Promoting Fairness in Classification of Quality of Medical Evidence

Simon Šuster[★] **Timothy Baldwin**[♠★] **Karin Verspoor**[♣]
★ School of Computing and Information Systems, The University of Melbourne
♣ School of Computing Technologies, RMIT University
♠ Department of Natural Language Processing, MBZUAI
simon.suster@unimelb.edu.au tb@ldwin.net karin.verspoor@rmit.edu.au

Abstract

Automatically rating the quality of published research is a critical step in medical evidence synthesis. While several methods have been proposed, their algorithmic fairness has been overlooked, even though significant risks may result when such systems are deployed in biomedical contexts. In this work, we study the fairness of two systems with respect to two sensitive attributes: participant sex and medical area. In some cases, we find important inequalities, leading us to apply various debiasing methods. Upon examining the interplay of predictive performance and fairness, as well as medically-critical selective classification capabilities and calibration performance, we find that it is possible to improve fairness through debiasing, but often at a cost to other performance measures.

1 Introduction

Automated methods for quality assessment of medical evidence have been developed to assist human experts in rating the quality of design, conduct, and reporting of published medical research. This includes predicting whether a study is affected by bias along several dimensions, or how strong the evidence is for a body of medical evidence constrained by a clinical question. A number of studies have proposed techniques and datasets for automated quality assessment (Millard et al., 2015; Marshall et al., 2020; Sarker et al., 2015; Šuster et al., 2023a), as well as follow-up research on practicality, expert acceptability, and reliability of these approaches (Gates et al., 2018; Soboczenski et al., 2019; Armijo-Olivo et al., 2020; Vinkers et al., 2021; Arno et al., 2022; Jardim et al., 2022; Šuster et al., 2023b).

Recent research has shown that machine learning (ML) techniques may suffer from bias when making decisions for people in different subgroups, which can lead to detrimental effects on the health and well-being of disadvantaged and underrepresented populations (Panch et al., 2019).[1] How to assess and mitigate such bias has been a topic of ongoing research in broader ML and natural language processing (NLP) contexts (Mehrabi et al., 2021), including in the biomedical domain (Pfohl et al., 2021; Thompson et al., 2021).

However, there has been a lack of research specifically addressing fairness and bias mitigation in automated quality assessment of medical evidence, despite unfair algorithmic decisions potentially having a large impact on either promoting or thwarting access to quality research for individual groups. A biased quality assessment classifier may systematically favour or discriminate against research conducted on participants of a specific sex or within a particular medical area. It could, for example, tend to systematically miss higher-quality evidence for Urology while working better for Cardiology.[2] By extension, the findings from studies conducted on a specific population (e.g. from Urology) or within a particular area would be undervalued or overlooked, and as a result, medical evidence relevant to those patients may not be recognized as such. An important reason for variable performance across across medical areas is the varying availability of medical evidence in the first place as well as the prevalence of higher-quality evidence (Šuster et al., 2023b). These may be grounded in different research practices and approaches to scientific assessment of interventions that have become established in medical fields (Victora et al., 2004). The consequences of such performance disparities and inequalities in the availability of medical evidence can be far-reaching, leading to outdated, ineffective, or even incorrect treatment recommendations.

[1] This notion of algorithmic bias needs to be distinguished from the bias stemming from methodology and reporting, which is formally assessed within the risk-of-bias and GRADE frameworks, as described in detail in Section 2.

[2] This example comes from our own findings.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 413–426

We aim in this paper to:

- analyse fairness of existing systems along two dimensions of protected attributes: (1) the *sex of participants*; and (2) the *medical area* of a study or a body of medical evidence. While the former is a standard attribute in the fairness literature (Sun et al., 2019), along with gender,[3] the latter extends the notion of a protected attribute to a highly multi-class group, with strong professional-ethics implications. Since medical practitioners or researchers typically work in a limited number of areas, the performance of an ML quality assessor on specific areas would be of immediate concern to them.

- show how debiasing affects different dimensions of performance, namely predictive performance, fairness, and selective classification performance (i.e., removing a model's less confident predictions), as well as how these interact in automated quality assessment.

While fairness can be understood in a number of ways (Mulligan et al., 2019), we take it to mean that all protected groups should have the same likelihood of being classified favourably (Hardt et al., 2016). That is, regardless of participant sex in a study or medical area within which a clinician works, the system should be equally likely to categorise that evidence as higher-quality. We apply bias mitigation techniques to either manipulate the data or the learning mechanism in an attempt to increase fairness.

We believe that investigating fairness in the context of quality classification of medical evidence can lead to more transparency, as well as raised awareness of potentially disparate outcomes on subgroups to which classifiers are applied.

2 Models and data

We work with two systems that differ in their intended use. EvidenceGRADEr rates the overall quality of a group of related studies (Šuster et al., 2023a; Guyatt et al., 2008), whereas TrialstreamerRoB (Marshall et al., 2015b) focuses on overall risk of bias (RoB) in a single clinical study. Next, we describe those two systems in more detail.

[3]We would like to note that Cochrane's Sex attribute used in our work refers to the biological traits, such as physiological characteristics, that generally distinguish males and females. The extent to which sex can be distinguished from gender is disputed (Tannenbaum et al., 2019).

2.1 EvidenceGRADEr

EvidenceGRADEr (Šuster et al., 2023a) is a machine learning system that performs quality assessment in the context of systematic reviews according to GRADE (Grading of Recommendations Assessment, Development and Evaluation) criteria (Guyatt et al., 2008). The system assesses a body of evidence — a set of studies included in a systematic review, grouped by a specific structured research question — and outputs predictions for various quality characteristics plus the overall quality of the body of evidence. In this work, we consider the binary classification task (low/very low vs. moderate/high quality of evidence). The system is composed of different encoders for each input feature type — a feed-forward neural network for numerical, an embedding layer for categorical, and the SciBERT language model (Beltagy et al., 2019) for textual inputs. The outputs of the encoders are composed by a top-level neural classifier. Such a system is expected to work alongside human experts to flag cases for which it is more confident, while other instances would require human review.

Data In our analysis, we use the dataset created by Šuster et al. (2023a) from a 2020 snapshot of the Cochrane Database of Systematic Reviews (CDSR) containing more than eight thousand reviews.[4] The dataset was developed by extracting and organising meta data of each review, textual parts of reviews (abstracts and summaries), tabular summaries of findings, and certain characteristics of primary studies. The two-tier grading dataset that we use in this work comes divided into 10 folds, each with its own train, development and test sets. We report the dataset statistics in Tables A1 and A2.

2.2 TrialstreamerRoB

For assessing overall RoB in a single clinical study, we broadly follow the approach in Marshall et al. (2020). We implement a system that takes as input an abstract describing the conduct and results of a clinical trial, and outputs a binary decision about whether the study is at low or high/unclear RoB. The abstract is encoded using SciBERT (Beltagy et al., 2019) and mapped to an RoB label using a feedforward neural network. The predictions of an abstract-based RoB classifier can be used to inform search rankings of medical literature in evidence exploration by clinicians or to quickly sift medical

[4]https://www.cochranelibrary.com/cdsr/about-cdsr

studies according to RoB before fine-grained RoB assessment during a systematic review.

Data We collect a large dataset of clinical trial abstracts from studies for which manual RoB annotations exist in CDSR, similarly to Marshall et al. (2015a). Starting with the PubMed identifiers for the studies included in CDSR, we then searched for abstracts using the *metapub* package,[5] obtaining a total of around 24,000 abstracts. We consider four Cochrane RoB 1 criteria (Higgins et al., 2011) modelled in previous work (Marshall et al., 2015b, 2020).[6] An overall RoB decision is labelled as "Low risk" whenever all individual criteria are at low risk, and "High risk" otherwise, following Higgins et al. (2019). Full dataset statistics appear in Tables A3 and A4.

2.3 Protected attributes

Since both datasets are derived from the same source (CDSR), we make use of the same two protected attributes readily available in CDSR:

- **Sex**, which is a subtype of Population annotated as part of Cochrane's ontology for study characterisation (Mavergames et al., 2023). It distinguishes between male and female populations, as well as allowing for an all-encompassing "male–female" category;
- Medical area (**Area**), obtained from Cochrane's topic annotations. As multiple labels are possible here, i.e., a single review can be described with more than one topic, we simply create one instance for each topic. This means that some instances are the same except for the assigned protected label.

The availability of protected group annotations varies by attribute, so we create different versions of datasets depending on the protected attribute (Sex or Area). These annotations are provided at the level of a systematic review, so we trivially linked them to data instances from our Evidence-GRADEr and TrialstreamerRoB datasets, which also have known systematic review identifiers.[7] We expect that these attributes are known ahead of prediction.

[5] https://pypi.org/project/metapub/

[6] 1) Random sequence generation, 2) allocation concealment, 3) blinding of participants and personnel, and 4) blinding of outcome assessment.

[7] As TrialstreamerRoB instances are built from individual clinical studies, we map the protected attribute obtained at review level to all the included studies.

3 Methodology and evaluation

3.1 Debiasing techniques

For our experiments, we select methods belonging to two groups of debiasing approaches based on where debiasing occurs: (1) in-data processing (pre-processing methods); or (2) in-model training (in-training methods) by adding constraints to model optimisation. To apply these techniques to our tasks, we extended the *fairlib* library (Han et al., 2022b).

Pre-processing methods. We use three different pre-processing methods: (1) Downsampling (**DownS**), which subsamples non-minority instances to derive a balanced training dataset according to a chosen objective (see next paragraph) (Kubat and Matwin, 1997; Wallace et al., 2011; Wang et al., 2019); (2) Resampling (**ReS**), which samples with replacement the instances in each subgroup to achieve a desired objective[8] (Zhao et al., 2018; Wang et al., 2019; Han et al., 2022a), and (3) Reweighting (**ReW**), which manipulates the weight of each instance in loss calculation during training. In this case, weights of different subsets of instances are derived from the empirical distribution in the training set, depending on the objective (Lahoti et al., 2020; Han et al., 2022a).

Pre-processing objectives While the above approaches describe the types of data manipulation, they can all work with different objective functions: Balanced Demographics (**BD**) (Zhao et al., 2018) encourages the model to equally focus on different demographic groups. The correlation between a group label and a class label is not explicitly captured. This objective is closely related to the Demographic Parity criterion (Dwork et al., 2012; Feldman et al., 2015). Balanced Targets (**BT**) encourages the trained model to be equally good on all target classes. In Conditional Balance of Demographics (**CBD**) (Wang et al., 2019), demographics are stratified according to the class distribution, capturing the conditional independence between a group and a target class. Conditional Balance of Targets (**CBT**) works analogously, but for target classes. In Joint Balance (**JB**) (Lahoti et al., 2020), demographics and target classes are jointly balanced. This is equivalent to using the combination of BT and CBD. Equal Opportunity (**EO**)

[8] For example, if the goal is to achieve balanced demographics and there are two protected groups, each group is under/oversampled so that their sizes are the same.

(Han et al., 2022a) balances the protected attributes within advantage classes through resampling instances based on equal opportunity objectives.

In-training methods We adopt the following approaches: (1) Adversarial training (**Adv**) (Elazar and Goldberg, 2018) extends the training objective with a discriminator component responsible for making the model unlearn the protected attributes; (2) Diverse Adversaries approach (**DAdv**) (Han et al., 2021) is a variant of Adv that adds multiple adversaries to the loss and subjects them to a diversity constraint; and (3) Fair Supervised Contrastive Loss (**FCL**) (Shen et al., 2022) builds on contrastive learning to encourage a latent space that separates instances based on target label, while mixing instances that share protected attributes.

3.2 Evaluation measures

Fairness Equality of opportunity is a widely used criterion (Hardt et al., 2016; Ravfogel et al., 2020; Han et al., 2022a). It measures the difference in true positive rate (TPR, or recall) across all groups, based on the notion that positive outcome represents a favourable decision. In our case, we view as favourable outcomes either higher quality of evidence (in the case of EvidenceGRADEr) or lower risk of bias (in case of TrialstreamerRoB). The difference (*gap*) in TPR reflects the degree to which different groups lack equal opportunity (De-Arteaga et al., 2019). The gap is calculated as variance across groups, where lower variance means greater equality. When evaluating fairness, we report $1-$gap so that higher numbers mean greater fairness. We refer to this measure as **Fairness**.

Predictive performance We report macro-averaged F1 scores. In the case of EvidenceGRADEr, the scores represent averages over 10 trials of cross-validation.

Selective classification performance Here, the model (or alternatively, the user) is granted the ability to decide which predictions should be trusted and kept (e.g. for subsequent processing by an expert), and which should be rejected (e.g. requiring a complete re-assessment). The intuition behind selective prediction is to reduce the error rate (risk) by sacrificing coverage, i.e., the proportion of all data points eligible for classification. In real-life applications, a practitioner would prefer — when comparing two models for selective prediction and for some maximum permissible error rate — the model with better coverage. Alternatively, coverage can be fixed and the model with better discrimination capability selected.

To evaluate a system's selective classification capabilities, we impose a confidence threshold τ on model predictions, keeping those that exceed it, and discarding others. The effect can then be captured in a risk–coverage curve that displays the trade-off between the risk of error and the coverage across the entire spectrum of τ (Ding et al., 2020; Geifman and El-Yaniv, 2017). To obtain a single-value conveying the significance of this trade-off, we calculate the area under the risk–coverage curve (AURC), where a smaller value indicates a better selective-prediction performance. Finally, we report $\mathbf{1-AURC}$ in our experiments for consistency with other evaluation metrics (i.e., higher is better).

Calibration One step towards understanding whether a model can be trusted is by analysing whether it is calibrated (Jiang et al., 2012; Desai and Durrett, 2020). A calibrated model gives us a signal that it "knows what it doesn't know", which can make the model easier to deploy in practice. A model is calibrated if the confidence estimates of its predictions are aligned with the empirical likelihood of the model being correct. The difference between the two is *calibration error* (Guo et al., 2017). In our analysis, we report the average over all predictions, known as expected calibration error (ECE), as well as the maximum calibration error (MCE). We empirically approximate calibration error by first discretising the probability interval into 20 bins containing an approximately equal number of predicted probabilities, a procedure known as "adaptive binning" (Nixon et al., 2019). We then calculate the average offset between the average confidence score and the proportion of samples belonging to the positive class (Guo et al., 2017). As above, we report $\mathbf{1-ECE}$ ($\mathbf{1-MCE}$) for consistency with other evaluation metrics.

3.3 Model selection

The results reported in the empirical part of the paper are based on test sets using a model found to perform best on a development set. All models are trained for 3 epochs with a patience of 1. We fine tune the debiasing hyperparameters individually for each model and for each protected attribute,[9] For pre-processing debiasing methods,

[9]For a total of $T \times P \times S = 172$ combinations, where T is the number of tasks (2), P the number of protected attributes

the hyperparameter space is defined as the set of objective functions described in Section 3.1, whereas for in-training methods we finetune the lambda parameters controlling the strength of debiasing as per the suggestions of Han et al. (2022b). Other hyperparameters are left as default values.

To select the best epoch and the best hyperparameters, we use a **DTO** (distance to optimum) criterion that combines three different measures of performance into a single figure of merit. The original formulation proposed in Han et al. (2022a) uses two criteria, namely Fairness (as defined in Section 3.2) and accuracy, to calculate Euclidean distance of normalised scores to a hypothetical system achieving perfect scores (Vincent et al., 1983). We make two adjustments to this formulation: (1) we replace accuracy with F1 as a preferred evaluation measure due to class imbalance; and (2) in addition to Fairness, we include 1−AURC as a measure of selective classification performance, adding a third criterion that we deem critical in our tasks. The calculation of Euclidean distance straightforwardly extends from two to three dimensions.

4 Results

The results for EvidenceGRADEr and TrialstreamerRoB with different protected attributes are shown in Tables 1 to 4. Looking at the models without debiasing first ("vanilla"), the predictive performance is somewhat higher for EvidenceGRADEr (around .71 F1) than TrialstreamerRoB (.66 F1).[10] Reasons for this are likely varied but could include the fact that inputs to TrialstreamerRoB are abstracts only in our setting, while the overall RoB may only be discernible from finer-grained judgments that require access to full texts. The models otherwise perform similarly in terms of selective classification and calibration.

The non-enhanced fairness of the vanilla models is highest for EvidenceGRADEr+Sex, where we find only small differences in TPR between groups (Figure 1). For TrialstreamerRoB+Sex, as well as for both models with the Area attribute (Figure 2),

Method	F1	Fair.	1−AURC	1−MCE	1−ECE
vanilla	.718	.940	.847	.845	.936
DownS	−.000	−.005	−.014	+.012	+.011
ReS	−.009	−.020	−.012	−.050	−.041
ReW	−.025	−.117	−.024	+.004	+.010
Adv	−.010	−.006	−.002	+.010	+.005
DAdv	−.010	+.009	−.002	+.041	−.002
FCL	−.310	+.030	−.276	−.001	+.007

Table 1: Main results for EvidenceGRADEr with Sex as protected attribute. Methods other than the "vanilla" method involve debiasing.

Method	F1	Fair.	1−AURC	1−MCE	1−ECE
vanilla	.714	.863	.839	.800	.911
DownS	−.001	+.011	+.001	+.031	+.010
ReS	−.012	+.009	−.017	−.054	−.050
ReW	+.007	+.013	−.013	+.047	+.027
Adv	−.000	+.013	+.005	+.016	+.002
DAdv	+.006	+.003	−.001	+.038	+.013
FCL	−.042	+.023	−.056	+.059	+.042

Table 2: Main results for EvidenceGRADEr with Area as protected attribute. Methods other than the "vanilla" method involve debiasing.

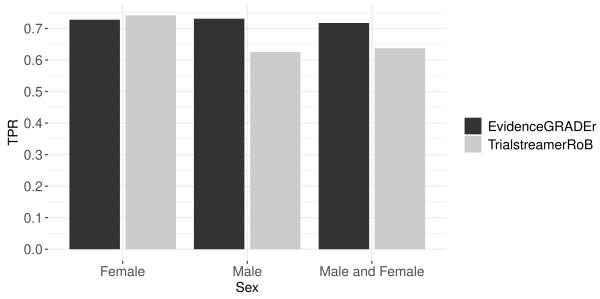


Figure 1: Variation in TPR per protected group (Sex), for the two models without fairness correction.

the differences are substantial. On Area, they range from as little as .47 up to .84.

In relation to this variability, we refer to relevant prior work on the characteristics of quality assessment data in Cochrane reviews (Šuster et al., 2023b). As the amount of evidence and prevalence of positive instances varies substantially, this may affect the ML outcomes that we observe.[11] For some groups, there is comparatively less research available. For example, sex-specific evidence is in

(2), and S is the total number of hyperparameter settings for different techniques (43). For EvidenceGRADEr, we tune hyperparameters only on the development set of the first fold of cross-validation, and use the best setting for the remaining folds.

[10] Marshall et al. (2020) report an F1∼0.5 for RoB assessment in Trialstreamer. Millard et al. (2015), whose approach is markedly different from ours (e.g. their model predicts individual criteria rather than overall risk), report AUC (∼0.69) instead of F1.

[11] For a related problem of spin, i.e., unjustified positive reporting of trial results, extensive literature exists that supports varying prevalence of this phenomenon across medical specialties: from lower (32–47%), found in anaesthesiology, surgical research, cancer, and obesity (Kinder et al., 2019; Fleming, 2016; Vera-Badillo et al., 2016; Austin et al., 2019); to higher (57–71%), found in cardiovascular research, otolaryngology, and wound care (Khan et al., 2019; Cooper et al., 2019; Lockyer et al., 2013).

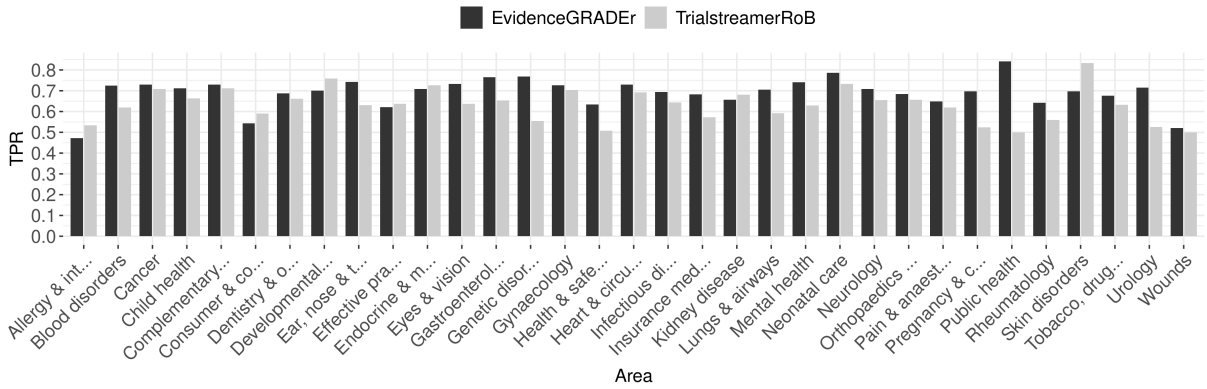


Figure 2: Variation in TPR per protected group (Area), for the two models without fairness correction.

Method	F1	Fair.	1−AURC	1−MCE	1−ECE
vanilla	.656	.862	.859	.760	.927
DownS	−.003	−.083	−.001	−.002	−.020
ReS	+.008	−.023	−.033	−.079	−.050
ReW	−.011	+.045	−.010	+.039	−.030
Adv	−.011	−.026	+.009	+.064	+.034
DAdv	+.012	−.056	+.005	+.096	+.002
FCL	−.015	+.047	−.072	+.041	+.007

Table 3: Main results for TrialstreamerRoB with Sex as protected attribute. Methods other than the "vanilla" method involve debiasing.

Method	F1	Fair.	1−AURC	1−MCE	1−ECE
vanilla	.663	.876	.855	.755	.914
DownS	+.007	−.015	−.027	−.003	−.028
ReS	−.011	−.002	−.022	−.050	−.015
ReW	+.008	−.004	−.006	−.067	−.025
Adv	+.016	−.015	−.002	+.053	−.001
DAdv	+.002	−.017	+.006	+.037	−.008
FCL	−.022	−.012	+.005	+.087	+.010

Table 4: Main results for TrialstreamerRoB with Area as protected attribute. Methods other than the "vanilla" method involve debiasing.

minority in our datasets with only around 13–22% of data points belonging to Female, and as few as 1–2% to Male (Tables A1 and A3 in the Appendix). We also see that higher-quality evidence is disproportionately low for some groups. An example is Public health, which can be explained by different research practices and nature of the area (Victora et al., 2004). However, such areas should not be disadvantaged according to the equal opportunity principle during ML-based quality assessment.

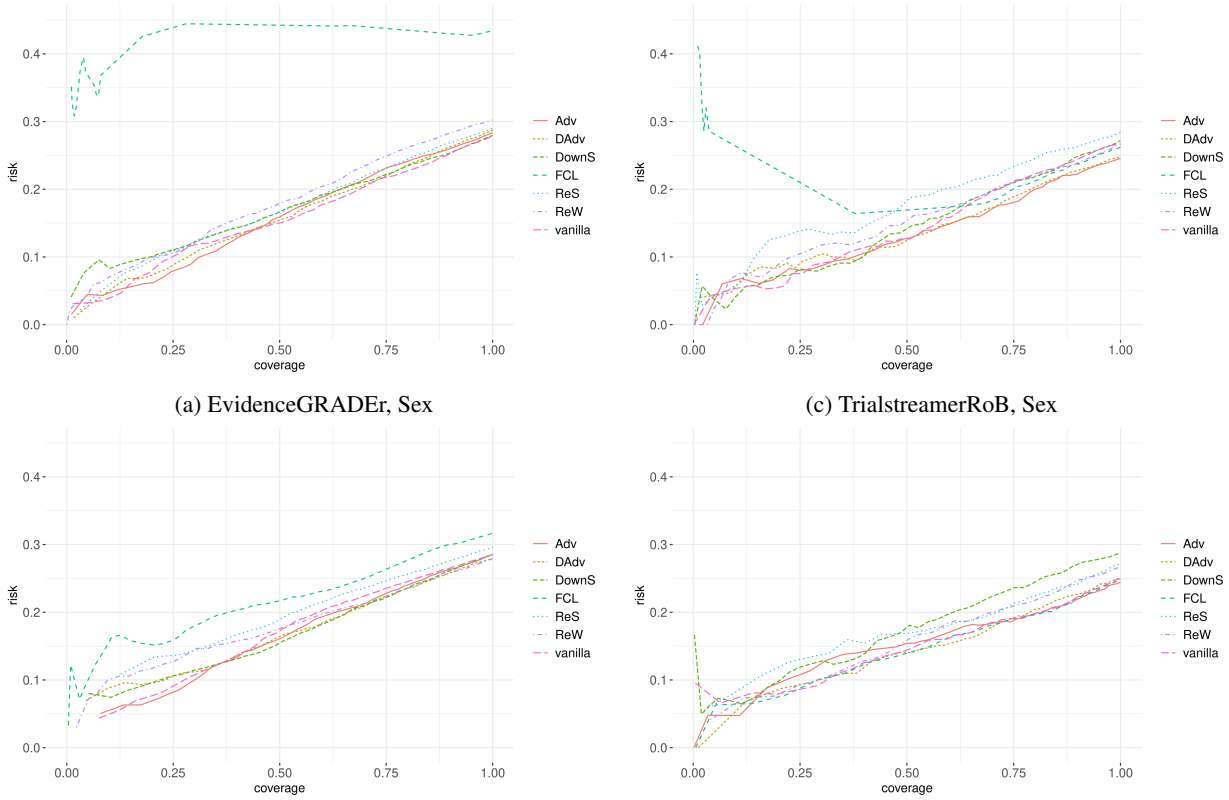
4.1 Effect of debiasing

As shown in Tables 1 to 4, debiasing can improve fairness in certain cases, especially for EvidenceGRADEr+Area. However, there is no single method that always works, which makes drawing any conclusions difficult. As the results are for models selected based on DTO (Section 3.3), this amounts to choosing a good all-rounder model. Because of that, aspects of performance other than fairness may sometimes increase, which can be seen in the results. There is no ideal situation where all main performance measures (F1, Fairness, 1−AURC) would increase, however. Often, enhanced fairness comes at a price of reduced predictive or selective classification performance, adding to the evidence on the accuracy–fairness trade-off (Han et al., 2021; Berk et al., 2023).

We inspect selective classification performance separately in Figures 3a to 3d. In most cases, the risk of error decreases steadily as we reduce coverage, which is the expected behaviour. Adversarial debiasing appears to work well, outperforming the vanilla model in the case of EvidenceGRADEr+Area and TrialstreamerRoB+Sex. Using no fairness correction still shows good risk–coverage trade-offs overall. It is clear from the figures that two debiasing techniques, namely FCL and ReS, cannot be recommended as they often lead to an increased risk of error.

4.2 Effect of model selection

Here, we explore whether choosing another selection criterion will affect fairness. The hyperparameter settings leading to the best results on the development sets are shown in Table A5. Compared to DTO-based results for TrialstreamerRoB, we find that optimising directly for Fairness leads to models that more often have substantially higher fairness (Tables 5 and 6). However, these large increases go hand in hand with even a larger drop in F1. As AURC is related to F1 (partly determined



(a) EvidenceGRADEr, Sex

(c) TrialstreamerRoB, Sex

(b) EvidenceGRADEr, Area

(d) TrialstreamerRoB, Area

Figure 3: Risk–coverage curves showing the effect of debiasing at various rejection thresholds.

Method	F1	Fair.	1−AURC	1−MCE	1−ECE
vanilla	.656	.862	.859	.76	.927
DownS	−.237	+.138	−.030	−.038	−.035
ReS	+.008	−.023	−.033	−.079	−.050
ReW	−.011	+.045	−.010	+.039	−.030
Adv	−.237	+.138	−.035	+.113	+.017
DAdv	−.002	−.047	+.007	+.076	+.006
FCL	−.237	+.138	+.072	+.052	−.022

Table 5: Results for TrialstreamerRoB when using Fairness as a model selection criterion. Protected attribute: Sex.

Method	F1	Fair.	1−AURC	1−MCE	1−ECE
vanilla	.663	.876	.855	.755	.914
DownS	−.187	+.04	−.029	+.048	+.021
ReS	−.011	−.002	−.022	−.050	−.015
ReW	−.006	−.000	−.008	+.112	+.025
Adv	−.249	+.124	−.009	+.033	−.004
DAdv	+.002	−.017	+.006	+.037	−.008
FCL	−.249	+.124	+.021	−.098	+.003

Table 6: Results for TrialstreamerRoB when using Fairness as a model selection criterion. Protected attribute: Area.

by F1 at full coverage), our experiments suggest that a similar trade-off exists between selective classification performance and fairness.

4.3 Gaps between groups

Next, we look at how debiasing reduces the TPR gaps (i.e., increases fairness) in cases where it works. How is it equalising TPR across groups? As the first case in point, we look at the application of ReW and FCL to TrialstreamerRoB+Sex (Table 3). They provide numerical evidence for substantially enhanced fairness, while maintaining competitive F1. However, when inspecting individual changes in TPR after debiasing (Table 7), we notice that TPR of *all* groups decreases. This is noteworthy because it implies that a fairer model is obtained by harming the TPR of each group.

We find a similar situation in the case of EvidenceGRADEr+Area. Here, TPR increases on several groups but decreases on others, as shown in Figure 4 when using Adv debiasing. We observe a similar pattern with other debiasing methods that display increased Fairness in Table 2.

Because of the above, we think it is necessary to look not only at the change of the aggregate fairness measure after debiasing but also at indi-

Sex	ΔTPR$_{ReW}$	ΔTPR$_{FCL}$
Male	−.042	−.042
Female	−.041	−.024
Male and Female	−.009	−.018

Table 7: Per-group changes in TPR after debiasing with either ReW or FCL. The results are for TrialstreamerRoB with Sex as protected attribute.

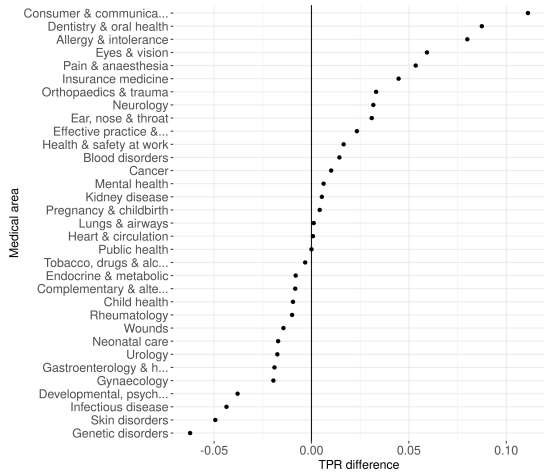


Figure 4: Positive values on the x-axis represent an increase in TPR obtained with Adv over the vanilla EvidenceGRADEr model. Negative values represent worsened TPR.

vidual group scores on which a fairness metric is based. The results of our experiments support the existence of the "levelling down" phenomenon described in Mittelstadt et al. (2023), which conveys that fairness is achieved by making every group worse off, or by bringing better performing groups down to the level of the worst off. Such solutions are unlikely to be acceptable in practice.

A road forward would be to incorporate value constraints on TPR, so that it never decreases under an admissible level. Another could be to stick to the vanilla classifier on pre-specified "advantaged" groups or groups with highest TPR, and use a fairness-enhanced classifier only on groups with lower TPR. We leave the implementation of these mechanisms for future work.

4.4 Intersectional groups

While we have investigated the protected attributes Sex and Area independently, it is possible that they may sometimes be confounding. To provide a possible explanation for varying TPR of TrialstreamerRoB from Figure 1, we examine the relationship between the groups constituting Sex and Area. Using instances with common PMIDs in the Sex and

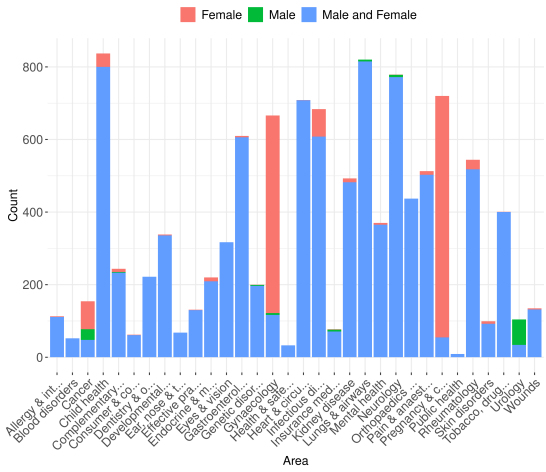


Figure 5: Contribution of each Sex group to the evidence within an Area. The counts are obtained from the intersection of training sets of TrialstreamerRoB.

Area RoB datasets, we can examine cases with both protected attributes. We calculate a contingency table based on these and show the results in a stacked bar plot (Figure 5).

There are a few areas with remarkably high occurrence of female-subject research (Gynaecology, Pregnancy & childbirth, Cancer, and Infectious diseases) and those with prominent research on male subjects (Urology, Cancer). As we saw in the fairness results for vanilla TrialstreamerRoB, females have higher TPR than other groups. As most of evidence on females is in areas with high TPR (Gynaecology, Cancer, and Infectious diseases) (Figure 2), this could help explain the high TPR in Female research. Debiasing along multiple dimensions is a complex but important avenue for future work (Subramanian et al., 2021; Lalor et al., 2022).

5 Conclusion and future work

We showed in this work that data rebalancing and training-based debiasing methods can sometimes improve fairness of quality assessment classifiers using sex or medical area as protected attributes. However, as this usually comes at the expense of predictive and selective classification performance, the decision about whether to mitigate bias should lie with domain experts who can consider the relative importance of different performance aspects.

In future work, we plan to consider using different weights for the contribution of different model selection criteria, or imposing hard (minimal) constraints on specific criteria (e.g. TPR of individual groups). Better understanding of this would enable

practitioners to "interact" with different aspects of performance in a more nuanced way.

6 Ethical considerations

In this work, we only considered two protected attributes, although many others could be used. Healthcare disparities encompass a wide range of other dimensions, including but not limited to socioeconomic status, insurance status, education status, language, age, gender, and sexual identity. The findings may differ depending on the attribute selected. Using Cochrane or PubMed meta data alone may conveniently provide many such attributes (e.g. population-related such as age, and intervention-related like disease).

7 Data and code availability

Details for obtaining CDSR data can be found in the Appendix (Section A.1). The repository with our code and the instructions to create the TrialstreamerRoB dataset is located at `https://github.com/SimonSuster/fairlib/tree/develop`.

8 Acknowledgments

We would like to thank Xudong Han and Artem Shelmanov for insightful discussions. A special thank you to Xudong for his generous assistance with the *fairlib* library.

References

Susan Armijo-Olivo, Rodger Craig, and Sandy Campbell. 2020. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Research Synthesis Methods*, 11(3):484–493.

Anneliese Arno, James Thomas, Byron Wallace, Iain J. Marshall, Joanne E. McKenzie, and Julian H. Elliott. 2022. Accuracy and efficiency of machine learning–assisted risk-of-bias assessments in 'real-world' systematic reviews. *Annals of Internal Medicine*.

Jennifer Austin, Christopher Smith, Kavita Natarajan, Mousumi Som, Cole Wayant, and Matt Vassar. 2019. Evaluation of spin within abstracts in obesity randomized clinical trials: a cross-sectional review. *Clinical obesity*, 9(2):e12292.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Richard A. Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. 2023. Fair risk algorithms. *Annual Review of Statistics and Its Application*, 10(1):165–187.

Craig M Cooper, Harrison M Gray, Andrew E Ross, Tom A Hamilton, Jaye Bea Downs, Cole Wayant, and Matt Vassar. 2019. Evaluation of spin in the abstracts of otolaryngology randomized controlled trials. *The Laryngoscope*, 129(9):2036–2040.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. 2020. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4–5.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Padhraig S Fleming. 2016. Evidence of spin in clinical trials in the surgical literature. *Annals of Translational Medicine*, 4(19).

Allison Gates, Ben Vandermeer, and Lisa Hartling. 2018. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the robotreviewer machine learning tool. *Journal of clinical epidemiology*, 96:54–62.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022a. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022b. FairLib: A unified framework for assessing and improving fairness. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–71, Abu Dhabi, UAE. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.

Julian P T Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, and Jonathan A C Sterne. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343.

Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

Patricia Sofia Jacobsen Jardim, Christopher James Rose, Heather Melanie Ames, Jose Francisco Meneses Echavez, Stijn Van de Velde, and Ashley Elizabeth Muller. 2022. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Medical Research Methodology*, 22(1):1–12.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. 2012. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Muhammad Shahzeb Khan, Noman Lateef, Tariq Jamal Siddiqi, Karim Abdur Rehman, Saed Alnaimat, Safi U Khan, Haris Riaz, M Hassan Murad, John Mandrola, Rami Doukky, et al. 2019. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: a systematic review. *JAMA network open*, 2(5):e192622–e192622.

NC Kinder, MD Weaver, Cole Wayant, and Matt Vassar. 2019. Presence of 'spin'in the abstracts and titles of anaesthesiology randomised controlled trials. *British Journal of Anaesthesia*, 122(1):e13–e14.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the International Conference on Machine Learning*, volume 97, page 179. Citeseer.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 33:728–740.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Suzanne Lockyer, Rob Hodgson, Jo C Dumville, and Nicky Cullum. 2013. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. *Trials*, 14(1):1–10.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2015a. Automating risk of bias assessment for clinical trials. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1406–1412.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2015b. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912.

Chris Mavergames, Julian Everett, Lorne Becker, Paul Wilton, and Silver Oliver. 2023. Cochrane PICO ontology. `https://data.cochrane.org/ontologies/pico/index-en.html`. Accessed: 14 April 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Louise AC Millard, Peter A Flach, and Julian PT Higgins. 2015. Machine learning to assist risk-of-bias assessments in systematic reviews. *International Journal of Epidemiology*, 45(1):266–277.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *Michigan Technology Law Review*.

Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR Workshops*.

Trishan Panch, Heather Mattie, and Rifat Atun. 2019. Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, 9(2).

Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. 2021. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113:103621.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2015. Automatic evidence quality prediction to support evidence-based decision making. *Artificial Intelligence in Medicine*, 64(2):89–103.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

Frank Soboczenski, Thomas A Trikalinos, Joël Kuiper, Randolph G Bias, Byron C Wallace, and Iain J Marshall. 2019. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making*, 19(1):96.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Simon Šuster, Timothy Baldwin, Jey Han Lau, Antonio Jimeno Yepes, David Martinez Iraola, Yulia Otmakhova, and Karin Verspoor. 2023a. Automating quality assessment of medical evidence in systematic reviews: Model development and validation study. *Journal of Medical Internet Research*, 25(e35568).

Simon Šuster, Timothy Baldwin, and Karin Verspoor. 2023b. Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area. *Journal of Clinical Epidemiology*, In press.

Cara Tannenbaum, Colleen M Norris, and M Sean McMurtry. 2019. Sex-specific considerations in guidelines generation and application. *Canadian Journal of Cardiology*, 35(5):598–605.

Hale M Thompson, Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M Churpek, Niranjan S Karnik, and Majid Afshar. 2021. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *Journal of the American Medical Informatics Association*, 28(11):2393–2403.

Francisco E Vera-Badillo, Marc Napoleone, Monika K Krzyzanowska, Shabbir MH Alibhai, An-Wen Chan, Alberto Ocana, Bostjan Seruga, Arnoud J Templeton, Eitan Amir, and Ian F Tannock. 2016. Bias in reporting of randomised clinical trials in oncology. *European Journal of Cancer*, 61:29–35.

Cesar G Victora, Jean-Pierre Habicht, and Jennifer Bryce. 2004. Evidence-based public health: moving beyond randomized trials. *American Journal of Public Health*, 94(3):400–405.

Thomas L Vincent, Walter Jervis Grantham, and W Stadler. 1983. Optimality in parametric systems. *Journal of Applied Mechanics*, 50(2):476.

Christiaan H Vinkers, Herm J Lamberink, Joeri K Tijdink, Pauline Heus, Lex Bouter, Paul Glasziou, David Moher, Johanna A Damen, Lotty Hooft, and Willem M Otte. 2021. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. *PLoS biology*, 19(4):e3001162.

Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. 2011. Class imbalance, redux. In *Proceedings of the 11th International Conference on Data Mining*, pages 754–763.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Obtaining the Cochrane dataset

Any requests from third parties to access the data set should be referred first to the Cochrane Collaboration by emailing `support@cochrane.org`. When Cochrane permits (at its discretion) the use of the data by the third party, it will grant a license to use the Cochrane Database of Systematic Reviews, including a clause that confirms that Cochrane allows us to grant third party access to the data set created in this work.

A.2 Additional details of datasets and experiments

We provide more context for the results in the following tables in the main paper.

	Train			Dev			Test		
	L	H	Total	L	H	Total	L	H	Total
Male	47	21	68	18	10	28	21	0	21
Female	738	597	1335	137	111	248	82	63	145
Male and Female	4628	3364	7992	441	418	859	460	433	893
Total	5413	3981	9395	596	540	1136	563	496	1059

Table A1: EvidenceGRADEr dataset statistics based on the first cross-validation split. Protected attribute: *Sex*. Columns: *L*: Lower quality evidence (very low and low GRADE); *H*: Higher quality evidence (moderate and high GRADE).

	Train			Dev			Test		
	Not-high	High-mod	Total	Not-high	High-mod	Total	Not-high	High-mod	Total
Allergy & ...	47	24	71	0	0	0	0	0	0
Blood diso...	161	134	295	25	9	34	31	14	45
Cancer...	395	529	924	77	62	139	105	102	207
Child heal...	2183	1661	3844	187	140	327	173	175	348
Complement...	642	420	1062	80	56	136	56	73	129
Consumer &...	34	24	58	1	2	3	3	0	3
Dentistry ...	134	48	182	17	0	17	17	7	24
Developmen...	154	121	275	15	25	40	8	7	15
Ear, nose ...	127	70	197	8	11	19	0	1	1
Effective ...	78	116	194	14	3	17	32	16	48
Endocrine ...	232	74	306	8	6	14	14	37	51
Eyes & vis...	299	142	441	8	5	13	7	12	19
Gastroente...	511	297	808	47	46	93	73	63	136
Genetic di...	74	40	114	0	10	10	14	8	22
Gynaecolog...	610	436	1046	64	50	114	53	14	67
Health & s...	182	52	234	11	6	17	19	0	19
Heart & ci...	386	430	816	43	75	118	58	86	144
Infectious...	500	460	960	34	32	66	81	58	139
Insurance ...	571	303	874	59	35	94	46	44	90
Kidney dis...	156	195	351	13	22	35	16	16	32
Lungs & ai...	509	663	1172	43	70	113	57	60	117
Mental hea...	758	592	1350	84	126	210	27	28	55
Neonatal c...	92	154	246	14	16	30	4	17	21
Neurology...	597	565	1162	26	98	124	78	115	193
Orthopaedi...	356	194	550	44	21	65	32	3	35
Pain & ana...	313	286	599	59	43	102	73	39	112
Pregnancy ...	232	235	467	50	53	103	17	47	64
Public hea...	84	26	110	0	0	0	3	16	19
Rheumatolo...	283	208	491	16	64	80	29	64	93
Skin disor...	310	214	524	61	77	138	21	19	40
Tobacco, d...	181	128	309	25	18	43	33	26	59
Urology...	213	137	350	39	13	52	21	4	25
Wounds...	120	21	141	9	8	17	9	4	13
Total	11525	9004	20529	1183	1202	2385	1212	1177	2389

Table A2: EvidenceGRADEr dataset statistics based on the first cross-validation split. Protected attribute: *Area*. Columns: *L*: Lower quality evidence (very low and low GRADE); **H**: Higher quality evidence (moderate and high GRADE).

	Train			Dev			Test		
	L	H	Total	L	H	Total	L	H	Total
Male	120	31	151	11	2	13	12	4	16
Female	1332	571	1903	146	54	200	151	65	216
Male and Female	8452	3319	11771	950	373	1323	1069	406	1475
Total	9904	3921	13825	1107	429	1536	1232	475	1707

Table A3: TrialstreamerRoB. Protected attribute: *Sex*. Columns: *L*: Lower quality evidence (high or unknown risk of bias); *H*: Higher quality evidence (low risk of bias).

	Train			Dev			Test		
	High/ unclear	Low	Total	High/ unclear	Low	Total	High/ unclear	Low	Total
Allergy & ...	108	30	138	7	4	11	15	5	20
Blood diso...	339	136	475	46	18	64	37	15	52
Cancer...	1057	475	1532	137	64	201	120	58	178
Child heal...	3818	1477	5295	408	174	582	452	198	650
Complement...	895	412	1307	86	53	139	119	51	170
Consumer &...	215	43	258	23	6	29	33	11	44
Dentistry ...	505	217	722	49	29	78	64	28	92
Developmen...	371	98	469	37	10	47	42	17	59
Ear, nose ...	199	82	281	18	14	32	35	8	43
Effective ...	465	201	666	49	25	74	56	22	78
Endocrine ...	639	237	876	84	30	114	80	33	113
Eyes & vis...	382	174	556	33	19	52	47	17	64
Gastroente...	1264	475	1739	161	61	222	160	59	219
Genetic di...	228	69	297	23	8	31	25	13	38
Gynaecolog...	772	360	1132	83	51	134	97	49	146
Health & s...	320	47	367	25	5	30	39	11	50
Heart & ci...	1661	763	2424	185	84	269	202	106	308
Infectious...	916	422	1338	96	54	150	103	65	168
Insurance ...	1101	319	1420	111	36	147	120	43	163
Kidney dis...	685	299	984	80	27	107	78	35	113
Lungs & ai...	1493	564	2057	155	76	231	177	84	261
Mental hea...	1397	509	1906	139	55	194	155	65	220
Neonatal c...	251	144	395	26	14	40	30	20	50
Neurology...	833	504	1337	93	59	152	108	66	174
Orthopaedi...	685	246	931	65	33	98	87	28	115
Pain & ana...	890	288	1178	88	36	124	87	35	122
Pregnancy ...	711	229	940	84	23	107	98	27	125
Public hea...	169	28	197	13	5	18	18	3	21
Rheumatolo...	678	292	970	84	36	120	103	39	142
Skin disor...	373	48	421	36	4	40	44	6	50
Tobacco, d...	708	317	1025	73	34	107	87	29	116
Urology...	373	96	469	39	14	53	55	8	63
Wounds...	215	24	239	22	6	28	26	1	27
Total	24717	9625	34342	2658	1167	3825	2998	1255	4254

Table A4: TrialstreamerRoB. Protected attribute: *Area*. Columns: *L*: Lower quality evidence (high or unknown risk of bias); *H*: Higher quality evidence (low risk of bias).

	EvidenceGRADEr		TrialstreamerRoB	
	Sex	Area	Sex	Area
DownS	BT	BT	BT	BT
ReS	CBT	BD	CBT	EO
ReW	CBD	EO	CBD	CBT
Adv	$10^{-1.2}$	$10^{-1.8}$	$10^{-.6}$	10^{-3}
DAdv	100	0.01	10	100
FCL	$10^{-2.6}$	$10^{-2.6}$	10^{-3}	10^{-3}

Table A5: Best hyperparameters setting per debiasing method based on DTO. For the pre-processing methods (first three), we indicate the chosen objective; for Adv, we include the chosen lambda parameter controlling the strength of adversarial regularisation; for DAdv, we include the chosen diverse lambda parameter which controls the strength of difference loss to encourage diversity of adversarial ensemble; for FCL, a single (joint) lambda parameter for strength of fair supervised contrastive loss.