Journal of the American Medical Informatics Association, 29(3), 2022, 415–423
https://doi.org/10.1093/jamia/ocab279
Advance Access Publication Date: 16 December 2021
Research and Applications

Research and Applications

Gender-sensitive word embeddings for healthcare

Shunit Agmon[1], Plia Gillis[2], Eric Horvitz[3], and Kira Radinsky[1]

[1]Computer Science Faculty, Technion - Israel Institute of Technology, Haifa, Israel, [2]Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel, and [3]Microsoft Research, Redmond, WA, USA

Corresponding Author: Shunit Agmon, MSc, Technion - Israel Institute of Technology, Haifa 3200003, Israel; shunit.agmon@gmail.com

Received 4 September 2021; Revised 30 November 2021; Editorial Decision 1 December 2021; Accepted 10 December 2021

ABSTRACT

Objective: To analyze gender bias in clinical trials, to design an algorithm that mitigates the effects of biases of gender representation on natural-language (NLP) systems trained on text drawn from clinical trials, and to evaluate its performance.

Materials and Methods: We analyze gender bias in clinical trials described by 16 772 PubMed abstracts (2008–2018). We present a method to augment word embeddings, the core building block of NLP-centric representations, by weighting abstracts by the number of women participants in the trial. We evaluate the resulting gender-sensitive embeddings performance on several clinical prediction tasks: comorbidity classification, hospital length of stay prediction, and intensive care unit (ICU) readmission prediction.

Results: For female patients, the gender-sensitive model area under the receiver-operator characteristic (AUROC) is 0.86 versus the baseline of 0.81 for comorbidity classification, mean absolute error 4.59 versus the baseline of 4.66 for length of stay prediction, and AUROC 0.69 versus 0.67 for ICU readmission. All results are statistically significant.

Discussion: Women have been underrepresented in clinical trials. Thus, using the broad clinical trials literature as training data for statistical language models could result in biased models, with deficits in knowledge about women. The method presented enables gender-sensitive use of publications as training data for word embeddings. In experiments, the gender-sensitive embeddings show better performance than baseline embeddings for the clinical tasks studied. The results highlight opportunities for recognizing and addressing gender and other representational biases in the clinical trials literature.

Conclusion: Addressing representational biases in data for training NLP embeddings can lead to better results on downstream tasks for underrepresented populations.

Key words: word embeddings, statistical models, bias, algorithms, gender

BACKGROUND AND SIGNIFICANCE

For decades, clinical trials excluded women participants.[1,2] A cited basis for this exclusion was the Thalidomide tragedy of the early 1960s, which led to pregnant women being considered as vulnerable research subjects, and for women of child-bearing potential to be excluded from early-phase clinical trials.[1] Another reason cited for excluding women was the added complexity of the menstrual cycle and its unknown effects on trial results.[3] 1993 was a turning point, when both a Food and Drug Administration guideline and the National Institutes of Health (NIH) Revitalization Act mandated that clinical trials must include women participants and to analyze results with respect to gender.[1] However, many years of clinical research performed before 1993 did not include representative numbers of women participants. Further, the gender bias in enrollment in

clinical trials has continued.[2] Despite NIH policies, clinical trial results are not analyzed with enough care about the representation and influences of gender, race, or ethnicity.[4] Although gender represents a wide spectrum and not a binary selection,[5–7] in this work we focus on binary biological sex determined by chromosomes and genitalia.

The poor representation of women in clinical trials can have grave consequences. For example, women can experience higher rates of adverse drug reactions than men.[8–10] A specific example is a sleep-inducing drug named Zolpidem. A postrelease study found that it takes longer for the drug to be cleared in women, potentially causing unexpected driving impairment on the day following use.[11]

We explore the identification and mitigation of gender biases in clinical trials data used to build predictive models with statistical natural language processing (NLP). NLP is being used in numerous healthcare applications, such as processing patient charts to predict diagnoses,[12] triaging patients in the Emergency Room,[13] and enhancing efficiencies of healthcare operations via automatic assignment of disease codes to patient records.[14] Constructing these models requires a large training corpus of text. Literature describing clinical trials has been used in prior work to construct models. For example, BioBERT,[15] a contextualized embedding model trained on PubMed abstracts and PubMed Central full-text articles,[16] achieved state-of-the-art performance in biomedical relation extraction, named entity recognition, and question answering tasks.

Diagnostic and predictive models constructed via machine learning and NLP have been shown to inherit biases latent in datasets.[17–19] Just as inferences made from gender-biased enrollments in clinical trials can lead to misdiagnoses and unexpected adverse drug reactions in women, models developed from NLP embeddings built from literature on gender-biased clinical trials can lead to gender-specific gaps in performance.[20]

Word embeddings[21–24] have grown to be central tools in learning and reasoning in language-centric applications. Embedding methods transform words into real-numbered vectors that capture their meaning in a semantic space of concepts. To date, a sizable amount of work has been done on biases in word embeddings. Numerous studies have focused on the detection and mitigation of unwanted stereotypical associations in word embeddings stemming from language usage that reflects long-term cultural biases.[25–30] Biases in word embeddings can manifest as unwanted proximities among words, for example, "homemaker" and "woman," or "engineer" and "man." Mitigations have focused on ways to remove or neutralize such unwanted associations from the embeddings and to make them gender neutral. One approach proposed for handling biases identified in clinical tasks is the strategy of removing the protected attribute (eg, gender).[20] However, for healthcare applications, gender has important physiological implications per disease prevalence and symptomatology. Gender is an important feature that can and should be used when diagnosing diseases, prescribing medications, and more.

In healthcare, unlike many other domains, important metadata about patients and situation is available and can be used to inform the training process. We describe methods in the context of gender, but the approach can be applied to other types of metadata about cases, such as demographic information, including age and race. We propose a method to train a gender-sensitive word embedding. In the approach, a specific paper abstract's contribution to the word embedding model is proportional to the number of female participants in the clinical trial. Intuitively, we wish to boost the impact of clinical trials that include more women on the embeddings, which will in turn be used for clinical prediction tasks for women. To achieve this, we merge 2 sources of data: the abstracts describing the clinical trials from PubMed, and the number of female/male participants in each respective clinical trial accessed from ClinicalTrials.gov. We draw large amounts of electronic health record (EHR) data from Maccabi Healthcare, the second largest healthcare provider in Israel, to validate our embeddings on comorbidity prediction tasks.

MATERIALS AND METHODS

Datasets

We use multiple datasets to analyze the bias in clinical trials, train gender-sensitive embedding models, and evaluate the models.

We leverage data from PubMed and ClinicalTrials.gov. PubMed[16] is a publicly available repository, containing more than 32 million citations of biomedical publications. We used the 2018 version of PubMed, which contains a total of 10 931 225 papers published between the years 2005 and 2018 (out of 18 789 150 papers in total). ClinicalTrials.gov,[31] also publicly available, contains the data and metadata regarding clinical trials. We make use of metadata on female and male participants included in the ClinicalTrials database. We used the NCT identifiers available in PubMed abstracts to match clinical trial abstracts with the metadata from ClinicalTrials.gov. After matching, the corpus contains 16 772 abstracts with available metadata.

For a source of real-world clinical outcomes and health information, we gained access to EHR data from Maccabi Health Services. This database provided diagnoses for more than 2 million patients, collected over the years 2003–2016. This database has been used in prior research.[32–35]

We compared the gender statistics on disease prevalence from the Maccabi database to those collected in the National Health Interview Survey (NHIS)[36] held in 2018 in the United States. We found that, for several main diseases, the percentages were similar, including coronary heart disease (NHIS: 4.8% in women, 8% in men; Maccabi: 4.5% in women, 9% in men), diabetes (NHIS: 9.9% in women, 10.9% in men; Maccabi: 9.2% in women, 10.6% in men), and asthma (NHIS: 15.2% in women, 11.4% in men; Maccabi: 14.6% in women, 15.6% in men). Out of the 2 million patients in the Maccabi EHR, 1.2 million patients had at least 2 diagnoses, and 51.6% of them are women. We randomly divided the patient data into 2 equally sized groups and used each one to build 2 evaluation tasks (see the "Comorbidity classification" section).

Another source of real-world clinical data is Medical Information Mart for Intensive Care III (MIMIC-III).[37] MIMIC-III is a large, freely available dataset containing de-identified data from over 40K critical care patients, collected over the years 2001–2012. It includes demographics, diagnoses, procedures, medications and more. We used this dataset to construct 2 additional clinical evaluation tasks for our models: hospital length of stay prediction ("Predicting hospital length of stay" section) and intensive care unit (ICU) readmission prediction ("ICU Readmission prediction" section).

Analyzing female inclusion in clinical trials

We hypothesize that the historical imbalance between female and male participants leads to different quantities of knowledge about women and men in healthcare. Numerous medical topics are studied less for women than for men, many times without correlation with the actual gender disease prevalence.

To validate this hypothesis, we extracted UMLS medical concepts (concept unique identifier [CUIs])[38] from the titles and abstracts of PubMed papers, using the MetaMap tool,[39] and

counted the number of men and women who participated in trials with each concept. We calculate the *female participant proportion* of a concept as the number of women in trials with this concept, divided by the total number of participants in trials mentioning this concept. Additionally, we extracted disease prevalence statistics for men and women as recorded in the EHR from Maccabi. We calculated the *female prevalence proportion* of a disease as the proportion of women in all patients diagnosed with the disease. The female prevalence proportion by diagnosis in the Maccabi EHR data and female participant proportions computed from the clinical trials data are presented in Figure 1, and an analysis over time (2008–2018) is shown for 3 diseases in Figure 2. We chose diseases from across the bias range: fibromyalgia (biased, too few men in research), diabetes mellitus (almost no bias), and spondylarthritis (biased, too few women in research).

Taking the gender-specific prevalence of diagnoses in the EHR data as representative of the larger worldwide population of prevalence, we compare the proportion of women for topics drawn from ClinicalTrials.gov with the EHR prevalence rates. In Figure 1, topics on the left show a large difference in prevalence rates in the EHR data and representation in clinical trials. We define *statistical bias* as the misalignment between gender participation in clinical trials on specific illnesses and the prevalence of the respective illnesses in a population. We see significant indications of such statistical bias. For example, 51% of Maccabi patients with liver cirrhosis are women, but only 30% or clinical trial participants on this topic are female. Topics on the right in Figure 1 have a large negative statistical bias; the proportion of female participants in trials is larger than the prevalence in female patients. Fibromyalgia specifically stands out: the prevalence in female Maccabi patients (56%) is consistent with the recent research[40] (close to 60%), while the research population is mostly composed of women (92%).

We conclude that indeed certain topics have been studied in a disproportion to the real-world prevalence of the disease in females. However, reconducting previous trials with the appropriate gender inclusion might be impractical. Excluding trials without the appropriate inclusion might lead to loss of useful scientific knowledge. In this work, we suggest augmenting the accumulated knowledge to better address the underlying bias.

Gender-sensitive embeddings

Word embeddings have been shown to be susceptible to biases in text and to aggravate them.[18,19] Thus, word embeddings trained on clinical trial papers with different participation by men and women are inherently biased, as shown in Rios et al.[41] By training *gender-sensitive* word embeddings for healthcare, we seek to address gender enrollment bias in healthcare applications.

We present a diagram of the process in Figure 3. Our approach includes a preprocessing method as follows. Each title and abstract are concatenated and processed by MetaMap[39] to identify UMLS concepts.[38] The range of words describing each concept is then replaced with the CUI, as done in Beam et al.[42] The rest of the text is lowered and tokenized, and punctuations are removed. For example, the following piece of text:

> "Effects of combination lipid therapy in type 2 diabetes mellitus."

is transformed into:

> "C1280500 of C0009429 C0023779 C0009429 in C0011860"

To create gender-sensitive embeddings, we use the number of female participants in a clinical trial abstract (available from ClinicalTrials.gov[31]) to determine the relevance of this abstract, as a unit, in the construction of a female-centric embedding. Intuitively, clinical trials with more female participants should have more influence on the embedding. The importance of an abstract is implemented as upsampling the abstract in the corpus on which the embedding is trained, thus giving the abstract more weight in the training process. We experimented with several heuristics and optimizations to determine each abstract importance (Supplementary Appendix A and D) and chose the one given below. We also train a baseline embedding, which we name *neutral embedding*, on a corpus containing each abstract exactly once, regardless of the number of participants in the trial.

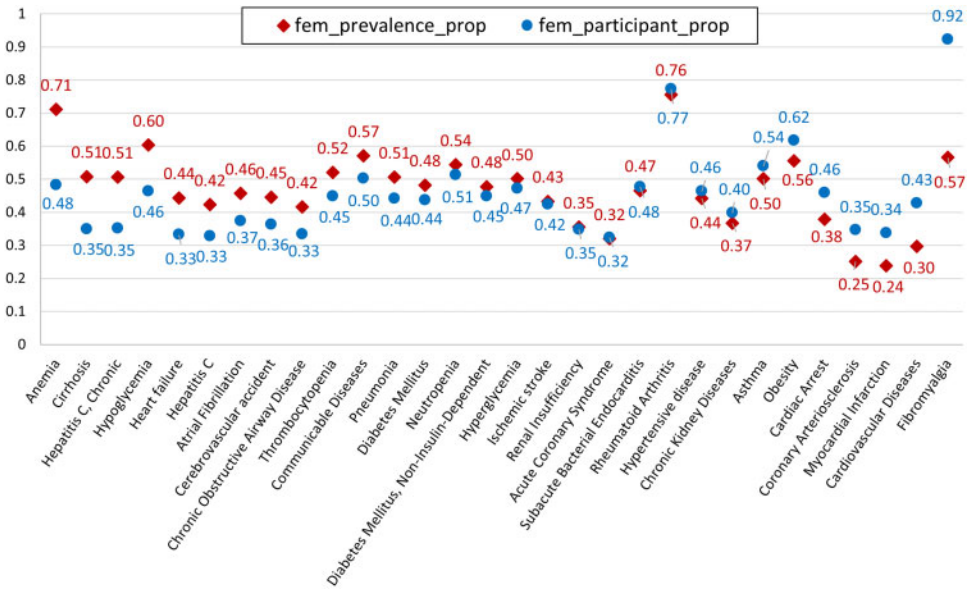

Figure 1. Proportion of female research participants drawn from ClinicalTrials.gov (red diamonds) vs computed prevalence of females with diagnoses in multi-year Maccabi electronic health record data (blue circles).

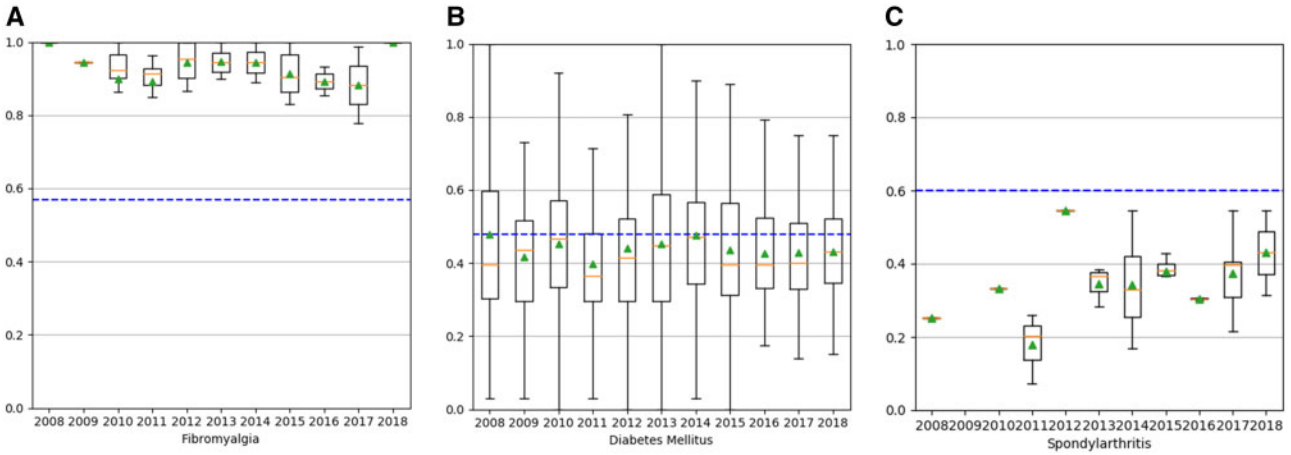

Figure 2. Gender trends over time: Proportion of female participants in clinical trials vs female prevalence drawn from electronic health record (EHR) data over a decade (2008–2018). The dashed blue line is the female prevalence calculated from Maccabi EHR data. The boxes/triangles represent the range from the first to third quartiles/means of the female participation percent in clinical trial papers by years.

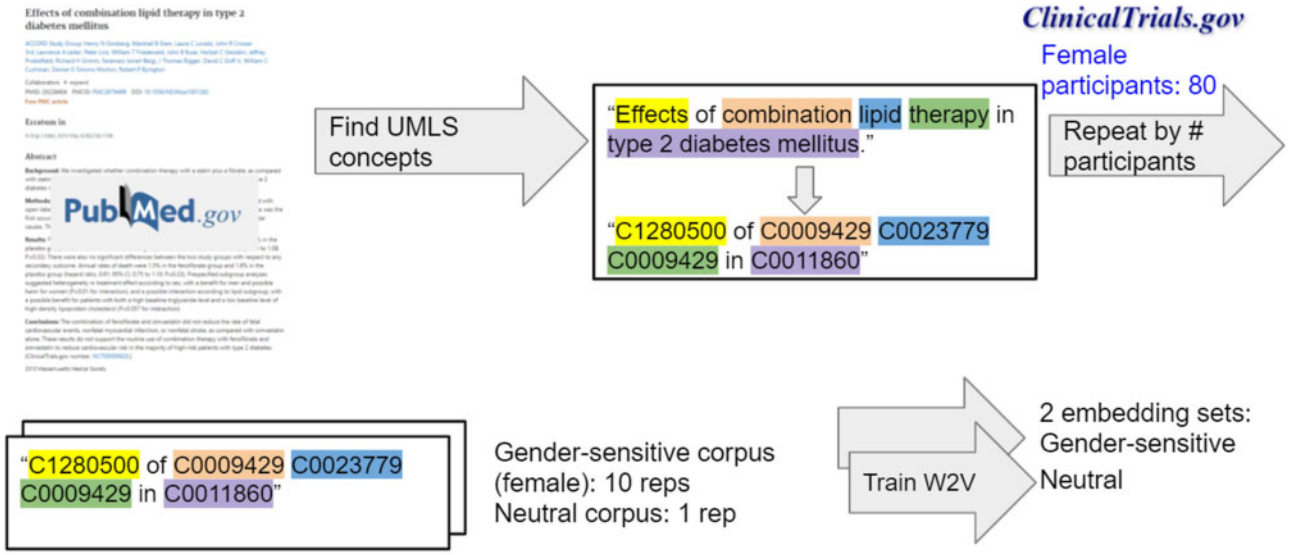

Figure 3. Training process of gender-sensitive embeddings.

The mapping from female participants to the number of repetitions is given by:

$$reps(x) = \begin{cases} 0, & x = 0 \\ 1, & 0 < x \leq 10 \\ 10, & 10 < x \leq 100 \\ 20, & 100 < x \end{cases}$$

We set a cap on the number of repetitions, to avoid giving a single abstract too much influence on the embeddings. We experimented with several other weighting policies; this policy performed best. See Supplementary Appendix A for details.

The training procedure yields a bias-sensitive embedding set and a baseline embedding set—2 different mappings from words to real-numbered vectors. We make our code and embeddings publicly available (https://github.com/shunita/gender_sensitive).

Finally, we filter only words that are CUIs and use them in our evaluation process. In the experiments below, we used embedding size of 40. There are 2118 concepts for which we have both versions of embeddings and appear in our evaluation data.

RESULTS

We evaluate our gender-sensitive embedding on a comorbidity prediction task, based on data from Maccabi EHR. Comorbidities are diseases that occur together frequently. In this task, the embedding is given as input to a model that aims to solve the task.

Next, we evaluate our embeddings on 2 tasks based on MIMIC-III data: hospital length of stay prediction and ICU readmission prediction. The embeddings are used to transform the patient's diagnoses into feature vectors, which are used as input to a prediction model. Additional implementation details on all downstream tasks and models are available in Supplementary Appendix E.

Comorbidity classification

Comorbidity classification is the task of predicting whether a disease or medical condition simultaneously presents with another or others in patients across aggregated knowledge over many patients. The task has been widely studied,[43–46] due to the increasing availability of large-scale EHR data. While most disease associations are known, others may still be uncovered via analysis of EHR data. For example, in Chaganti et al,[43] the authors find through EHR mining previously unknown connections between autism spectrum disorder and glaucoma and between Alzheimer's disease and prior inflammatory processes. We aim to evaluate the opportunities to harness the gender-sensitive embedding ability to improve the performance of a comorbidity classifier: a model which detects if 2 diseases are comorbidities.

The comorbidity binary label for our task is calculated using the 2 million patient data from Maccabi EHR. To calculate the label, we perform a statistical test for proportion difference (z-test), comparing 2 proportions: the probability of a person with disease A to be diagnosed with B, versus the probability of a person without disease A to be diagnosed with B. If A increases the chances of getting B, and B increases the chances of getting A we consider A and B as comorbidities and assign them a positive label. The comorbidity label can be calculated on a subset of the population. We calculated women's comorbidities, by counting only female patients in the compared proportions. We filter out diseases that appear in less than 30 patients, to maintain the statistical test validity.

We employ for the classifier a single-layer neural network with 50 neurons. Its input features are the concatenated embeddings of the 2 diseases, and its output is binary (whether the diseases are comorbidities).

We compared the performance of the model when given each embedding set (gender-sensitive and neutral baseline) as input features. We used 5-fold cross validation, calculated the metrics on each fold, and finally averaged them. A comparison of the average accuracy and the average area under the receiver-operator characteristic (AUROC) curve for women's comorbidities is shown in Table 1.

The gender-sensitive embedding performs statistically significantly better than the neural baseline. The performance for diseases of 3 categories (cardiovascular, autoimmune, and other commonly misdiagnosed diseases) is presented in Table 2, along with the average proportion of female participants in clinical trials and the disease prevalence in women. Diseases where the difference between the 2 models' area under the curves (AUCs) was statistically significant according to Delong's Test[47] were marked with an asterisk sign (*). The full table is available in Supplementary Appendix C.

Out of 265 diseases with at least 10 abstract mentions, in 77 (27%) diseases the gender-sensitive model performed better; in 185 (69.8%) diseases there was no significant difference; and in only 2 (1.1%) diseases the neutral model performed better. These 2 were Alzheimer's disease and gastroenteritis. A particularly interesting disease category is cardiovascular diseases, in which we saw a significantly higher AUROC in half of the diseases (7 out of 15). In the other cardiovascular diseases, the gender-sensitive model's AUROC was higher, but with P-value $> .05$.

Table 1. Accuracy and area under the receiver-operating characteristic curve (AUROC) on women's comorbidity classification task

Input features	Accuracy	AUROC
Neutral baseline	0.737	0.814
Gender-sensitive embedding	0.776	0.860

We also examined the performance on autoimmune diseases,[48] which are commonly misdiagnosed in women.[49] Our model performed significantly better on comorbidities of psoriatic arthritis, rheumatoid arthritis, and systemic lupus. In other commonly misdiagnosed diseases,[50] the performance of the gender-sensitive model was also higher.

Next, we compared the 2 embedding sets over disease sets with different levels of bias (Figure 4). We define the bias as the difference between prevalence and participants. The female prevalence of diseases is the proportion of female patients out of all patients diagnosed with the illnesses. The gender-sensitive embedding is consistently better than the neutral embedding for all disease groups. Additional analyses of the AUROC differences are in Supplementary Appendix B.

As many pair-level datasets[51] have relations between the train and the test (eg, a training set might have a pair of disease (A, B) and (B, C) and the test might have (A, C)), we repeated the experiment for patient-level comorbidity prediction based on diagnoses history, like the task of Folino and Pizzuti.[52] Each patient's previous diagnoses were aggregated by an Long Short-term Memory neural network, and fed to N binary classifiers, one for each possible future diagnosis. When averaged over diseases with participatory statistical bias (the female participant proportion is lower in the contributing studies than the female prevalence), the average AUC of the gender-sensitive model on female patients was 0.68, compared to 0.66 for the neutral model.

Predicting hospital length of stay

Next, we turn to evaluating the models on the task of predicting hospital length of stay. In this task, the goal is to predict a patient's length of stay in the hospital, based on the patients' diagnoses from the previous admissions, primary diagnosis from the current admission, and demographic features. Estimating a patient's length of stay is important in hospital planning around the allocation of rooms and resources. The predictions can also be taken as indications of severity and need for different levels of care and recovery.

The features used in this prediction task were patient demographics (gender, age, ethnicity), previous diagnoses embeddings, and primary diagnosis embedding (the first diagnosis in the admission). The embedding was done using the models evaluated as described above. We also harnessed a set of aggregated numerical features, including the number of previous admissions, number of previous procedures, number of previous diagnoses, and the number of days since last admission.

The sum of previous diagnoses was concatenated to the primary diagnosis embedding and to the other features. The combined feature vector was fed into a 3-layered neural network.

The mean absolute error (MAE) with 95% confidence intervals for both embeddings is shown in Figure 5. The gender-sensitive model achieved a lower MAE (4.60 vs 4.65). Most of the error improvement was for the female patient visits (4.59 vs 4.66) with lower levels of improvement for male patient visits (4.60 vs 4.64). We further analyzed the results by age and gender (Supplementary Appendix F), and found that, for most age groups, the gender-sensitive embedding improved the error for all patient visits, with larger improvements for female patient visits. The improvements were larger for older ages, which may be attributed to more complex

Table 2. Area under the receiver-operating characteristic curve (AUROC) of both models on women's comorbidity classification task for cardiovascular (CV), autoimmune (A), and commonly misdiagnosed diseases (M)

Disease	Neutral AUC	Gender-sensitive AUC	Female participants	Female prevalence	Category
Acute coronary syndrome (*)	0.81	0.91	0.32	0.32	CV
Aortic stenosis symptomatic	0.89	0.91	0.53	0.48	CV
Aortic valve insufficiency (*)	0.67	0.73	0.56	0.47	CV
Acute myocardial infarction	0.88	0.91	0.25	0.24	CV
Cardiac event	0.83	0.88	0.37	0.33	CV
Aortic valve stenosis	0.82	0.85	0.45	0.48	CV
Atrial fibrillation	0.82	0.84	0.37	0.46	CV
Cardiovascular diseases (*)	0.75	0.84	0.43	0.30	CV
Coronary occlusion	0.87	0.87	0.29	0.22	CV
Heart failure w. normal ejection fraction (*)	0.82	0.89	0.57	0.52	CV
Chronic heart failure	0.80	0.84	0.27	0.44	CV
Cardiac arrest (*)	0.79	0.87	0.46	0.38	CV
Heart failure, systolic (*)	0.86	0.9	0.31	0.33	CV
Acute heart failure (*)	0.85	0.90	0.35	0.44	CV
Arthritis, psoriatic (*)	0.76	0.84	0.49	0.51	A
Psoriasis	0.72	0.80	0.38	0.50	A
Inflammatory bowel diseases	0.82	0.83	0.48	0.53	A
Rheumatoid arthritis (*)	0.74	0.86	0.77	0.76	A
Lupus erythematosus, systemic (*)	0.59	0.80	0.88	0.83	A
Lupus erythematosus	0.79	0.78	0.90	0.83	A
Multiple sclerosis	0.61	0.75	0.69	0.67	A
Polycystic ovary syndrome	0.46	0.58	0.97	1	M
Sleep apnea syndromes	0.84	0.89	0.39	0.27	M
Fibromyalgia	0.70	0.71	0.92	0.57	M
Irritable bowel syndrome	0.75	0.74	0.75	0.64	M
Sleep apnea, obstructive (*)	0.75	0.85	0.38	0.22	M

Note. Diseases with an asterisk sign are ones where the difference in AUROC between the 2 models is statistically significant (P-value $< .05$ in Delong test[47]).

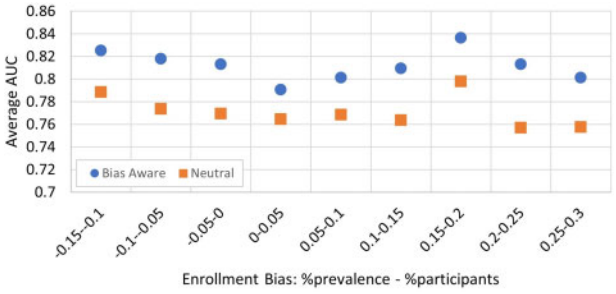

Figure 4. Average AUROC in the comorbidity classification task for the bias aware model and neutral model, analyzed by bias (the difference between female prevalence percent and female participant percent). Bins with less than 10 diseases are not shown.

relationships among multiple diseases in older ages, which are better captured by the gender-sensitive embedding.

ICU readmission prediction

Another important task on which we evaluate the gender-sensitive embedding is the prediction of unplanned readmission of a patient to the ICU, at the time of their discharge. Such readmissions indicate an unexpected deterioration in the patient's state. Detecting such cases in advance can improve the quality of care for the patients based on the prospect of allocating special programs and resources that address reasons for readmission. Studies to date have focused on predicting the likelihood of unplanned readmissions for patients, per the goal of informing decisions about programs aimed at reducing the likelihood of readmissions for patients predicted at being at high risk for readmission.[53] We follow Lin et al[54] for the definition of an unplanned readmission: Patients that were transferred from the ICU to low-level wards or discharged, but within 30 days they either returned to the ICU or died.

We use the same features as in "Predicting hospital length of stay" section, along with all the diagnoses of the current admission. Diagnoses are given as ICD9 codes, matched to CUI (if possible) and then embedded using the evaluated embedding. The final feature vector is composed of the sum of previous diagnoses vectors, the sum of current diagnoses vectors, and the additional feature vector. The vector is fed to a classifier (a 3-layer neural network).

The AUROC for both models can be seen in Figure 6. The AUROC was improved for all patients, but more so for female patients, leading to similar performance on men and women. The gender-sensitive model achieved an AUROC of 0.686. We experimented with additional features from the patient chart events table. Adding the gender-sensitive embeddings of diagnoses to these commonly used features for ICU readmission increased the AUROC from 0.68 to 0.72 for female patient visits (see Supplementary Appendix G). We deduce that adding the gender-sensitive embeddings of current previous diagnoses can improve the results of previous works as well.[55,56]

DISCUSSION

The increasing use of machine learning models in healthcare can help reduce workload of caregivers, reduce delays, and save medical

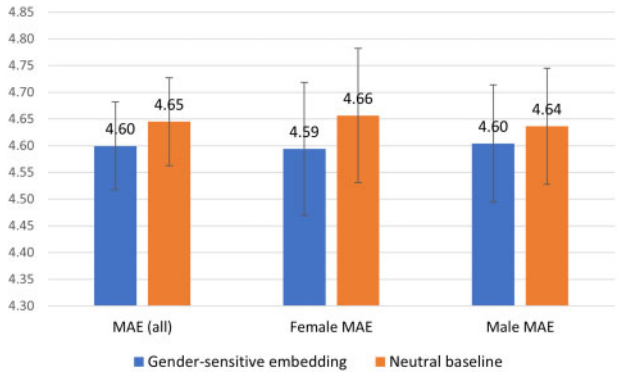

Figure 5. Mean absolute error (MAE) in prediction of length of hospital stay for the gender-sensitive embedding and the neutral baseline. The leftmost column represents the overall MAE for all visits in the test set, while the other 2 columns represent the MAE over female patient visits and male patient visits. Error bars represent a 95% confidence interval.

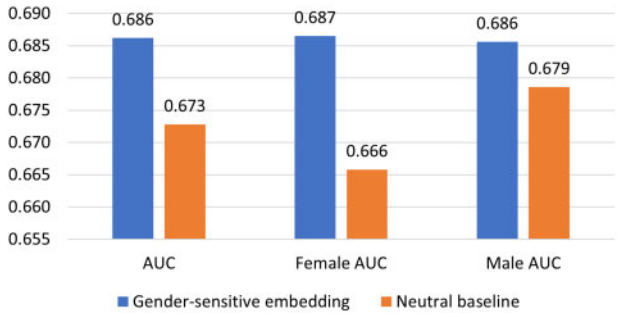

Figure 6. AUROC in ICU readmission prediction for the gender-sensitive embedding and the neutral baseline. The leftmost column represents the overall AUROC for all visits in the test set, while the other 2 columns represent the AUROC over female patient visits and male patient visits.

costs. But the corpora used to train such models contain decades-worth of underrepresentation bias.

We analyzed the female inclusion in a decade of clinical trials (2008–2018) and found gaps in research in both directions. Some topics have not been researched on enough women: liver diseases, anemia and more. In other topics, like fibromyalgia, there are more female participants than the actual prevalence.

We describe a novel approach to use clinical trial metadata to reduce the bias in machine learning models trained on them. We have designed a data augmentation method that enables the training of gender-sensitive word embeddings. Since our method is based on re-weighting of corpus samples, it can be applied easily to any embedding algorithm. We validated our method on 3 clinical prediction tasks, that were created with EHR data from 2 datasets.

In the comorbidity classification task, our model achieved an AUC of 0.86 compared to 0.81 achieved by word embeddings without data augmentation. For nearly all diseases, use of the gender-sensitive embedding either leads to either better results or to results competitive with the neutral baseline. We specifically examined cardiovascular diseases, autoimmune diseases and commonly misdiagnosed diseases and found an increase in the AUC in nearly all diseases of these categories. This suggests that the data augmentation method can lead to higher performance in health downstream tasks based on aggregated data from many female patients.

In both clinical prediction tasks based on the MIMIC-III data, the gender-sensitive embeddings improved the performance of the prediction model overall, but the largest improvement was for female patients. In the ICU readmission prediction task, the neutral baseline performed worse on female patients than on male patients; the gender-sensitive embeddings equalized the performance on both genders, by improving the performance on female patients.

We note that our study has several limitations. First, the use of disease prevalence data from Maccabi Healthcare may not be representative of worldwide disease prevalence, and suggest that the comorbidity experiments should be repeated with other EHR datasets. Second, the single binary gender variable available in both clinical trials and EHR datasets inherently limits our method to a binary concept. In ClinicalTrials.gov gender is defined as "a person's classification as female or male based on biological distinctions." In the MIMIC-III dataset, gender is defined as "the genotypical sex of the patient." These narrow definitions are at the foundation of health informatics, and they limit healthcare research per blindness to more comprehensive notions of gender and the potential influences of gender in health and wellness. We argue that dataset designers should modify these definitions to include broader notions of genders. Third, we restricted our implementation and evaluation to gender bias in clinical trials data, but the method can be tested for overcoming other types of bias in trials such as race and age.

Other than single source biases, another issue is intersectionality: biases that stem from the combination of 2 or more attributes, such as race and gender together.[57,58] Future work should explore adjusting the methods we have presented to analyze and mitigate intersectional biases. A possible method of doing so is to train a designated embedding set for each underrepresented group defined by an intersection of properties, for example, for each combination of race and gender. However, the presence of individuals of an intersectional group in clinical trials may be even lower than each single-attribute group (eg, there are fewer black women in clinical trials than women in general), leading to the availability of insufficient data to train an embedding set for each group. However, it is possible to train a specialized embedding for each population group (eg, for each gender and for each race), and to combine these embeddings to achieve a representation of each intersectional group.

Our work gives rise to several applications and future research directions. First, the performance difference between gender-sensitive and neutral word embeddings can be used to highlight diseases that should be researched more for genders poorly represented in research, typically women. As such, it can be used in a system that periodically surveys the medical research automatically and highlights research gaps.

Second, the improved performance in the tasks studied, including for comorbidity classification, suggests a possible usage of word embeddings as part of a recommendation system that detects risk factors and suggests diseases that a patient is at risk for, as well as for hospital stay prediction, and ICU readmissions.

CONCLUSION

For decades, clinical trials had poor representation of women participants. We have introduced a method for leveraging content from gender-biased clinical trials to build language-based representations for clinical classification and prediction tasks in women. We proposed and evaluated a method aimed at addressing the historical gender imbalance in clinical trials to build gender-sensitive word2-vec word embeddings. The method leverages content about clinical trials along with metadata about the degree of female representation in studies. The procedure assigns a higher weight to the content

from clinical trials that included more women. We demonstrated that use of the gender-sensitive embeddings achieves better results than the baseline, in a global clinical prediction task and 3 local (patient-level) clinical prediction tasks based on 2 EHR datasets. To our knowledge, this work represents the first effort to incorporate clinical trial metadata about the representation of participants in clinical trials to train word embeddings. We hope the work will stimulate follow-up effort, including the use of other available metadata about representation of patients in clinical trials, including patient race, ethnicity, and age, and about the type and quality of research as measured by various metrics and scales.[59]

AUTHOR CONTRIBUTIONS

SA, KR, and EH were involved in drafting or critically revising the presented work. PG provided clinical annotations and was involved in the analysis of the results. SA and KR designed the algorithm and the experiments, and EH supervised upon the algorithms and their correctness. All authors gave final approval of the submitted manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The PubMed data are publicly available from https://pubmed.ncbi.nlm.nih.gov. The clinical trials metadata is publicly available from https://clinicaltrials.gov. MIMIC-III is publicly available (https://physionet.org/content/mimiciii/1.4/) pending the completion of a training course. This study was approved by the Maccabi Healthcare Services Institutional Ethics Committee.

REFERENCES

1. Liu KA, Dipietro Mager NA. Women's involvement in clinical trials: historical perspective and future implications. *Pharm Pract (Granada)* 2016; 14 (1): 708.
2. Feldman S, Ammar W, Lo K, *et al.* Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA Netw Open* 2019; 2 (7): e196700.
3. McGregor AJ. Sex bias in drug research: a call for change. *Evaluation* 2016; 14 (34): 7887.
4. Geller SE, Koch AR, Roesch P, *et al.* The more things change, the more they stay the same: a study to evaluate compliance with inclusion and assessment of women and minorities in randomized controlled trials. *Acad Med* 2018; 93 (4): 630–5.
5. 46,XX testicular disorder of sex development. https://medlineplus.gov/genetics/condition/46xx-testicular-disorder-of-sex-development/. Accessed October 2021.
6. Matsuno E, Budge SL. Non-binary/genderqueer identities: a critical review of the literature. *Curr Sex Health Rep* 2017; 9 (3): 116–20.
7. Dhejne C, Van Vlerken R, Heylens G, *et al.* Mental health and gender dysphoria: a review of the literature. *Int Rev Psychiatry* 2016; 28 (1): 44–57.
8. Tran C, Knowles SR, Liu BA, *et al.* Gender differences in adverse drug reactions. *J Clin Pharmacol* 1998; 38 (11): 1003–9.
9. Zopf Y, Rabe C, Neubert A, *et al.* Women encounter ADRs more often than do men. *Eur J Clin Pharmacol* 2008; 64 (10): 999–1004.
10. Whitley H, Lindsey W. Sex-based differences in drug activity. *Am Fam Physician* 2009; 80 (11): 1254–8.
11. Farkas RH, Unger EF, Temple R. Zolpidem and driving impairment—identifying persons at risk. *N Engl J Med* 2013; 369 (8): 689–91.
12. Liang H, Tsui BY, Ni H, *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019; 25 (3): 433–8.
13. Horng S, Sontag DA, Halpern Y, *et al.* Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017; 12 (4): e0174708.
14. Arifoğlu D, Deniz O, Aleçakır C, Meltem TY, *et al.* CodeMagic: semi-automatic assignment of ICD-10-AM codes to patient records. In: Czachórski T, Gelenbe E, Lent R, eds. *Information Sciences and Systems*. Cham: Springer; 2014: 259–68.
15. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
16. PubMed. https://pubmed.ncbi.nlm.nih.gov. Accessed June 2021.
17. Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
18. Garg N, Schiebinger L, Jurafsky D, *et al.* Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A* 2018; 115 (16): E3635–44.
19. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017; 356 (6334): 183–6.
20. Zhang H, Lu AX, Abdalla M, *et al.* Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020; July 23–24, 2020.
21. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. *arXiv Preprint* 2013; arXiv:1301.3781.
22. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25–29, 2014; Doha, Qatar.
23. Peters M, Neumann M, Iyyer M, *et al.* Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); June 1–6, 2018; New Orleans, Louisiana.
24. Devlin J, Chang MW, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.
25. Bolukbasi T, Chang KW, Zou JY, *et al.* Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv Neural Inform Process Syst* 2016; 29: 4349–57.
26. Zhao J, Zhou Y, Li Z, *et al.* Learning gender-neutral word embeddings. *arXiv Preprint* 2018; arXiv:1809.01496.
27. Gonen H, Goldberg Y. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv Preprint* 2019; arXiv:1903.03862.
28. Kurita K, Vyas N, Pareek A, *et al.* Measuring bias in contextualized word representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing; August 2, 2019; Florence.
29. Basta C, Costa-Jussá MR, Casas N. Evaluating the underlying gender bias in contextualized word embeddings. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing; August 2, 2019; Florence.
30. Ravfogel S, Elazar Y, Gonen H, *et al.* Null it out: Guarding protected attributes by iterative nullspace projection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; July 6–10, 2020.

31. ClinicalTrials.gov – information on clinical trials and human research studies. https://clinicaltrials.gov/. Accessed June 2021.
32. Eisenberg V, Weil C, Chodick G, *et al.* Epidemiology of endometriosis: a large population-based database study from a healthcare provider with 2 million members. *BJOG* 2018; 125 (1): 55–62.
33. Levkovitch-Verbin H, Goldshtein I, Chodick G, *et al.* The Maccabi Glaucoma Study: prevalence and incidence of glaucoma in a large Israeli health maintenance organization. *Am J Ophthalmol* 2014; 158(2): 402–8.
34. Weil C, Nwankwo C, Friedman M, *et al.* Epidemiology of hepatitis C virus infection in a large Israeli health maintenance organization. *J Med Virol* 2016; 88 (6): 1044–50.
35. Weitzman D, Chodick G, Shalev V, *et al.* Prevalence and factors associated with resistant hypertension in a large health maintenance organization in Israel. *Hypertension* 2014; 64 (3): 501–7.
36. National Health Interview Survey. United States, 2018. https://www.cdc.gov/nchs/nhis/ADULTS/www/index.htm. Accessed July 2021.
37. Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 1–9.
38. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.
39. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium; November 3–7, 2001; Washington, DC.
40. Wolfe F, Walitt B, Perrot S, *et al.* Fibromyalgia diagnosis and biased assessment: sex, prevalence and bias. *PLoS One* 2018; 13 (9): e0203755.
41. Rios A, Joshi R, Shin H. Quantifying 60 years of gender bias in biomedical research with word embeddings. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical *Language Processing*; July 2020.
42. Beam AL, Kompa B, Schmaltz A, *et al.* Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv Preprint* 2018; arXiv:1804.01486.
43. Chaganti S, Welty VF, Taylor W, *et al.* Discovering novel disease comorbidities using electronic medical records. *PLoS ONE* 2019; 14 (11): e0225495.
44. Engels EA, Parsons R, Besson C, *et al.* Comprehensive evaluation of medical conditions associated with risk of non-Hodgkin lymphoma using Medicare claims ("MedWAS"). *Cancer Epidemiol Biomarkers Prev* 2016; 25 (7): 1105–13.
45. Hanauer DA, Rhodes DR, Chinnaiyan AM. Exploring clinical associations using '-omics' based enrichment analyses. *PLoS One* 2009; 4 (4): e5203.
46. Holmes AB, Hawson A, Feng L, *et al.* Discovering disease associations by integrating electronic clinical data and medical literature. *PLoS One* 2011; 6 (6): e21132.
47. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44 (3): 837–45.
48. What are autoimmune disorders? https://www.webmd.com/a-to-z-guides/autoimmune-diseases. Accessed June 2021.
49. Women's health problems doctors still miss. https://edition.cnn.com/2009/HEALTH/10/19/undiagnosed.women.problem/index.html. Accessed June 2021.
50. Inside the epidemic of misdiagnosed women. https://www.prevention.com/health/a32085516/common-misdiagnosis-women/ [Accessed June 2021].
51. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019; 47 (D1): D607–13.
52. Folino F, Pizzuti C. A comorbidity-based recommendation engine for disease prediction. In: IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS); October 12–15, 2010; Perth, WA.
53. Bayati M, Braverman M, Gillam M, *et al.* Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One* 2014; 9 (10): e109264.
54. Lin YW, Zhou Y, Faghri F, *et al.* Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS One* 2019; 14 (7): e0218942.
55. Desautels T, Das R, Calvert J, *et al.* Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. *BMJ Open* 2017; 7 (9): e017199.
56. Nguyen DP. Accurate and reproducible prediction of ICU readmissions. *medRxiv*, 2021. https://doi.org/10.1101/2019.12.26.19015909.
57. Crenshaw K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1989; 1989 (1): 139.
58. Roberts D, Jesudason S. Movement intersectionality: The case of race, gender, disability, and genetic technologies. *Du Bois Rev* 2013; 10 (2): 313–28.
59. Olivo SA, Macedo LG, Gadotti IC, *et al.* Scales to assess the quality of randomized controlled trials: a systematic review. *Phys Ther* 2008; 88 (2): 156–75.