Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes

Chiman Salavati[1], Shannon Song[2], Willmar Sosa Diaz[1], Scott A. Hale[3,4], Roberto E. Montenegro[5], Fabricio Murai[2*], Shiri Dori-Hacohen[1*]

[1]University of Connecticut, Storrs, Connecticut, USA
[2]Worcester Polytechnic Institute, Worcester, Massachusetts, USA
[3]Meedan, San Francisco, California, USA
[4]University of Oxford, Oxford, UK
[5]University of Washington, Seattle, Washington, USA
{chiman.salavati, willmar.sosa_diaz, shiridh}@uconn.edu, {smsong, fmurai}@wpi.edu,
scott@meedan.com, roberto.montenegro@seattlechildrens.org

Abstract

Biased information (recently termed *bisinformation*) continues to be taught in medical curricula, often long after having been debunked. In this paper, we introduce BRICC, a first-in-class initiative that seeks to mitigate medical bisinformation using machine learning to systematically identify and flag text with potential biases, for subsequent review in an expert-in-the-loop fashion, thus greatly accelerating an otherwise labor-intensive process. We have developed a gold-standard BRICC dataset throughout several years containing over 12K pages of instructional materials. Medical experts meticulously annotated these documents for bias according to comprehensive coding guidelines, emphasizing gender, sex, age, geography, ethnicity, and race. Using this labeled dataset, we trained, validated, and tested medical bias classifiers. We test three classifier approaches: a binary type-specific classifier, a general bias classifier; an ensemble combining bias type-specific classifiers independently-trained; and a multi-task learning (MTL) model tasked with predicting both general and type-specific biases. While MTL led to some improvement on race bias detection in terms of F1-score, it did not outperform binary classifiers trained specifically on each task. On general bias detection, the binary classifier achieves up to 0.923 of AUC, a 27.8% improvement over the baseline. This work lays the foundations for debiasing medical curricula by exploring a novel dataset and evaluating different training model strategies. Hence, it offers new pathways for more nuanced and effective mitigation of bisinformation.

1 Introduction

The field of medicine is marred by a long, painful, and deleterious history of overt and covert forms of social injustice, bias, and racism, as illustrated by the American Medical Association's recent pledge to take action to confront systemic racism (Saini 2019; Madara 2020). Studies continue to demonstrate that physicians possess implicit biases in a number of different areas such as race/ethnicity, gender, sex, age, weight, substance use and mental illness (FitzGerald and Hurst 2017). Bias in medicine leads to harmful public health costs, which are borne disproportionately by women, the economically disadvantaged, and other minoritized (Wingrove-Haugland and McLeod 2021) groups. Harmful biases are widespread among clinicians, such as assuming psychogenic causes for physical symptoms for women, minimizing pain in people of color, or other well-documented and entrenched biases in medicine (Norman 2018). This medical biased information (recently termed *bisinformation*, Dori-Hacohen et al. 2021) among clinicians persists with pernicious effects on health inequities despite refuting evidence.

Unfortunately, a main vector of transmission for bisinformation originates from materials that continue to be taught in medical schools, even long after being debunked by research (Redacted for anonymity, Under Review). Despite numerous calls to action to deracialize and debias medical curricula and assessment content, most medical institutions continue to teach biased medicine in the preclinical years (Tsai et al. 2016; Rodney 2016; Halman, Baker, and Ng 2017). Many educators, for example, continue to inappropriately use race as a proxy for genetics or ancestry, or even as a "risk factor" for numerous health outcomes often erroneously associated with race (e.g., Salt Gene Hypothesis) while ignoring social or structural determinants of health (SSDoH), such as systemic racism or income inequities (Ali-Khan et al. 2011; Hunt, Truesdell, and Kreiner 2013; Acquaviva and Mintz 2010; Karani et al. 2017; Metzl and Roberts 2014). Likewise, the inappropriate use of gender and sex terms perpetuates the idea that sex and gender are binary or stagnant (vs. fluid), which can potentially alienate gender-nonconforming students and patients alike.

By equating social identifiers to biology without social or structural context, medical educators are unknowingly perpetuating a curriculum that medicalizes social identities like race or gender; reifies the false conceptualization that race and gender are a biological reality rather than social constructs; and perpetuates biased knowledge and inappropriate language use. Bias reduction in curricular and assessment content is key for educating future physicians in evidence-based medicine (Le et al. 2020; Ripp and Braun 2017).

Despite the urgent need for debiasing curricular content,

[*]Equally contributing senior corresponding authors.

Bias Type	Quote (biased or potentially biased)	Annotator Comment
Gender	*Often, significant changes in a child's growth reflect significant events in the family unit such as a mother going to work, parents separating, moving to a new home or a significant family illness*	*This statement reinforces traditional family structures which stigmatizes mothers going to work, or families without a mother or two mothers. Consider omitting gendered clause.*
Sex	*most common in adolescent females with BMI $>$ 30, often treated with acetazolamide and repeated therapeutic lumbar punctures, and a weight loss program.*	*Consider addressing why weight and sex are relevant factors here - does being female predispose a [patient] to idiopathic intracranial hypertension, and if so, providing a citation prevents interpretation as bias.*
Race	*Although the incidence is lower in patients of color, the morbidity/mortality can often be higher*	*No source for claim. Consider explaining more beyond "patients of color" as this may come across as grouping every minority group into one and making a generalized statement*
Ethnicity	*While the components of genetic versus environmental risk have not been fully established, note the increased incidence of colorectal cancer in the Alaskan native population.*	*Include citation, stratify why Alaskan Native Population is disproportionately affected*
Age	*New onset solid food dysphagia in anyone over 40 especially with a long history of heartburn should be considered to be esophageal cancer until proven otherwise.*	*[I]s this a medical fact regarding this age group?*
Geography	*Had any other members of his travel group suffered the same symptoms, either in Brazil or after returning?*	*No significance of Brazil indicated in the case study*

Table 1. Sample biased quotes from medical educational content for various types of biases, along with the annotator comment on how to make them less biased or back up identity-specific factors with citations. Note that many quotes were labeled for two or more types of bias.

there are several reasons why institutions continue to struggle with this issue (Krishnan et al. 2019): (i) faculty educators may have a significant knowledge gap; (ii) faculty educators of all backgrounds may resist confronting their own implicit biases and privilege (Frey 2020); (iii) examining educational content for bias and language misuse often entails a manual review of a cross-sectional sample (Tsai et al. 2016; Martin et al. 2016), leading to high variation in assessing for bias and likely results in continued bias, since not all materials are or can be examined.

Addressing and dismantling these entrenched prejudices is paramount for advancing equitable healthcare. Thus, the objective of this research is twofold. Firstly, we undertake an empirical investigation, drawing on the expertise of trained medical professionals to systematically gather, mark, and annotate instances of bias within medical texts. This process resulted in a robust and reliable dataset that reflects the multi-faceted nature of bias within medical education. Secondly, we harness this new dataset to train artificial intelligence (AI) models for discerning bias in medical text. This approach aligns with the recently introduced *Fairness via AI* framework (Dori-Hacohen et al. 2021), in which AI is used carefully and deliberately to support equitable outcomes in society. Our contributions are as follows:

- We curate a comprehensive labeled dataset from medical curricula encompassing two years' worth of instructional content (e.g. lecture notes, PowerPoint slides, articles) comprising more than 4,000 quotes annotated with codes covering a wide spectrum of biases. Quotes include medical excerpts that were labeled by experts as *biased* (Table 1) or *non-biased* (Table 2) with respect to certain social categories such as race, gender, etc.
- We map those codes to labels of interest through a multi-stage preprocessing procedure guided by our experts. We then identify the most prevalent types of biases—those related to gender, sex, race, ethnicity, age, and geographic location—to be considered in our study.
- To augment the set of 'non-bias' examples, we use lexicons of social identifiers associated with each bias type to extract additional 4,391 quotes from the medical curricular content. Filtering with those identifiers ensures that no trivial samples are included.
- We implement and evaluate four approaches to detect bias and if present, the type of bias: (i) a type-specific classifier, (ii) a general bias classifier, (iii) an ensemble of bias type-specific classifiers, and (iv) a multi-task classifier that predicts both general and type-specific biases.
- We provide a thorough comparison of the models with respect to accuracy, precision, recall, F1-score, F2-score and area under ROC curve (AUC). Additionally, we contrast the performance of a Transformer model (DistilBERT) with machine learning models trained on static textual embeddings (FastText).
- We develop a first-in-class bias detection classifier by fine-tuning a pre-trained DistilBERT model on our curated dataset, achieving 0.923 of AUC at detecting (general) bias within medical curricula.

By intersecting rigorous data curation with advanced computational techniques, this work illuminates sources of

Bias Type	Quote (non-biased)	Type of Negative
Sex	*In 1971, Raisman and Field reported that female rats have more dendritic spines in the preoptic area of the hypothalamus (POA) than do males.*	Explicit (EN)
Age	*In general, trends attributed to colorectal cancer screening in patients $>$ 50 although potential impact from changes in modifiable risk factors.*	Explicit (EN)
Geography	*Oral and oropharyngeal cancer is diagnosed each year in 40,000 Americans and kills 8,000 of them.*	Implicit (IN)
Sex	*The idea that some adult behavior is influenced by the sex-steroidal milieu during development had its origins in a classic 1959 study by Phoenix and co-workers, who showed that in male guinea pigs, testosterone acts during a narrow window of time in fetal development to permanently 'organize' the brain's ability to express stereotypic sexual behavior in adulthood*	Implicit (IN)
Race	*[..] ED physicians generally prescribe fewer opioids to African Americans regardless of clinical disorder and presentation than to non-Hispanic whites [citation]. [..]*	Extracted (XN)
Gender	*Physical exam shows a woman in moderate distress with mild jaundice and a fever of 39°c.*	Extracted (XN)
Inappropriate Use of Language	*His actions were impulsive with little regard for consequences with some reports suggesting he became an alcoholic and drifter.*	Remaining (RN)

Table 2. Sample non-biased quotes for various bias types. The non-biased quotes are further divided into four types of negatives: explicit, implicit, remaining, and extracted, which will be described in Section 5.3. The final sample is an example of a sentence labeled for 'inappropriate use of language,' but was not labeled for bias; this code is common in the Remaining Negatives group.

bias in medical education and provides a scalable solution to mitigate their influence on future generations of healthcare professionals.

2 Related Work

There is growing concern among researchers about the risks and ethical implications of applying AI in healthcare without proper governance (Althubaiti 2016; Gianfrancesco et al. 2018; Nelson 2019). A major cause for this concern is that current practices used for gathering training data can ultimately lead to models that will produce biased outputs.

In this context, the term "bias" can refer to a variety of different meanings—statistical biases, psychological biases, and societal biases—all of which raise questions over the validity and reliability of some medical research. Althubaiti (2016) examines the first two types of biases in the context of epidemiological and medical research, whereas others (Gianfrancesco et al. 2018; Nelson 2019; Mittermaier, Raza, and Kvedar 2023) discuss how issues with data quality used for training models may favor certain populations in detriment of others.

Some proposals for reducing biases in medical research advocate for changes in medical education (Cavallo and Montenegro 2017; Stanciu et al. 2017). Althubaiti (2016) suggests that bias awareness should be integrated into medical education early on and stress the importance of transparency in reporting research findings. Stone and Moskowitz (2011) address the problem of non-conscious racial and ethnic bias in healthcare, proposing a workshop-based intervention for health care professionals.

Research in trustworthy and explainable AI also aims to create more reliable, fair and equitable models for healthcare in trustworthy and explainable AI. By carefully integrating quality, bias risk, and data fusion, it is possible to train more dependable models, while empowering them with the capability of providing explanations along with predictions, which can help to identify and mitigate residual biases (Albahri et al. 2023). In a similar vein, Kiyasseh et al. (2023) has developed a strategy which helps surgical AI systems avoid algorithmic bias in assessing surgical skills across diverse hospital settings and across surgeon sub-cohorts.

Additionally, advocacy for formal AI governance arises to ensure responsible use and accountability, alongside potential federal legislation for assessing algorithmic bias risks. Nelson (2019) argues that, in order to prevent bias and misuse, clinicians must play a critical role in overseeing AI algorithms and validating their use in healthcare, whereas Kiyasseh et al. (2023) calls for the need of explainable models so that regulatory bodies, such as the FDA, can provide and validate frameworks to manage these biases effectively.

Complementary to the previous ones, Dori-Hacohen et al. (2021) present a multifaceted approach towards establishing fairness in AI applications by merging insights from medical education, sociology, and antiracism, as part of a new framework. They introduce "bisinformation" as a new concept distinct from misinformation and call for research into its nature and mitigation. They advocate for the use of AI to identify and rectify biased or harmful information that adversely affects minority groups. This is the approach we adopt in this paper, in which we demonstrate how AI models can be used for detecting biases in medical text. Our work is related to, but distinct from, the concept of Technology Assisted Review (TAR) and high-recall applications Zou and Kanoulas (2020); Kusa et al. (2024); Song, Lee, and Afshar (2019); Gray et al. (2024) in that like TAR, we also rely on expert review.

3 Problem Definition

Motivated by the mandate to debias the curriculum at the University of Washington Medical School, our goal is to re-

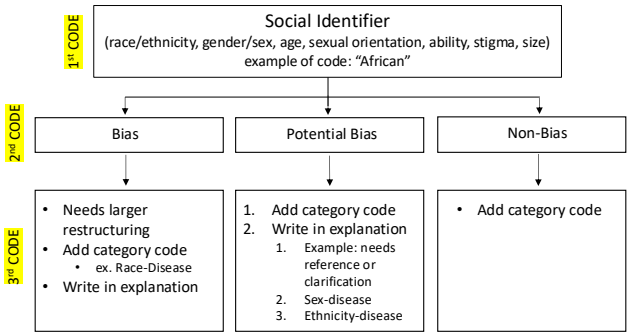

Figure 1. BRICC Coding/Labeling Procedure. If a social identifier is present in an excerpt, annotators coded the sample in 3 levels. 1st code: tokens denoting social identifier (e.g. "African American"); 2nd code: bias label ("bias", "potential bias" or "non-bias"); and, if applicable, 3rd code: identity types (e.g., "race-disease," "gender-disease") that are relevant to the bias, along with an explanation for biased quotes. An additional "review" label is reserved for cases where the annotators were unsure or could not reach consensus.

	Counts
Number of PDF Files	509
Total Number of Pages	12,647
Annotated Excerpts	4,105
Annotated Positives	1,116
Annotated Negatives	2,989
Extracted Negatives	4,391

Table 3. BRICC Dataset characteristics.

view medical educational content in an automated fashion and flag sentences with biased or potentially biased content, which will then be provided to experts for careful review[1].

We will define the problem formally as follows. Let x be an educational medical text excerpt, which can contain a claim, case report, incidence statistics, or any other textual content. Let t be a social identity which can be the target of bias, such as *"gender," "race,"* etc. Now, let $bias(x, t) \in \{true, false\}$ be a binary variable denoting whether excerpt x exhibits bias with respect to a social identity t, or not.

As one example, consider an excerpt in a skin cancer module that reads: *"Although the incidence is lower in patients of color, the morbidity/mortality can often be higher"*. This excerpt was labeled by our expert annotators (see Section 4.1 for more details) as **biased** with respect to **race**, but not with respect to other identities (gender, geography, etc.). The annotator comment indicates: *"No source for claim. Consider explaining more beyond 'patients of color' as this may come across as grouping every minority group into one and making a generalized statement."*

By extension, we can also define $bias(x, \mathcal{T}) \in \{true, false\}$ with respect to a set of social identities $\mathcal{T} = \{t_1, t_2..t_m\}$ in the following manner:

$$bias(x, \mathcal{T}) = true \iff \exists t \in \mathcal{T} \text{ s.t. } bias(x, t) = true$$

Given these definition, we now define a set of **bias classification tasks** as follows:

- **Type-specific Bias Classification:** Given a social identity type t and a text excerpt x, construct a classifier that produces a prediction for $bias(x, t)$.
- **General Bias Classification:** Given a set of social identities $\mathcal{T}$ and a text excerpt x, construct a classifier that produces a set of predictions $bias(x, \mathcal{T})$.

Therefore, given a set of $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$, we wish to construct $m+1$ classifiers (m type-specific classifiers and one general one).

4 Dataset

4.1 Data Collection and Annotation

Our team, which includes Dr. Montenegro MD PhD, an expert in sociology and race relations in medicine, has collected two years of instructional material from the University of Washington School of Medicine, totaling approximately 500 documents and over 12,000 pages, including both text and graphical content. The instructional materials include syllabi, lecture notes, and articles on a variety of topics in the medical school curricula including modules on "Lifecycle", "Mind, Behavioral and Brain", etc. Table 3 shows a summary of the dataset statistics.

The data was annotated using ATLAS.ti software by annotators, who were at least at the 4th year medical student level. The annotators were trained by Dr. Montenegro, who wrote the first draft of the guidelines, which were then refined alongside the annotators, based on their feedback. The training and annotations included a detailed coding manual of over 20 pages, which was designed in accordance with established Content Analysis techniques (Krippendorff 2018; Neuendorf 2017). Each piece of content was reviewed by a minimum of 2 annotators, often 4 or more, and discussed among the annotators until consensus labels were reached. Infrequently (only 73 examples; approx. 5.8% of annotated samples), consensus was not reached, signaling the annotators' uncertainty on this more challenging example. Overall, the team marked and subsequently annotated over 4,000 textual instances from the data. Annotators identified a broad swath of problematic text types, not all of which are related to bias. Specifically, they identified 17 different types of bias, as varied as religious bias, occupation bias, socioeconomic status bias and so forth. In this paper, we focus only on the 6 most common bias types, namely gender, sex, race, ethnicity, geography, and age bias. In what follows, we provide a brief overview of the annotation schema followed by the team, which is illustrated in Fig. 1. Two sample pages from the coding manual is reproduced in the Supplemental Materials [2].

[1] This "expert-in-the-loop" approach means that we have a preference for higher recall, and are willing to tolerate some reduced precision in order to cast a wide net. False positives will subsequently be corrected by expert review at a later stage.

[2] (Salavati et al. 2024)

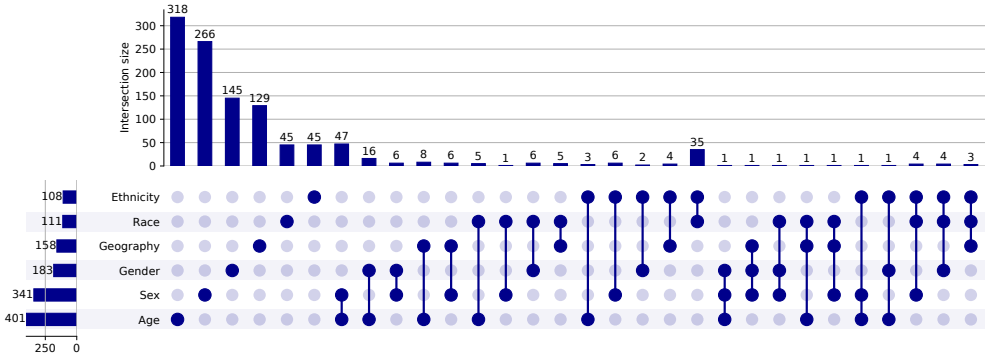

Figure 2. Histogram of intersection between sets of biased quotes. A filled-in circle indicates inclusion of specific bias type.

The **1st-level codes** highlight social identifiers present in the excerpt. The **2nd-level codes** categorize the claim in the excerpt for the presence or absence of bias, including possible potential bias[3]. At the 2nd-level, the annotators used four distinct categories—'Bias,' 'Potential Bias,' 'Non-bias,' and 'Review.' They encode the following meaning:

- **Bias:** Flagging the use of stereotypes, theories of inherent group difference and advocacy of differential medical treatment based on social identities. A statement with this code definitively needs significant restructuring at minimum to become non-biased.
- **Potential Bias:** Flagging a statement that would be non-biased, if clearly cited, appropriate data exists and provides sufficient factual basis for this claim.
- **Non-bias:** Use of social identifiers in a manner that falls under the previous categories. If assertions are based on social identities, they are based on clearly cited and medically sound, compelling evidence.
- **Review:** The annotators were unsure or could not reach a consensus on the labels, and preferred to have a senior attending physician review their work.

Then, **3rd-level codes** indicate an association between a medical condition and one or more social identifier categories (e.g., race), indicating which type of identity was discussed (whether in a biased or non-biased way). The excerpt will then be labeled with one of more codes of the form "$\$type$-disease," where $\$type$ is the specific social identifier. Our dataset included 17 different types of bias, but some only had a handful of examples in the entire corpus. Therefore, we focus on the six most frequent categories in our data: sex, gender, race, ethnicity, age, and geographical location (denoted as *geography* for brevity).

Our panel of medical experts outlined the recommended use of these social identifiers in curricular content:

- **Sex:** sex terms such as 'female', 'male', 'AMAB', 'AFAB' are acceptable when referring to population biology and are often more appropriately referenced by anatomy and genetics. Anatomic or phenotypic terms should be used with a clear purpose and are preferred (e.g., "People with ovaries"). Information about chromosomes, sex assigned at birth, and identity should be added as needed to support respectful descriptions, inclusion, and clinical reasoning.
- **Gender:** captures social identity such as 'man', 'woman', 'boy', and 'girl' for individuals. Use the patient's personally articulated gender term. Do not use gender terms when discussing population trends or outcomes– sex terms are more appropriate and specific for these descriptions.
- **Ethnicity:** inappropriate use of ethnicity (perceived shared culture) for race (perceived shared ancestry-externally imposed identity)
- **Race:** language that mistakes race (perceived shared ancestry- externally imposed identity) for ethnicity (perceived shared culture).
- **Age:** use numerical ages in describing individuals, although in some cases a descriptive term such as neonate or pediatric might be used, as is common practice in a given discipline.
- **Geography:** refers to the disproportionate representation or emphasis on diseases, conditions, treatments, and health practices that are prevalent in specific geographic regions, to the potential exclusion or marginalization of those that are more common in other regions.

In Tables 1 and 2 we showcase examples of excerpts that were identified by our annotators as either biased/potentially biased, or non-biased. Additionally, in Figure 2, we show the prevalence of various bias types in our dataset including the incidence of excerpts labeled in more than one type. This suggests that approaches learned from samples annotated for different bias types can be beneficial for type-specific bias detection. The largest intersection occurred between the following pairs: Sex and Age; Ethnicity and Race; and Gender and Age.

[3]A separate dimension of coding indicates when inadequate language was used to describe social identities, such as outdated or offensive language, which is out of scope for this paper, which focuses exclusively on bias detection.

4.2 Data Extraction and Pre-processing

The ATLAS.ti format of the BRICC dataset is optimized for human readability and annotation, but not for automated processing; therefore, we exported the data into machine readable formats. Once processed, the dataset contained the annotated excerpts, their source (document name and page number), assigned codes, and the annotators' comments. For this paper, we focused exclusively on the labels relevant to identifying biased claims, while setting aside the challenge of detecting cases in which the language is inappropriate.

Since the dataset utilizes multiple codes in order to label each excerpt, we set out to standardize these codes in a manner conducive to a machine learning setting. Figure 3 illustrates the key stages of our data collection and preprocessing pipeline: 1) data collection, as described above; 2) positive bias filter, where data samples with biases are identified and categorized; 3) negative filters, where data samples that are non-biased are are identified and categorized; 4) applying a stratified K-fold split, in order to maintain class distribution as well as better generalization.

Finally, to ensure the integrity of our evaluation and prevent any potential data leakage between the train and test sets, we also carefully checked and confirmed that text excerpts from the same document do not appear in both training and test sets. This validation step ensures that our model's performance is evaluated on truly unseen data, thereby providing a robust assessment of its generalization capabilities.

Positive Filter. For the positive and negative filters, we select the subset of the excerpts that we consider to be biased (non-biased) based on a combination of possible labels.

Let x be a text excerpt containing a medical claim and x_{codes} be the assigned codes by the annotators. Let $type_{codes}$ be the set of "$\$type$-disease" labels for a given set of types. We define a **positive** label as

$$y_{any} = \begin{cases} 1 & \text{if } \exists c_1, c_2 \in x_{codes} \text{ s.t. } c_1 \in \{\textit{bias}, \textit{potential bias}, \textit{review}\} \text{ and } c_2 \in type_{codes} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Let c_t = "$\$t$-disease" for each type t, so for example, c_{race} is the code "race-disease". We can then define our **bias type-specific labels** as follows:

$$y_t = \begin{cases} 1 & \text{if } y_{any} = 1 \text{ and } c_t \in x_{codes} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

We opt to combine 'bias', 'potential bias,' and 'review' categories because the machine learning model is intended to be used as the first-stage of an expert-in-the-loop system. Hence the model's goal is to flag (with high recall rates) sentences to be reviewed by an expert in the second stage, so we would like to flag any sentences that might possibly be biased, rather than only those certainly biased. Table 1 shows some positive samples for different types of biases from the data.

Negative Filter. The dataset contains four different types of negative (i.e., not biased) samples. We will describe each in turn. Explicit Negatives (EN) are annotated quotes that are coded with both 'non-bias' and "$\$type$-disease". Implicit Negatives (IN) are annotated quotes that are coded with "$\$type$-disease" without being explicitly labeled biased or non-biased. Collectively, we refer to these two categories as 'hard negatives' because the annotators took note of them. Remaining Negatives (RN) are additional annotated quotes that have not been coded as "$\$type$-disease", but possess other labels such as 'sex-misuse', 'gender-misuse' or 'inappropriate use of language'. Finally, all the curricular text was reviewed by our annotators, yet the vast majority of it was not labeled for bias. Therefore, we can consider any non-labeled data to be additional non-biased samples. To extract the most relevant negatives, we first remove all annotated ("positive") quotes from the texts. After that, we filter the samples which have at least one social identifier. Those sentences will compose the Extracted Negatives (XN). We refer to all four types collectively as All Negatives (AN).

Table 2 exemplifies different types of negative samples and Table 3 shows the overall statistics of our dataset.

5 Experimental Setup

To understand how we can construct effective machine learning (ML) models for detecting bias in medical curricular content, we investigate a few different strategies for building and training models. In this section, we present these strategies and detail our experimental setup.

As described in Section 3, we refer to a total of 7 tasks:

Tasks 1-6: bias type-specific classification. Given a bias type t, it consists of training a model C_t, with parameters θ_t, to predict y_t from some x containing social identifiers associated with t. Predicted label is $\hat{y}_t = \mathbb{1}\{C_t(x) > 0.5\}$.

Task 7: general bias classification. Consists of training a model C, with parameters θ, to predict y_{any} from x. The model output $C(x)$ is a probability, and the predicted label is $\hat{y}_{\text{any}} = \mathbb{1}\{C(x) > 0.5\}$, where $\mathbb{1}(.)$ is an identity function.

Figure 4 illustrates our model training and evaluation process, depicting the structured process of training an ensemble of binary bias classifiers and a multi-task bias classifier using a shared feature extraction layer across multiple folds, and ensemble classification. Each step integrates advanced text processing and machine learning techniques to enhance the reliability of bias detection.

5.1 Strategies for Building Classifiers

We consider three strategies for training models (see Fig. 4):

1. **Binary:** In order to predict general bias, we train the model based on y_{any}, without regarding specific bias types. For type-specific bias detection, we train models using y_t, for each bias type t.
2. **Ensemble:** We take the bias type-specific classifiers and combine their predictions $\hat{y}_t$ in a simple way, using a "logic OR". The rationale is that if any model trained for identifying a specific bias type predicts 'bias', then the sample should be reviewed by experts. Formally, the ensemble prediction is hence defined as follows:

$$\hat{y}^{\text{ens}}_{\text{any}}(x) = \hat{y}_1(x) \vee \hat{y}_2(x) \vee \ldots \vee \hat{y}_m(x).$$

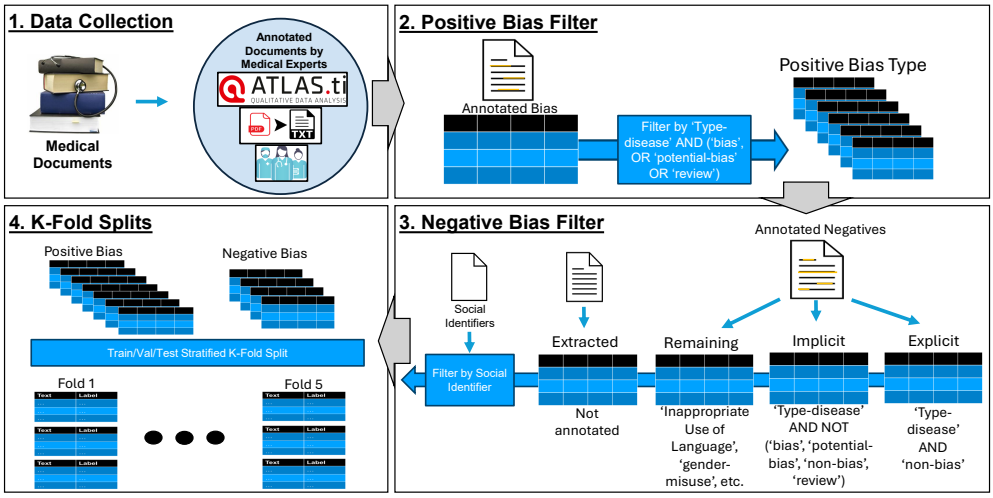

Figure 3. Overview of the data pipeline. In part 1, we collected an annotated corpus from a team of medical experts and consolidated this to one file. In part 2, we filter the positive bias types by their respective type specific -'disease'. In part 3, we filter the negative bias subsets by different labeled conditions as well as by social identifiers for respective positive bias. In part 4, we split our data into training, validation, and testing sets by using a stratified K-Fold split.

3. **Multi-task learning (MTL) classifier.** We train a single classifier capable of performing all Tasks 1-7 simultaneously via multi-task learning (Goodfellow, Bengio, and Courville 2016). The goal is to enhance the performance of each task by leveraging shared knowledge across some or all of these tasks.

   To do so, we build a neural network that has a shared backbone which splits into task-specific heads, each with one binary classification layer at the end. This allows the model to generate multiple outputs for the same input x. Denote the function computed by the backbone by $B(.)$ and those computed by task-specific heads respectively by $H_t(.)$ for tasks $t \in [1..6]$ and $H(.)$ for task 7. The MTL classifier outputs are calculated as

$$\hat{y}_t^{\text{mtl}} = \mathbb{1}\{H_t(B(x)) > 0.5\},\ t \in [1..6], \quad (3)$$

$$\hat{y}_{\text{any}}^{\text{mtl}} = \mathbb{1}\{H(B(x)) > 0.5\}. \quad (4)$$

   Loss Function. We choose to use a Weighted Binary Cross-Entropy as the loss function for each task to address class imbalance. This function achieves this goal by applying class-dependent weights to examples in the training data. These weights are dynamically adjusted based on the class distribution observed in the training data, ensuring that minority classes are appropriately emphasized during model training. The overall loss for the model is computed as a weighted sum of the losses from each task-specific classifier.

   Weighted Binary Cross-Entropy Loss (WBCE) for a Single Task. Let y denote the true label for a training example and $C(x) = P_C(y = 1|x)$ denote the probability that the example's label is positive according to classification model C. The binary cross-entropy loss for this example is given by:

$$\text{BCE}(y, C(x)) = -y \log(C(x)) - (1-y) \log(1-C(x)).$$

   When incorporating class weights, assume w_0 and w_1 are the weights for the classes 0 and 1 respectively. The weighted binary cross-entropy loss for an example (x, y) becomes:

$$\text{WBCE}(y, C(x)) = w_y \cdot \text{BCE}(y, C(x)).$$

   where w_y is selected based on the true class label y (i.e., $w_y = w_0$ if $y = 0$ and $w_y = w_1$ if $y = 1$).

   Training Loss for the Multi-Task Model. In MTL-based frameworks, it is common to assign different weights to the loss associated with each task. However, since each input is used to generate predictions for all heads and there isn't a clear ordering of the tasks in terms of importance, we opt to assign equal weights to each.

   While the inputs in the training data are the same for every task, the split between positive and negative instances differs because the positive samples for one bias type will be negative samples for another (except for multi-labeled samples). Therefore, we opt to use the task-specific class weights $w_{t,0}$ and $w_{t,1}$. Thus, the total loss used for training the MTL model is

$$L = \frac{1}{N(m+1)} \sum_t \text{WBCE}(y, C(x)), \quad (5)$$

   where $m + 1$ is the total number of tasks.

5.2 Model Construction and Architecture

We train several models based on DistilBERT (Sanh et al. 2019), which is a Transformer architecture based on the popular BERT model. We opt to use DistilBERT because it has been shown to achieve similar performance to BERT despite having 40% less parameters. This largely speeds up the training process, allowing us to experiment with several configurations and perform K-fold cross-validation.

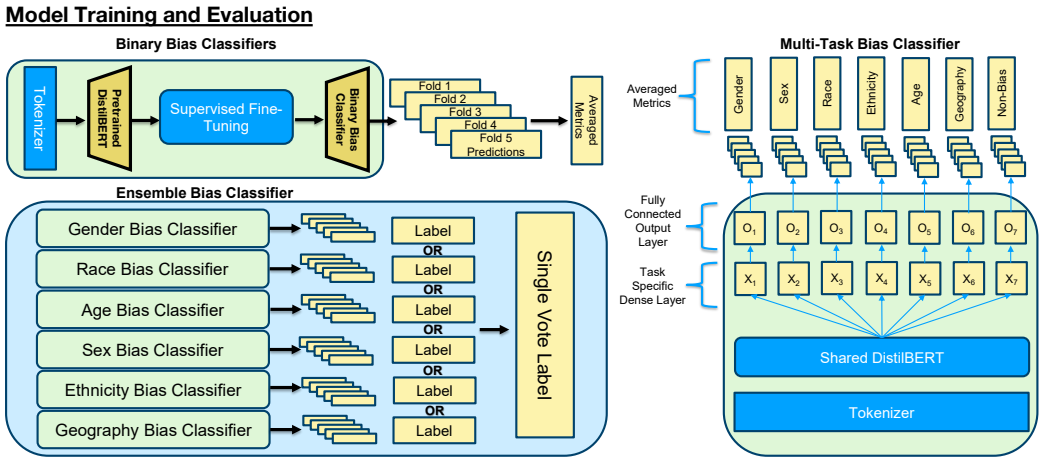

Figure 4. An overview of the model training and evaluation stage. Binary bias classifier model was used for detecting general bias as well as task-specific bias detection. The ensemble bias classifier is composed of the task-specific bias models and multitask model has 7 different tasks include bias detection as general and specific bias types.

More specifically, we fine-tune the pre-trained DistilBERT model by adding classification heads, consisting of a dense layer and classification layer. This technique, which is one variant of transfer learning (Goodfellow, Bengio, and Courville 2016), allows us to leverage the knowledge DistilBERT has acquired from large-scale language modeling tasks and apply it to our specific classification tasks. The original DistilBERT layers are frozen, while the new ones are trained by setting number of epochs to 10, learning rate to 4e-5, and batch size to 32.

Finally, we compare all our experiments to a baseline approach consisting of using FastText (Bojanowski et al. 2017) to obtain text embeddings and training a XGBoost (Chen and Guestrin 2016) (binary) classification model. The baseline was used to create 7 classifiers total, one for each bias type and one for the general bias task.

The training process was conducted over a maximum of 20 epochs with a dynamically calculated number of steps per epoch based on the batch size. The validation data, repeated and batched similarly to the training data, was used to evaluate the model performance at the end of each epoch.

For each setting, we set the prediction threshold to 0.5 and report the model performances using the standard metrics of accuracy, precision, recall, F1-score, and ROC-AUC, averaged across the five folds. Since we desire to avoid false negatives, we also include F2-score, which differs from F1 for placing more emphasis on recall than on precision.

5.3 Negative Training Set Variations

We assess the impact of different types of negative samples on the model's ability to distinguish biased from non-biased excerpts. To do so, we created **three variations of the training set** corresponding to different combinations of negative types. It is worth noting that the test set is kept constant across all different experiments associated with a given task.

The three variations we consider for defining the negative samples to be included the training data are as follows:

1. **Extracted Negatives (XN):** we use only (negative) samples with social identifiers that were automatically extracted from our medical documents.
2. **All Negatives (AN):** we use all types of negatives, including EN, IN, RN and XN.
3. **All but Remaining Negatives (AN-RN):** we exclude samples from RN but consider all the others. This is due to RN containing a variety of confounding sentences with inappropriate use of language, which we suspected may reduce the models' performance.

In contrast to most works using deep neural networks, we opt to use K-Fold cross-validation to compute the average performance based on the entire data. More precisely, we use a Stratified K-fold split with $K = 5$ to ensure consistency in class distribution across folds.

For the binary classifier for specific bias type $bias(x, t)$, we used all positive bias samples for type t and concatenated those with their corresponding negative samples in each experiment. Since the ensemble model consists of the predictions made from the specific bias classifiers, they shared the same filtering protocols as their individual classifiers. For MTL, all positive cases were used with multiple labels ($bias(x, t)$ for each type t, as well as $bias(x, \mathcal{T})$) and concatenated with their corresponding negative samples in each experiment. Finally, the binary classifier for the general bias detection was trained with the same sets of samples as MTL, but only using the $bias(x, \mathcal{T})$ label.

6 Results

In this section, we present the results of our proposed type-specific, general, ensemble and multi-task learning models and compare their performance with a baseline model in the context of specific and general bias detection tasks. Through a detailed analysis employing ROC curves and AUC metrics, we evaluate the effectiveness of our models across different

Method (Type-specific)	Precision			Recall			F1-Score			F2-Score			AUC		
	XN	AN	AN-RN	XN	AN	AN-RN	XN	AN	AN-RN	XN	AN	AN-RN	XN	AN	AN-RN
Gender	.292	**.430**	.303	**.902**	.742	.885	.434	**.539**	.448	.623	**.643**	.633	.948	.948	**.949**
Sex	.340	**.583**	.297	.839	.801	**.868**	.480	**.663**	.440	.643	**.735**	.623	.932	**.951**	.929
Race	.606	**.613**	.564	.873	.864	**.945**	.708	**.709**	.699	.796	.791	**.825**	.945	.948	**.950**
Ethnicity	.615	**.626**	.597	**.889**	.861	.852	**.725**	.722	.700	**.814**	.798	.783	**.919**	.917	.916
Age	**.452**	.445	.444	.875	**.905**	.860	.594	**.595**	.585	.735	**.747**	.723	.902	**.910**	.904
Geography	.522	**.582**	.545	**.892**	.773	.849	.657	.659	**.662**	**.780**	.721	.762	.894	**.896**	**.896**

Table 4. Performance of type-specific models for different sets of negative examples (XN: extracted negatives, AN: all negatives, AN-RN: all except remaining negatives). Best result for each pair (bias type, metric) is bold-faced.

	Method	P	R	F1	F2	AUC
Gender	Binary-Type	**.430**	**.742**	**.539**	**.643**	**.948**
	MultiTask	.416	.406	.396	.397	.868
	Baseline	.067	.333	.111	.187	.617
Sex	Binary-Type	.583	**.801**	**.663**	**.735**	**.951**
	MultiTask	**.632**	.478	.531	.496	.886
	Baseline	.097	.199	.176	.188	.576
Race	Binary-Type	.613	**.864**	.709	**.791**	**.948**
	MultiTask	**.764**	.756	**.753**	.753	**.948**
	Baseline	.018	.324	.034	.074	.548
Ethnicity	Binary-Type	.626	**.861**	**.722**	**.798**	**.917**
	MultiTask	**.712**	.695	.695	.693	.908
	Baseline	.020	.370	.038	.082	.568
Age	Binary-Type	.445	**.905**	**.595**	**.747**	**.910**
	MultiTask	**.580**	.383	.445	.403	.839
	Baseline	.117	.428	.184	.280	.634
Geography	Binary-Type	.582	**.773**	**.659**	**.721**	**.896**
	MultiTask	**.693**	.615	.645	.625	.887
	Baseline	.041	.538	.076	.158	.651

Table 5. Performance of Binary (type-specific), Multi-Task & Baseline (FastText+XGBoost) models on detection of each bias type, trained on all negatives (AN). Metrics: Precision, Recall, F1-score, F2-score, AUC. Best result for each tuple (bias type, model, metric) is bold-faced.

folds of validation, thereby ensuring robustness and consistency in our evaluations.

Firstly, we discuss the performance of models in the detection of specific bias types. The first comparison focuses on the binary classification models. Table 4 shows their performance across all of our experiment settings. We found that training with only the XN negative category leads to the high recall, at the detriment of lower precision (except for type 'ethnicity'), ultimately leading to low F1-score. Between training with AN and AN-RN, we find the results to be generally comparable, however, AN usually yields AUC that is higher or comparable to that achieved by AN-RN.

Based on the previous observation, we choose the setting AN to compare the binary classifier with the Multi-Task and baseline methods. Table 5 presents the results of the compar-

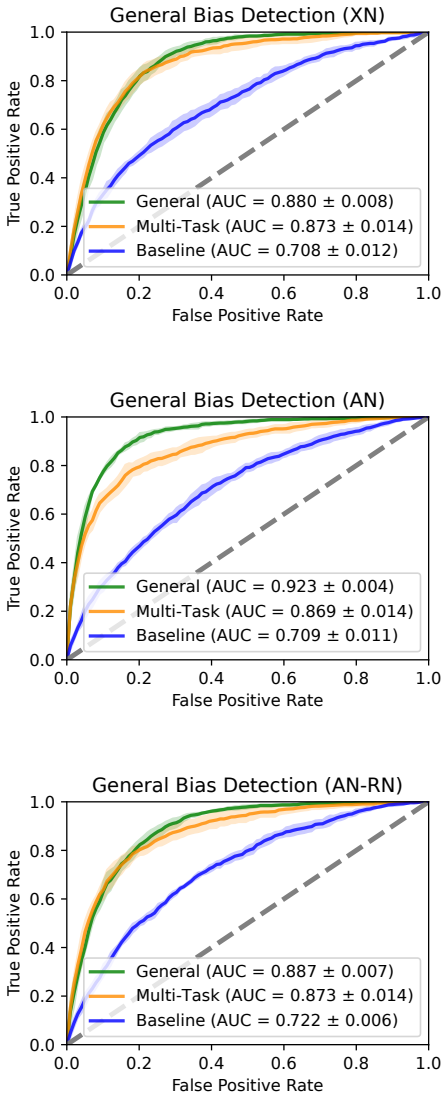

Figure 5. Mean ROC curves and standard deviation for general bias detection when training with different sets of negative samples (Top: XN, Middle: AN; Bottom: AN-RN).

Method	Precision			Recall			F1-Score			F2-Score			AUC		
	XN	AN	AN-RN	XN	AN	AN-RN	XN	AN	AN-RN	XN	AN	AN-RN	XN	AN	AN-RN
Binary-General	.314	**.504**	.335	**.925**	.812	.893	.468	**.615**	.486	.665	**.717**	.668	.880	**.923**	.887
Ensemble	.206	**.232**	.219	**.954**	.936	.951	.338	**.371**	.356	.552	**.580**	.570	-	-	-
MultiTask	.400	**.624**	.438	**.780**	.479	.724	.527	.530	**.541**	**.654**	.495	.636	**.874**	.869	**.874**
Baseline	**.288**	.263	.260	.461	.454	**.527**	**.354**	.333	.348	.411	.396	**.437**	.708	.709	**.722**

Table 6. Performance of general bias detection for different sets of negative examples (XN: extracted negatives, AN: all negatives, AN-RN: all except remaining negatives). Best result for each pair (bias type, metric) is bold-faced. Overall best for each metric is underlined. AUC is not applicable to the Ensemble method, since it is generated using a logical OR between binary outputs.

ison. Both the type-specific and multitask model greatly outperformed the baseline model. While MTL often led to the best precision, the binary classifier has outperformed MTL in all the other metrics. Furthermore, the AUC for the latter model across different tasks varied from 0.896 to 0.951, which is considered very effective at detecting bias.

Last, we compare the proposed models and the baseline on the general bias classification task. Table 6 shows the performance results for the different experiment settings and metrics. Once again, the baseline did not perform well. Considering only the proposed methods, the AN setting generally led better precision, whereas the XN setting led to better recall. As expected, the Ensemble method achieves the best overall recall, since the logical OR causes the model to be highly sensitive, leading to the lowest precision in our results. Based on F1-Score, F2-Score and AUC, the best result is achieved by the binary classifier model with the AN setting. This result is consistent with that for type-specific bias classification. For other settings (XN and AN-RN), MTL yields similar F2-Score and AUC to the binary classifier.

Figure 5 shows ROC curves for the baseline and the proposed methods (except for Ensemble, since their predictions are based on the logical OR and cannot be directly converted to probabilities). These figures highlight the superiority of the binary classifier, particularly with the AN setting. Based on Figure 5 (middle), we can see that this combination allows one to choose a threshold that achieves a high true positive rate without incurring in high false positive rates.

7 Conclusions

Our investigation into bisinformation in medical education has revealed a significant issue that, if unaddressed, can continue to negatively impact health quality, equity, and outcomes. Through our detailed empirical study, we have made substantial progress in detecting and mitigating bias in medical education materials related to gender, sex, age, geography, race, and ethnicity. This paper is, to the best of our knowledge, a first-in-class effort to eliminating biases from medical curricula using AI. In this context, our models make the first strides towards leveraging AI to promote equitable outcomes in medical education in a scalable manner, by addressing biases that disproportionately affect vulnerable communities, particularly those underserved and often misrepresented in medical literature.

This paper provides several important contributions: a robust dataset and bias detection classifier, as well as a first operationalization of the Fairness via AI framework (Dori-Hacohen et al. 2021), combining AI techniques with a deep understanding of medical educational material and systematic biases in society. Our proposed method has shown a significant capability in distinguishing biased content, showing a 27.8% improvement in AUC over a zero-shot baseline.

Limitations & Future Work. Our models are specifically trained on medical instructional materials, and designed to be deployed in an expert-in-the-loop—sometimes referred to as technology assisted review (Gray et al. 2024)—settings, and as such, we prioritize recall over precision. For obvious reasons, we are developing this software to be used with human intervention, requiring a final stage of expert review; we would never recommend that AI be solely response for changing these curricular materials directly.

Our research and annotation process has also revealed several additional issues with medical curricular data that we did not address in this work. For example, our annotators identified a large number of medical claims that were phrased in disrespectful, outdated, or non-patient-centered language; we intend to explore this orthogonal problem in future work.

We also did not train or evaluate these classifiers on other genres of content such as patient EMR data, online articles, and/or social media; we leave such efforts to future work, and expect that a transfer learning approach may be beneficial in order to leverage the annotated data and trained classifiers on the medical curricular genre. We also hope to investigate other potential biases in medical texts, such as those related to socioeconomic status, disability, neurodiversity, and intersectionality, to further understanding—and mitigate—the multifaceted and complex nature that bias plays in medical education.

Acknowledgements. This material is based upon work supported in part by the National Science Foundation Grant IIS-2147305 and the National Board of Medical Examiners Stemmler Fund 2021-3132. We thank Keen Sung for his significant contributions to an earlier version of the project, and UW BRICC research assistants and faculty who played a critical role in compiling this data set: Jadrien Gonzalez, Richard Chung, Henry Hilt, Judith Wong, Nihar Mahajan, Mike Reynolds, Michela Blain, Michelle Terry, and MB Valdez. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Board of Medical Examiners.

Ethics Statement

We include the following information for context on the ethics of our project.

Ethical Considerations. This study is rooted in making strides toward facilitating institutional change in the medical field. We are developing this software to be used with human intervention, requiring a final stage of a human review. These algorithms should not decide whether specific material contains biased information but rather be used as a tool for educators to be aware of possible bias within their curricula. It is up to each institution to ensure that we move beyond identifying bias to implementing strategies to reduce bias in curricular content. Likewise, institutions should be held responsible for ensuring their faculty work on minimizing thier own biases to avoid perpetuating bisinforamation that can potentially harm patients and their health outcomes. It is up to medical institutions to provide appropriate faculty development and resources to accomplish this large undertaking. Lastly, it's important not to burdened our under-represented faculty with this task given and inadvertently adding to the minority tax burden.

Adverse Impacts This work should never be taken as a way to penalize educators but rather a tool for fostering growth - to identify bias and allow the opportunity to correct it. When implementing this approach in practical scenarios, it is crucial to consider the context of the findings and not automatically assume that correlation implies causation. For instance, discovering higher levels of potential bias in a school's curriculum does not necessarily mean that the faculty or students are inherently more biased.

Conflicts of Interest. At the time this project was initiated, Shiri Dori-Hacohen held a significant financial interest in AuCoDe.

Researcher Positionality. The primary investigators and trainees have multiple identities that have diverse backgrounds and abilities. Participants in this work stem from multiple backgrounds including computer science, social science and medicine. These identities have influenced our specific lens on these issues and as such we acknowledge that there are both benefits and limitations stemming from our intersecting identities.

The authors note that ChatGPT was used for light editing only. The authors remain responsible for all content.

References

Acquaviva, K. D.; and Mintz, M. 2010. Perspective: are we teaching racial profiling? The dangers of subjective determinations of race and ethnicity in case presentations. *Academic Medicine*, 85(4): 702–705.

Albahri, A.; Duhaim, A. M.; Fadhel, M. A.; Alnoor, A.; Baqer, N. S.; Alzubaidi, L.; Albahri, O.; Alamoodi, A.; Bai, J.; Salhi, A.; et al. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*.

Ali-Khan, S. E.; Krakowski, T.; Tahir, R.; and Daar, A. S. 2011. The use of race, ethnicity and ancestry in human genetic research. *The HUGO journal*, 5: 47–63.

Althubaiti, A. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, 211–217.

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.

Cavallo, J.; and Montenegro, R. E. 2017. Addressing Discrimination and Bias in Medical Education. *The ASCO Post*.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Dori-Hacohen, S.; Montenegro, R. E.; Murai, F.; Hale, S. A.; Sung, K.; Blain, M.; and Edwards-Johnson, J. 2021. Fairness via AI: Bias Reduction in Medical Information. In *The 4th FAccTRec Workshop on Responsible Recommendation at RecSys*.

FitzGerald, C.; and Hurst, S. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18: 1–18.

Frey, W. R. 2020. White fragility: Why it's so hard for white people to talk about racism Robin DiAngelo.

Gianfrancesco, M.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178 (11): 1544–1547.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Gray, L.; Lewis, D. D.; Pickens, J.; and Yang, E. 2024. High Recall Retrieval Via Technology-Assisted Review. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2987–2988.

Halman, M.; Baker, L.; and Ng, S. 2017. Using critical consciousness to inform health professions education: A literature review. *Perspectives on medical education*, 6: 12–20.

Hunt, L. M.; Truesdell, N. D.; and Kreiner, M. J. 2013. Genes, race, and culture in clinical care: racial profiling in the management of chronic illness. *Medical anthropology quarterly*, 27(2): 253–271.

Karani, R.; Varpio, L.; May, W.; Horsley, T.; Chenault, J.; Miller, K. H.; and O'Brien, B. 2017. Commentary: racism and bias in health professions education: how educators, faculty developers, and researchers can make a difference. *Academic Medicine*, 92(11S): S1–S6.

Kiyasseh, D.; Laca, J.; Haque, T. F.; Otiato, M.; Miles, B. J.; Wagner, C.; Donoho, D. A.; Trinh, Q.-D.; Anandkumar, A.; and Hung, A. J. 2023. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *npj Digital Medicine*, 6(1): 54.

Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Krishnan, A.; Rabinowitz, M.; Ziminsky, A.; Scott, S. M.; and Chretien, K. C. 2019. Addressing race, culture, and structural inequality in medical education: a guide for revising teaching cases. *Academic Medicine*, 94(4): 550–555.

Kusa, W.; Peikos, G.; Staudinger, M.; Lipani, A.; and Hanbury, A. 2024. Normalised Precision at Fixed Recall for Evaluating TAR. In *The 10th ACM SIGIR/The 14th International Conference on the Theory of Information Retrieval*.

Le, T.; Bhushan, V.; Sochat, M.; Vaidyanathan, V.; Schimansky, S.; Abrams, J.; and Kallianos, K. 2020. First aid for the USMLE step 1 2020. *(No Title)*.

Madara, J. L. 2020. America's health care crisis is much deeper than COVID-19. *American Medical Association. Retrieved March*, 17: 2022.

Martin, G. C.; Kirgis, J.; Sid, E.; and Sabin, J. A. 2016. Equitable imagery in the preclinical medical school curriculum: findings from one medical school. *Academic Medicine*, 91(7): 1002–1006.

Metzl, J. M.; and Roberts, D. E. 2014. Structural competency meets structural racism: race, politics, and the structure of medical knowledge. *AMA Journal of Ethics*, 16(9): 674–690.

Mittermaier, M.; Raza, M. M.; and Kvedar, J. C. 2023. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine*, 6(1): 113.

Nelson, G. S. 2019. Bias in artificial intelligence. *North Carolina medical journal*, 80(4): 220–222.

Neuendorf, K. A. 2017. *The content analysis guidebook*. sage.

Norman, A. 2018. *Ask me about my uterus: A quest to make doctors believe in women's pain*. Bold Type Books.

Redacted for anonymity,. Under Review. Bias in Infectious Disease Fellowship Curriculum. *Clinical Infectious Diseases*.

Ripp, K.; and Braun, L. 2017. Race/ethnicity in medical education: an analysis of a question bank for step 1 of the United States Medical Licensing Examination. *Teaching and learning in medicine*, 29(2): 115–122.

Rodney, R. 2016. Decolonization in health professions education: Reflections on teaching through a transgressive pedagogy. *Canadian Medical Education Journal*, 7(3): e10.

Saini, A. 2019. *Superior : the return of race science*. Boston: Beacon Press. ISBN 0807076945.

Salavati, C.; Song, S.; Diaz, W. S.; Hale, S. A.; Montenegro, R. E.; Murai, F.; and Dori-Hacohen, S. 2024. Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes. arXiv:2407.12680.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Song, J. J.; Lee, W.; and Afshar, J. 2019. An effective high recall retrieval method. *Data & Knowledge Engineering*, 123: 101603.

Stanciu, C. N.; Ahmed, S.; Rivers Allen, J.; Bridge, W.; Fisher, A. H.; Kim, A.; Co, S.; Makani, R.; Parikh, T.; Saw, C.; et al. 2017. II. Sexual Orientation, Gender Identity, and Sex Development Competencies in Medical Education: Implications for Public Psychiatry.

Stone, J.; and Moskowitz, G. B. 2011. Non-conscious bias in medical decision making: what can be done to reduce it? *Medical education*, 45(8): 768–776.

Tsai, J.; Ucik, L.; Baldwin, N.; Hasslinger, C.; and George, P. 2016. Race matters? Examining and rethinking race portrayal in preclinical medical education. *Academic Medicine*, 91(7): 916–920.

Wingrove-Haugland, E.; and McLeod, J. 2021. Not "Minority" but "Minoritized". *Teaching Ethics*, 21(1).

Zou, J.; and Kanoulas, E. 2020. Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS)*, 38(3): 1–35.