. For males, minimizing false negatives is prioritized due to the typically more physical nature of their aggression. It was emphasized that increasing the false negative rate for males to improve fairness metrics is unacceptable, whereas doing so for females might be more tolerable given the potential harm to staff members as an outcome.
- *Awareness of gender biases in clinical practices:* There is an acknowledgment of historical and ongoing gender biases in clinical practices, such as the use of male-centric criteria for diagnoses (Lai et al. 2015). In this case, biases may also arise from the way the AI model or problem is defined, such as targeting predominantly male-centric symptoms for a diagnosis of mental disorders, unknowingly reflecting biases present in clinical knowledge. This highlights the need for fairness in not just algorithmic models, but also in the broader clinical diagnostic criteria and treatment practices.
- *Communication and knowledge among clinicians:* Ensuring that clinicians are aware of any performance differences in models by gender is vital for informed decision-making. Such transparency enables better utilization of models in clinical practice, ensuring that interventions are as effective and fair as possible. The importance of transparency and explainability was mentioned by a few participants (n=4 over 8 participants). It was highlighted that capabilities and accuracy of AI models should be communicated to clinicians. If it is known that the model works better for some groups, clinicians can be educated and informed to use these models cautiously to prevent disparities.
- *Fairness beyond gender:* There is a consensus that while gender should be considered to ensure fairness, it is far from being sufficient. Fairness for other groups (e.g., age groups) and subgroups (e.g., young females) should be studied. Additionally, recognizing the distinctiveness of

each patient requires integrating their individual histories into decision models. To ensure individual-level fairness, the focus should be customizing interventions to the individual, going beyond simple representations of patients with a few categorical variables.

Throughout the interviews, we identified the need to define goals and interventions for each model to define fairness in the given contexts. This motivated us to formulate some assumptions regarding the use-cases. For instance, in violence risk assessment, we told participants that we assume that the objective is to identify *outward* violent incidents, with potential interventions including *restricting patients' social activities or prescribing medications*. Conversely, in depression phenotype recognition, clinicians were informed that the model would serve as an *alerting tool*, rather than directly influencing diagnosis.

Beyond qualitative analysis, we also report statistics from multiple-choice questions. For each fairness measure, we asked participants to rate its importance on a 4-point scale before deploying models for clinical use. The results are shown in Table 4. We show the number of participants who considered the measure important (indicating it should or must be satisfied) out of the total participants, on the 'important' row.

For the violence use case, there is full agreement on the importance of the $PredEq$ measure, and partial agreement on the satisfaction of $EqOpp$ and $EAcc$. In contrast, counterfactual token fairness is not strongly supported by any participants. On the other hand, for the depression use case, there is less distinction among the fairness measures. All participants consider all measures important and think they should be satisfied before deploying models.

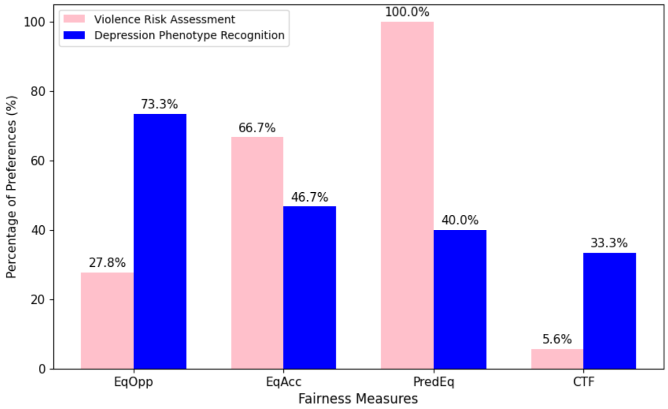

Figure 2: Comparison of relative preferences for fairness measures across use-cases.

The results of pairwise comparison questions are shown in Figure 2. The bar plots illustrate the number of preferences over the total number of pairwise questions for each fairness measure. For the violence use-case, $PredEq$ is the most preferred measure, with a 100% preference rate, followed by $EqAcc$, $EqOpp$, and CTF.

For the depression use-case, while all participants consider all measures important, pairwise comparisons help to identify the most critical measures, which is crucial in case of trade-offs. As expected, the preference percentages are more evenly distributed. However, $EqOpp$ is the most preferred measure, followed by $EqAcc$, $PredEq$, and CTF.

Comparison of bias mitigation methods

Table 5 presents a comparison of various bias mitigation methods, evaluated through fairness and performance measures. Each measure shows average scores from 5×10 fold cross-validation experiments. To assess the difference between male and female groups, independent t-tests are applied to female and male pairs of True Positive Rates, False Positive Rates, and F1 scores for the $EqOpp$, $PredEq$, and $EAcc$ measures, respectively. When a significant difference ($p < 0.05$) is found—in other words, the bias direction is significant—the group with the higher score is indicated (F: female, M: male). To ensure that the values are in the [0, 1] range, the higher scores were used as the denominator for all measures.

We also conduct two-tailed paired t-tests to measure the significance between the original model and each bias mitigation method in terms of fairness (R_m, $EqOpp$, $PredEq$, $EAcc$) and overall performance (F1). Significant differences ($p < 0.05$) are shown in bold: with a + indicated next to values that are better and a - next to values that are worse.

The violence risk assessment model, trained on the UMCU Violence Dataset, shows significant bias toward male groups, resulting in higher True Positive and False Positive Rates. This result is not surprising due to the highly gender-class imbalanced dataset. However, the $PredEq$ measure, chosen as a key fairness metric, is 0.45, indicating the model generates over twice as many false positives for males than females. This highlights the potential for significant harm to certain groups when fairness is disregarded.

All approaches show significant improvements over the original method in terms of the $PredEq$ measure. While no method enhances the $EqAcc$ measure, Reweighing and ROC also significantly improve $EqOpp$. Despite its success across three fairness dimensions, the ROC method is unsuitable for this task, as clinicians explicitly oppose trading performance for fairness.

Using gender-specific models, as expected, significantly worsens the mismatch ratio, due to using two distinct models for decisions, resulting in a natural mismatch. However, most participants view the mismatch ratio as irrelevant to the violence risk assessment task. Consequently, changes in this dimension can be disregarded for this problem. Considering these factors, the Reweighing method (followed by gender-specific models) appears to be the best choice, given the trained model, task, and the importance of fairness measures.

In contrast, while most participants view $EqOpp$ as the most important fairness measure for depression phenotype classification, there is no significant distinction between the importance of the remaining fairness measures for this task. Gender-specific models and ROC methods are disregarded, as they both significantly lower the overall performance. None of the remaining methods significantly im-

	UMCU-Violence					MIMIC-Depression				
	R_m↓	$EqOpp$↑	$PredEq$↑	$EAcc$↑	$F1$↑	R_m↓	$EqOpp$↑	$PredEq$↑	$EAcc$↑	$F1$↑
Original	0.08	0.68_M	0.45_M	0.92	0.61	0.04	0.79	0.64_F	0.89	0.69
Gender-specific	$\mathbf{0.18}^-$	0.76	$\mathbf{0.64}_M{}^+$	0.93	0.62	$\mathbf{0.35}^-$	0.80	0.62_F	0.87_M	$\mathbf{0.67}^-$
Data Aug.	$\mathbf{0.03}^+$	0.66_M	$\mathbf{0.52}_M{}^+$	0.93	0.62	$\mathbf{0.0}^+$	0.79	0.66	0.87	0.70
Data Neutr.	$\mathbf{0.0}^+$	0.71_M	$\mathbf{0.52}_M{}^+$	0.93	0.61	$\mathbf{0.0}^+$	0.79	0.69	0.88	0.68
Reweighing	$\mathbf{0.06}^+$	$\mathbf{0.76}_M{}^+$	$\mathbf{0.61}_M{}^+$	0.93	0.61	$\mathbf{0.02}^+$	0.80	0.68	0.88	0.69
ROC.	$\mathbf{0.07}^+$	$\mathbf{0.81}_M{}^+$	$\mathbf{0.65}_M{}^+$	0.94	$\mathbf{0.59}^-$	$\mathbf{0.03}^+$	0.81	$\mathbf{0.73}^+$	0.89	$\mathbf{0.67}^-$

Table 5: Results from stratified 5×10-fold cross-validation, comparing fairness and performance measures across original NLP models and various bias mitigation methods. Scores significantly better (indicated with +) or worse (indicated with -) than those of the original model, based on a two-tailed paired t-test ($p < 0.05$), are shown in **bold**. The letter 'F' (or 'M') indicates that the measure for the female (or male) group is significantly higher than that of the other group, based on an independent t-test ($p < 0.05$). R_m represents mismatch ratio, $EqOpp$ equal opportunity, $PredEq$ predictive equality, and $EAcc$ equal accuracy. F1 denotes the macro-averaged F1 score for the overall model.

prove $EqOpp$, $PredEq$, or $EqAcc$. However, Data Augmentation and Data Neutralization approaches markedly improve the mismatch ratio (R_m), making them suitable for this problem.

In this study, we compare approaches based on their relative improvement over the original method, incorporating input from clinicians on the relative importance of these measures. However, acceptable thresholds for these measures should be discussed with domain and ethics experts.

Discussion

In conclusion, our research contributes to understanding fairness within automated decision-making systems in the clinical setting, with a focus on the mental health domain. Through an insightful exploration of clinicians' perceptions of fairness in two use-cases—violence risk assessment and depression phenotype recognition—we gain an understanding of the importance of fairness objectives and capabilities of bias mitigation methods in clinical decision-making.

Fairness objectives in mental health. Through qualitative analysis of interviews, we identified seven themes that reveal the complexity of fairness in mental health. To ensure fairness in mental health, one should consider many aspects and associated parameters, while selecting appropriate fairness measures. The importance of fairness measures in the mental health domain varies significantly depending on the context, use-case, and the goal of the model (e.g., alerting, diagnosis), as well as possible interventions. In assessing performance parity, participants prioritized fairness measures considering potential harm to the patients. For instance; for assessments of outward violence risk (with the aim of prevention through medication or patient restriction), the False Positive Rate is viewed as more harmful by participants thus *Predictive Equality* is chosen as the most critical fairness measure. However, if the problem was inward violence, the consequences of false negatives (such as suicide attempts) may be more damaging than those of false positives (such as unnecessary medication).

We compared various bias mitigation strategies for selected fairness objectives. The analysis revealed that although some fairness measures show improvement, the most critical fairness measure might remain unmet (e.g. *Equal Opportunity* for depression phenotype recognition). This lack of understanding could result in misguided improvements in model behavior. The results highlight the importance of the selection and ranking of fairness measures as this is not merely a technical decision. This underscores the need for an approach to prioritizing fairness measures that are most aligned with the ethical considerations and practical implications in the clinical context.

Label bias. The commonly used fairness measures rely on ground-truth labels to assess the fairness of models. In the case of violence risk assessment use cases, the task is to predict future incidents. Conversely, depression phenotype recognition relies on annotations by clinicians. One should note that intrinsic biases may be reflected in the annotation process (Şahin et al. 2024). To fairly assess these aspects of the model and improve the quality of the labels, it is important to involve all groups in the annotation process. In this study, we assume that labels are ground-truth annotations.

How to choose an appropriate bias mitigation method for such tasks? Ensuring the fairness of models across various demographic groups is crucial, yet achieving high accuracy remains paramount for clinical issues. With insights from clinicians, we've highlighted the importance of fairness measures and recommended certain bias mitigation methods. We observed that the *reweighing* method is suited for the violence risk assessment model, while *data augmentation/neutralization* approaches work best for depression phenotype recognition. These methods are suggested for their ability to enhance critical fairness dimensions without deteriorating performance. However, this knowledge alone is insufficient, as achieving a perfectly fair system (e.g., Equal Opportunity = 1) is often not possible. Therefore, it is important to establish and communicate an acceptable threshold for fairness measures with the domain and ethics experts for the specific problem at hand.

Generalizability of data augmentation and neutralization methods. To measure counterfactual token fairness, we use a pre-defined list of gender pronouns to augment the test dataset. The data augmentation and data neutralization methods also utilize this dictionary to augment or neutralize the training set. However, this introduces a generalizability problem common to all dictionary-based approaches: these methods might not generalize well when test instances contain gendered terms not previously seen. Earlier studies have shown that data neutralization can even amplify existing downstream biases, likely due to gender biases in pre-trained word embeddings (Sogancioglu, Kaya, and Salah 2023). If counterfactual token fairness is a critical measure for a given problem, this limitation can be addressed by implementing more sophisticated approaches, such as counterfactual logit pairing (Garg et al. 2019).

Limitations. To understand clinicians' fairness perceptions for given use-cases, we relied on expert interviews, and since experts' time was valuable, we had a small set of experts to draw upon. From the eight experts we consulted, seven resided in a single country. This might have introduced a cultural bias, as culture significantly shapes people's values and perceptions, potentially influencing the interview results. Furthermore, since none of these use-cases have yet been implemented in clinics, several assumptions were made regarding interventions, model goals, and other aspects, which, as discussed in the interviews, greatly influence fairness decisions.

Another limitation was the use of a binary definition of gender, as gender pronouns are likely chosen based on the patient's biological sex in the dataset. While this study makes an important step towards understanding and equalizing outcomes for binary gender groups, it is crucial to consider individual fairness using patient histories and other groups and subgroups, such as ethnicities, age groups, and non-binary identities. Addressing these broader dimensions can lead to a more comprehensive understanding of fairness in these contexts. The importance of this issue was also highlighted by the participants during interviews.

The counterfactual token fairness measure relies on gender pronouns, which is straightforward to implement for many Western languages (as demonstrated in English and Dutch in this study). However, it requires alternative approaches for languages with different gendered structures (e.g., Arabic has gendered adjectives and verbs).

Finally, while this study focused on clinicians' perceptions, it is important to involve not only domain experts, but also ethics experts and patients who will be affected by these automatic tools to ensure fairness (Banja et al. 2023).

Acknowledgments

We would like to thank all participants for their time and valuable contributions. Additionally, we extend our gratitude to Joppe Kooistra for his efforts in preparing the list of Dutch gender pronouns and gender-neutral alternatives.

References

Aguirre, C.; Harrigian, K.; and Dredze, M. 2021. Gender and Racial Fairness in Depression Research using Social Media. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2932–2949. Online: Association for Computational Linguistics.

Bagheri, A.; Giachanou, A.; Mosteiro, P.; and Verberne, S. 2023. Natural Language Processing and Text Mining (Turning Unstructured Data into Structured). In *Clinical Applications of Artificial Intelligence in Real-World Data*, 69–93. Springer.

Bailey, A.; and Plumbley, M. D. 2021. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 596–600. IEEE.

Banja, J.; Gichoya, J. W.; Martinez-Martin, N.; Waller, L. A.; and Clifford, G. D. 2023. Fairness as an afterthought: An American perspective on fairness in model developer-clinician user collaborations. *PLOS Digital Health*, 2(11): e0000386.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NEURIPS*, 29: 4349–4357.

Cheong, J.; Kuzucu, S.; Kalkan, S.; and Gunes, H. 2023. Towards gender fairness for mental health prediction. *IJCAI 2023*.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Czarnowska, P.; Vyas, Y.; and Shah, K. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9: 1249–1267.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 219–226.

Gehrmann, S.; Dernoncourt, F.; Li, Y.; Carlson, E. T.; Wu, J. T.; Welt, J.; Foote Jr, J.; Moseley, E. T.; Grant, D. W.; Tyler, P. D.; et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2): e0192360.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; and Ur, B. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 392–402.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-W. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Kamiran, F.; Karim, A.; and Zhang, X. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, 924–929. IEEE.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *NEURIPS*, 30.

Lai, M.-C.; Lombardo, M. V.; Auyeung, B.; Chakrabarti, B.; and Baron-Cohen, S. 2015. Sex/gender differences and autism: setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54(1): 11–24.

Liu, M.; Ning, Y.; Teixayavong, S.; Mertens, M.; Xu, J.; Ting, D. S. W.; Cheng, L. T.-E.; Ong, J. C. L.; Teo, Z. L.; Tan, T. F.; et al. 2023. A translational perspective towards clinical AI fairness. *NPJ Digital Medicine*, 6(1): 172.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14867–14875.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Menger, V.; Scheepers, F.; and Spruit, M. 2018. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6): 981.

Moseley, E.; Celi, L. A.; Wu, J.; and Dernoncourt, F. 2020a. Phenotype annotations for patient notes in the MIMIC-III database. *PhysioNet*.

Moseley, E. T.; Wu, J. T.; Welt, J.; Foote, J.; Tyler, P. D.; Grant, D. W.; Carlson, E. T.; Gehrmann, S.; Dernoncourt, F.; and Celi, L. A. 2020b. A Corpus for Detecting High-Context Medical Conditions in Intensive Care Patient Notes Focusing on Frequently Readmitted Patients. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1362–1367. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Mosteiro, P.; Kuiper, J.; Masthoff, J.; Scheepers, F.; and Spruit, M. 2022. Bias Discovery in Machine Learning Models for Mental Health. *Information*, 13(5): 237.

Mosteiro, P.; Rijcken, E.; Zervanou, K.; Kaymak, U.; Scheepers, F.; and Spruit, M. 2021. Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes. *Journal of Artificial Intelligence for Medical Sciences*, 2: 44–54.

Rejaibi, E.; Komaty, A.; Meriaudeau, F.; Agrebi, S.; and Othmani, A. 2022. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71: 103107.

Saha, D.; Schumann, C.; Mcelfresh, D.; Dickerson, J.; Mazurek, M.; and Tschantz, M. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*, 8377–8387. PMLR.

Şahin, D.; Kambeitz-Ilankovic, L.; Wood, S.; Dwyer, D.; Upthegrove, R.; Salokangas, R.; Borgwardt, S.; Brambilla, P.; Meisenzahl, E.; Ruhrmann, S.; et al. 2024. Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis. *The British Journal of Psychiatry*, 224(2): 55–65.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3–13.

Sogancioglu, G.; Kaya, H.; and Salah, A. A. 2023. The effects of gender bias in word embeddings on patient phenotyping in the mental health domain. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. IEEE.

Wei, P.-C.; Peng, K.; Roitberg, A.; Yang, K.; Zhang, J.; and Stiefelhagen, R. 2023. Multi-modal Depression Estimation Based on Sub-attentional Fusion. In Karlinsky, L.; Michaeli, T.; and Nishino, K., eds., *Computer Vision – ECCV 2022 Workshops*, 623–639. Cham: Springer Nature Switzerland. ISBN 978-3-031-25075-0.

World Health Organization. 2024. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Technical Report ISBN: 978-92-4-008475-9, Chief Scientist and Science Division (SCI), Health Ethics & Governance (HEG).

Yoon, J.; Kang, C.; Kim, S.; and Han, J. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12226–12234.

Zhang, H.; Lu, A. X.; Abdalla, M.; McDermott, M.; and Ghassemi, M. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, 110–120.

Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; and Lu, Z. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, 6(1): 52.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

2 (Short Papers), 15–20. New Orleans, Louisiana: Association for Computational Linguistics.