A Comprehensive Study of Gender Bias in Chemical Named Entity Recognition Models

Anonymous ACL submission

Abstract

Chemical named entity recognition (NER) models influence numerous downstream tasks, from adverse drug reaction identification to pharmacoepidemiology. However, it is unknown whether these models work the same for everyone. Performance disparities can potentially cause harm rather than the intended good. This paper assesses gender-related performance disparities in chemical NER systems. We develop a framework for measuring gender bias in chemical NER models using synthetic data and a newly annotated corpus of over 92,405 words with self-identified gender information from Reddit. Our evaluation of state-of-the-art biomedical NER models reveals evident biases. For instance, synthetic data suggests female-related names are frequently misclassified as chemicals, especially with datasets rich in brand names. Additionally, we observe significant performance disparities between female- and male-associated data in both datasets. Many systems fail to detect contraceptives such as birth control. Our findings emphasize the biases in chemical NER models, urging practitioners to be aware of and address these biases in application.

1 INTRODUCTION

Chemical named entity recognition (NER) is the extraction of chemical mentions (e.g., drug names) from the text. Chemical NER is essential in many downstream tasks, from pharmacovigilance (O'Connor et al., 2014) to facilitating drug discovery by mining biomedical research articles (Agarwal and Searls, 2008). For instance, Chemical NER systems are the first step in pipelines developed to mine adverse drug reactions (ADRs) (Farrugia and Abela, 2020). However, manually collecting ADRs is challenging due to limitations in clinical trials, such as insufficient participants for rare ADRs, limited durations, inability to test all drug combinations swiftly, and drug repurposing leading to unexpected ADRs (Mammì et al., 2013). Hence, using chemicals to mine ADR mentions at scale can have a positive impact. However, it is unknown whether these systems perform the same for everyone. Who benefits from these systems, and who can be harmed? We present a comprehensive analysis of gender-related performance disparities of Chemical NER in this paper.

Performance disparities have recently received substantial attention in the field of NLP. For example, recent research shows performance drops in text classification models across different subpopulations such as gender, race, and minority dialects (Dixon et al., 2018; Park et al., 2018; Badjatiya et al., 2019; Rios, 2020; Lwowski and Rios, 2021; Mozafari et al., 2020). Performance disparities can manifest in multiple parts of NLP systems, including training data, resources, pre-trained models (e.g., word embeddings), and their downstream applications (Zhao et al., 2019; Goldfarb-Tarrant et al., 2021; Zhao et al., 2017). However, while previous research has explored these disparities, the focus has been largely on synthetic data and non-biomedical applications (Mehrabi et al., 2020). Our study addresses this literature gap by providing a comprehensive examination of gender-related performance disparities in Chemical NER, focusing on both synthetic and real-world data.

This paper studies a similar task like Mehrabi et al. (2020) with two primary distinctions. First, our focus is on Chemical NER, a less studied area in Biomedical NLP despite its significant bias implications. Second, while Mehrabi et al. (2020) uses synthetic data and templates (e.g., NAME in LOCATION) for bias analysis, we delve deeper into the potential biases in chemical naming, especially how they contribute to false positives. Lieven et al. (2015) highlighted a preference for linguistically feminine brand names in the market, leading drug companies to adopt such naming conventions. This can inadvertently cause models to misclassify female names as chemicals. However, only using templates might not capture the diverse writing of different groups. If a model favors certain names, how does it affect real individuals? Do biases from template data affect some groups disproportionately? For example, Sundbom et al. (2017) shows that women are more frequently prescribed antidepressants than men. If models struggle to detect these drugs, often mentioned by females, it may cause gender-specific biases in their performance. Other studies, like Riley III et al. (1998), reveal gender differences in pain sensitivity and opioid prescriptions, with women receiving opioids twice as often. The template method would not capture these differences in the model in chemical detection performance for certain classes of drugs.

Therefore, this paper presents a dual approach: we explore template data but also assemble and annotate a real-world dataset with self-identified gender information. [1] Our approach is influenced by the concerns raised by Blodgett et al. (2021) regarding many biased studies needing a sufficient understanding of the potential harm posed by the models. While we cannot fully conceptualize all potential harms, this paper moves beyond prior work focused on non-realized or synthetic datasets. We believe exploring data from people who have self-identified their demographic information is better. This will provide a more realistic understanding of how these models will perform based on how people write and what they write about.

Our main contributions are threefold.

1. We introduce and will publicly release a novel annotated Chemical NER dataset for social media data. Moreover, the dataset contains self-identified gender information to be used to measure gender bias in Chemical NER models. To the best of our knowledge, this is the first Reddit-based Chemical NER dataset. *Moreover, it is the first Chemical NER dataset with self-identified gender information.*

2. We introduce a comprehensive testing framework for gender bias in Chemical NER using both synthetic and real-world data. To the best of our knowledge, our results are the first to conduct bias analysis for chemical NER in biomedical application. This allows a better understanding of modern chemical NER techniques. *Moreover, it spurs a discourse about how information extraction methods can be biased, how the biases can be measured, and provides a framework for bias mitigation technique development.*

3. Finally, we provide a comprehensive error analysis and discussion to better understand *how* Chemical NER models can be biased. The study links biases to both chemical naming conventions and limits in current datasets with regard to gender specific chemical mentions (e.g., contraceptives).

2 RELATED WORK

Prior work extensively curated labeled data for chemical NER and developed domain-specific models. For example, the CHEMDNER corpus (Krallinger et al., 2015) was created for the 2014 BioCreative shared task on chemical extraction from text. Researchers recognize the importance of these systems and are working to make them as fair and accurate as possible. Likewise, the CDR (Li et al., 2016) dataset was developed to detect chemical-disease relations for the 2015 shared task. Similar to traditional NER tasks (Li et al., 2020), a broad range of approaches have been proposed to detect Chemicals (Rocktäschel et al., 2012; Chiu et al., 2021; Lee et al., 2020; Sun et al., 2021; López-Úbeda et al., 2021; Weber et al., 2021), from traditional conditional random fields to deep learning methods. Many recent neural network-based advances can be broken into three main groups of models, word, character, and contextual embedding-based models. For instance, Lee et al. (2020) trained a biomedical-specific BERT model that improved on many prior state-of-the-art results. HunFlair (Weber et al., 2021) introduced a method that matches the word, contextual, and character embeddings into a unified framework to achieve state-of-the-art performance. In this paper, we evaluate several state-of-the-art systems. Particularly, we focus on systems that use word embeddings, sub-word embeddings, and character embeddings, which allows us to understand the impact of morphological features of the chemical names on gender bias.

Several previous works have measured and highlighted bias in different NLP tasks. For instance, Sap et al. (2019) measures the bias of offensive language detection models on African American English. Likewise, Park et al. (2018) measures gender bias of abusive language detection models and evaluates various methods such as word embedding

[1] The dataset and code will be released publicly upon acceptance.

debiasing and data augmentation to improve biased methods. Davidson et al. (2019) shows racial and ethnic bias when identifying hate speech online and that tweets in the black-aligned corpus are more likely to be assigned hate speech. Gaut et al. (2020) creates the WikiGenderBias dataset to evaluate the gender bias in the relation extraction (RE) model, confirming that the RE system behaves differently when the target entities are of different genders. Cirillo et al. (2020) demonstrate that biases in biomedical applications can stem from various sources, such as skewed diagnoses resulting from clinical depression scales that measure symptoms more prevalent in women, potentially leading to a higher reported incidence of depression among this group (Martin et al., 2013). Other sources include the underrepresentation of minority populations such as pregnant women (Organization and for Women's Health in Society, 2009), non-representative samples in AI training data, and inherent algorithmic discrimination, all potentially contributing to inaccurate and unfair results.

Overall, several metrics have been proposed to measure gender bias. One of the most commonly used metrics involves measuring bias by examining model performance disparities on male and female data points (Kiritchenko and Mohammad, 2018). Performance disparities have been observed across a wide array of NLP tasks such as detecting virus-related text (Lwowski and Rios, 2021), language generation (Sheng et al., 2019), coreference resolution (Zhao et al., 2018), named entity recognition (Mehrabi et al., 2020), and machine translation (Font and Costa-jussà, 2019). Most related to this study, researchers have shown that traditional NER systems (i.e., to detect people, locations, and organizations) are biased concerning gender (Mehrabi et al., 2020). Specifically, Mehrabi et al. (2020) demonstrates that female-related names are more likely to be misidentified as a location than male names. This stream of research underscores the importance of our investigation into performance disparities in NLP.

Finally, while not directly studied in prior NER experiments. It is important to discuss some background about morphological elements of chemical names. Morphological elements often representing masculinity or femininity are frequently used in chemical naming conventions. According to Lieven et al. (2015), consumers perceive linguistically feminine brand names as warmer and likelier. For instance, adding a diminutive suffix to the masculine form of the name usually feminizes it. The masculine names such as Robert, Julius, Antonio, and Carolus (more commonly Charles today) are feminized by adding the suffixes “a”, “ia”, “ina”, or “ine” to generate Julia, Roberta, Antonia, and Caroline, respectively. The suffixes “ia” and “a” is commonly used for inorganic oxides such as magnesia, zirconia, silica, and titania (Hepler-Smith, 2015). Likewise, “ine” is used as the suffix in many organic bases and base substances such as quinine, morphine, guanidine, xanthine, pyrimidine, and pyridine. Hence, while these practices were not originally “biased” in their original usage, they can potentially impact model performance (e.g., feminine names can be detected as chemicals). Therefore, the patterns can cause biased models. As part of our approach to investigate this potential source of bias, we propose using synthetic data to quantify this phenomenon.

	# of Chemical Mentions	# Sentences	# Words
CDR	4,409	14,306	346,001
CHEMDNER	84,355	87,125	2,431,247
AskDoc MALE	1,501	2,862	52,221
AskDoc FEMALE	1,774	2,151	40,184
AskDoc ALL	3,275	5,013	92,405
Synthetic MALE	2,800,000	2,800,000	25,760,000
Synthetic FEMALE	2,800,000	2,800,000	25,760,000
Synthetic ALL	5,600,000	5,600,000	51,520,000

Table 1: Dataset statistics.

3 DATASETS

In this section, we describe the four main datasets used in our experiments: two are publicly-released datasets based on PubMed, and two are newly curated datasets, one using social media data and another based on templates. Table 1 provides their statistics. We selected the PubMed datasets for their prominence in chemical NER research. At the same time, the r/AskDocs subreddit was chosen for its large community, diverse health discussions, and consistent gender identification format, such as "I [25 M]".

CDR (Li et al., 2016) We use the BioCreative V CDR shared task corpus. The CDR corpus comprises 1,500 PubMed articles with 4,409 annotated chemicals, 5,818 diseases, and 3,116 chemical disease interactions. This corpus is designed to address two distinct tasks: Relation classification and NER. For this study, we focus on the NER for chemical entities. The annotator agreement for this corpus was .87. Finally, we used the same train, validation, and test splits from the shared task for our experiments.

CHEMDNER (Krallinger et al., 2015) The CHEMDNER corpus includes abstracts from 10000 chemistry-related journals published in 2013 on PubMed. Each abstract was manually annotated for chemical mentions. These mentions were categorized into seven subtypes: abbreviation, family, formula, identifier, multiple, systematic, and trial. The BioCreative organizers divided the corpus into training (3500 abstracts), development (3500 abstracts), and test (3000 abstracts) sets. The BioCreative IV CHEMDNER corpus comprises 84,355 chemical mention annotations across 10,000 abstracts, with an inter-annotator agreement of .91 (Krallinger et al., 2015). For this study, we only use the major Chemical annotations and ignore the subtypes for consistency across corpora. Finally, we use the same train, validation, and test splits used in the shared task for our experiments.

Synthetic (Template) Data We designed a new synthetic dataset to quantify the gender bias in the Chemical NER models. Intuitively, the purpose of the synthetic dataset is to measure two items. First, do gender-related names and pronouns get incorrectly classified as Chemicals (i.e., cause false positives)? Second, does the appearance of gender-related names/pronouns impact the prediction of other words (i.e., cause false negatives)? Specifically, we create templates such as “[NAME] said they has been taking [CHEMICAL] for illness.”. In the "[NAME]" column, we filled in the names associated with the male and female genders based on the 200 most popular baby names provided by the Social Security Administration [2]. We recognize that gender is not binary and that names do not equal gender. Hence, we refer to these “gender-related” names in this paper. This is a similar framework used by Mishra et al. (2020) and other gender bias papers (Kiritchenko and Mohammad, 2018). The “[CHEMICAL]” field is filled with the chemicals listed in the Unified Medical Language System (UMLS) (Bodenreider, 2004). For example, completed templates include “John said they has been taking citalopram for illness.” and “Karen said they has been taking citalopram for illness.” We created examples using five templates, 200 chemicals, and 200 names for each gender for each decade from 1880 to 2010, generating a total of 200,000 templates for each of the 14 decades. A list of additional templates is shown in Table 2. This dataset is only used for evaluation.

Templates
[NAME] said they has been taking [CHEMICAL] for illness.
Did you hear that [NAME] has been using [CHEMICAL].
[CHEMICAL] has really been harming [NAME], I hope they stop.
I think [NAME] is addicted to [CHEMICAL].
[NAME], please stop taking [CHEMICAL], it is bad for you.

Table 2: Templates used to create the synthetic dataset.

AskDocs We develop a new corpus using data from the Reddit community r/AskDocs. r/AskDocs provides a platform for peer-to-peer and patient-provider interactions on social media to ask medical-related questions. The providers are generally verified medical professionals. We collected all the posts from the community with self-identified gender mentions. To identify self-identified gender, we use a simple regular expression that looks for mentions of “I” or “My” followed by gender, and optionally age, e.g., “I [F34]”, “My (23F)”, “I [M]”. Next, following general annotation recommendations for NLP (Pustejovsky and Stubbs, 2012), the annotation process was completed in two stages to increase the reliability of the labels. First, two graduate students annotated chemicals in the dataset resulting in an inter-annotator agreement of .874, achieving a similar agreement score as CDR and CHEMDNER. Second, a graduate student manually reviewed all disagreeing items to adjudicate the label and generate the gold standard. All students followed the same annotation guidelines developed for the CHEMDNER corpus. Contrary to the synthetic dataset, the actual data will allow users to measure biases arising from text content differences across posts with different self-identified gender mentions.

4 EXPERIMENTAL DESIGN AND METHODS

The goal of NER is to classify words into a sequence of labels. Formally, given an input sequence $\mathcal{X} = [x_1, x_2, \ldots, x_N]$ with N tokens, the goal of NER is to output the corresponding label sequence $\mathcal{Y} = [y_1, y_2, \ldots, y_N]$ with the same length, thus modeling the probabilities over a sequence $p(\mathcal{Y}|\mathcal{X})$. For this task, we conducted an experiment evaluating out-of-domain models on the AskDoc corpus.

[2] https://www.ssa.gov/oact/babynames/

Specifically, models were trained and optimized on the CHEMDNER and CDR datasets and then applied to the AskDoc dataset. All models are evaluated using precision, recall, and F1. To measure bias, we use precision, recall, and F1 differences (Czarnowska et al., 2021). Specifically, let m be Males' performance metric (e.g., F1), and f represent the Female metric. The bias is measured using the difference $f - m$.

4.1 MODELS

We evaluate three distinct models: Word Embedding models (Mikolov et al., 2013), Flair embedding models (Akbik et al., 2018), and BERT-based models (Devlin et al., 2019a). While the embeddings for each model type vary, the sequence processing component is the same for each method. Specifically, following best practices for state-of-the-art NER models (Akbik et al., 2019a), we use a Bidirectional long short-term memory network (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) due to its sequential characteristics and capability to capture long-term dependencies. Recent research has shown that Bi-LSTM models can produce state-of-the-art performance when combined with contextual embeddings and Conditional Random Fields (CRFs) (Mueller et al., 2020; Veyseh et al., 2022). Hence, in this paper, we use the Bi-LSTM+CRF implementation in the Flair NLP framework (Akbik et al., 2019b). The Bi-LSTM+CRF model is flexible because it can accept arbitrary embeddings as input. It is not constrained to traditional word embeddings (e.g., Word2Vec). We describe the embeddings we experiment with in the next Section.

4.2 EMBEDDINGS

We explore three sets of embeddings: Word2Vec, Flair, and BERT. Social media texts are brief and informal. Drugs and chemicals are typically described in descriptive, nontechnical language with spelling errors. These issues challenge social media NER. Some medications, like "all-trans-retinoic acid", contain morphologically difficult parts. Yet, similar-structured phrases still generally represent similar things (Zhang et al., 2021). Hence, how we represent words (i.e., the embeddings we use) can directly impact performance and bias. We describe each embedding we use below:

4.2.1 Word2Vec (Pyysalo et al., 2013)

We use Word2Vec embeddings pre-trained on PubMed and PubMed Central. The embeddings are publicly released as part of the FLAIR package. It is important to state that word embeddings have a major limitation. Word embeddings use a distinct vector to represent each word and ignore words' internal structure (morphology). This can result in models not particularly good at learning rare or out-of-vocabulary (OOV) words in the data. The growing number of emerging chemicals/drugs with diverse morphological forms makes recognizing chemical entities on social media platforms particularly challenging. Another challenge posed by user-generated content is its unique characteristics and use of informal language, typically short context, noisy, sparse, and ambiguous content. Hence, we hypothesize that word embeddings would perform worse than other methods. However, it is unclear how these differences can impact bias.

4.2.2 HunFlair (Weber et al., 2021)

Weber et al. (2021) recently proposed a Flair contextual string embeddings (a character-level language model). Specifically, we use the embeddings in the HunFlair extension of the Flair package (Weber et al., 2021), which is pre-trained on a corpus of three million full-text articles from the Pubmed Central BioC text mining collection (Comeau et al., 2019) and about twenty-five million abstracts from PubMed. Unlike word embeddings mentioned above, Flair embeddings are a contextualized character-level representation. Flair embeddings are obtained from the hidden states of a bi-directional recurrent neural network (BiRNN). They are trained without any explicit notion of a word. Instead, Flair models a word as sequences of characters. Moreover, these embeddings are determined by the text surrounding them, i.e., the same word will have different embeddings depending on its contextual usage. The variant of the Flair embedding used in this study is the Pooled Flair embedding (Weber et al., 2021; Akbik et al., 2018). Furthermore, we use the forward and backward representations of Flair embeddings returned from the BiRNN. Intuitively, character-level embeddings can potentially help improve model predictions with better OOV handling.

4.2.3 BERT (Devlin et al., 2019b)

We also evaluate transformer-based embeddings. Specifically, we use the BERT variant "bert-base-uncased" available Flair and HuggingFace (Wolf et al., 2020). BERT was pre-trained using the BooksCorpus (800M words) and English

	Precision	Recall	F1
CDR + Word	**.8544**	.7989	.8230
CDR + Flair	.8793	.8733	.8761
CDR + BERT	**.8978**	**.9023**	**.9000**
CHEMDNER + Word	.8638	.7916	.8211
CHEMDNER + Flair	**.8929**	**.8652**	**.8783**
CHEMDNER + BERT	.8184	.7363	.7632

Table 3: Overall Results on CDR and CHEMDNER.

Wikipedia (2,500M words) (Devlin et al., 2019b). Furthermore, BERT embeddings are based on subword tokenization, so BERT can potentially handle OOV better than word embeddings alone. Intuitively, it fits somewhere between Flair (generating word embeddings from character representations) and Word2Vec (which independently learns embeddings for each word). Likewise, each word representation is context-dependent. Hence, BERT is better at handling word polysemy by capturing word semantics in context.

4.2.4 Hyper-Parameter Settings

In this section, we report the best hyperparameter for each model. Similar to random hyperparameter search (Bergstra and Bengio, 2012), we generate 100 samples using different parameters for each dataset-model combination (e.g., we generate 100 versions of BERT for the CDR dataset). For the specific hyper-parameters, we used sample dropout from 0.1 to 0.9, hidden layer sizes from {128, 256, 512, 1024}, learning rates selected from 1e-4 to 1e-1 at random, and the option of whether to fine-tune the embedding layers (i.e., True vs. False). In addition, we trained all models for 25 epochs with a mini-batch size set to 32, where only the best model on the validation dataset is saved after each epoch. Finally, all experiments were run on four NVidia GeForce GTX 1080 Ti GPUs.

5 RESULTS

In this section, we report the performance of our model on the original CDR and CHEMDNER test datasets and the synthetic and real-data bias results.

5.1 CDR and CHEMDNER Results

Table 3 reports the average recall, precision, and F1 scores for each embedding type for the CDR and CHEMDNER datasets. The scores are averaged over the various random seeds and hyperparameters used to train the models. The Flair embeddings result in the best performance for the CDR dataset. While in the CHEMDNER corpus, the Flair outperforms the BERT embeddings (.8783 vs. .7632). For the CHEMDNER results, we found that BERT is highly sensitive to hyperparameters, resulting in poorly performing models. The best-performing BERT models can perform similarly to the Flair model. See the supplementary material for details (e.g., max, min, and median scores).

5.2 Synthetic (Template) Results

In Table 4, we report the average synthetic dataset results and bias scores for each model trained on three different datasets (CDR, CHEMDNER, and AskDocs) with the three different embeddings (Word, Flair, and BERT). Overall, NER models have a substantial bias against female-related names. Specifically, nine out of nine models (1.000) have a lower precision for female-related templates, with an average precision bias of .0204 against female-related names. Likewise, seven out of nine (.7778) dataset-model pairs have lower F1 scores for female-related templates. The recall scores are similar for male- and female-related templates, with an average score near 0. The AskDoc dataset has the largest bias scores against female-related names (e.g., .0555 for precision). Yet, the CDR and CHEMDNER datasets also have substantial biases with differences as high as .0367. These results indicate that most bias differences are caused by female-related names being more likely to be classified as a chemical. This finding is consistent with prior research on naming conventions for brands (Lieven et al., 2015). To further investigate this, we randomly sample 100 chemicals from all three datasets and measured the number of brand name mentions. Overall, we found one brand name in the CHEMDNER dataset, 19 in the CDR dataset, and 32 in the ASKDOC dataset, which generally matches the bias performance differences in Table 4. Moreover, the Word Embedding (Word2Vec) models have the lowest bias scores. Word2Vec models are not impacted by the morphological structure of the chemical names. Hence, the models using word embeddings do not confuse names for chemicals. We find similar patterns for word embeddings on models trained on each dataset.

5.3 AskDoc Results

The AskDoc results are reported in Table 5. The results in Table 5 come from a model trained on PubMed data. As seen in Table 5, there is no sig-

	Male			Female			Difference		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
CDR + Word	1	.8230	.9029	1	.8230	.9029	.0000	.0000	.0000
CDR + Flair	.9711	.9486	.9597	.9344	.9494	.9418	.0367	-.0008	.0179
CDR + BERT	.9867	.8493	.9128	.9728	.8444	.9041	.0138	.0048	.0087
CHEMDNER + Word	.9990	.8625	.9257	.9968	.8622	.9246	.0021	.0003	.0011
CHEMDNER + Flair	.9982	.8836	.9374	.9885	.8852	.9340	.0097	-0.007	.0034
CHEMDNER + BERT	.9913	.8768	.9306	.9680	.8762	.9198	.0233	-.0006	.0107
ASKDOC + Word	.9739	.9330	.9530	.9739	.9330	.9530	.0000	.0000	.0000
ASKDOC + Flair	.8833	.9523	.9164	.8278	.9519	.8852	**.0555**	.0005	**.0312**
ASKDOC + BERT	.9394	.9288	.9340	.8967	.9282	.9121	.0427	.0006	.0220
				Aggregate Measures					
AVG							.0204	-.0002	.0106

Table 4: Synthetic (Template) Data Results. The smallest bias score for each dataset is marked in blue and the biggest is marked in red. The overall largest scores are in **bold**. The bottom section reports aggregate result measures, specifically the average differences and the percent of the DATASET + MODEL combinations that are biased against the female-related text.

	Male			Female			Difference		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
CDR + Word	.8375	.5605	.6548	.8400	.5495	.6499	-.0025	.0110***	.0049
CDR + Flair	.8320	.6557	.7285	.8293	.6256	.7081	.0026	.0302***	.0204***
CDR + BERT	.8724	.6582	.7500	.8693	.6215	.7244	.0030	**.0367*****	.0256***
CHEMDNER + Word	.8693	.5444	.6609	.8751	.5305	.6521	-.0058	.0139***	.0089***
CHEMDNER + Flair	.8791	.6120	.7206	.8611	.5830	.6939	**.0180****	.0290**	.0267***
CHEMDNER + BERT	.7995	.5942	.6717	.7964	.5648	.6438	.0031	.0295**	**.0279*****
				Aggregate Measures					
AVG							-.0056	.0251	.0189

Table 5: AskDoc Results. The smallest bias score for each dataset is marked in blue and the biggest is marked in red. The overall largest scores are in **bold**. The bottom section reports aggregate result measures, specifically the average differences and the percent of the DATASET + MODEL combinations that are biased against the female-related text. Statistically, significant differences based on the Wilcoxon Signed Rank test are marked with * (p-value $<$ .05), ** (p-value $<$ .01), and *** (p-value $<$ .001).

nificant difference in precision between male and female datasets for most models, suggesting that precision remains consistent regardless of gender. However, recall displays a consistent bias against the female group. Likewise, F1 scores indicate a bias against the female group, except for the Word Embedding model trained on the CDR corpus. In contrast, Table 4 mirrors the results from Table 5, but the biases observed are in precision and F1 scores rather than recall and F1 scores. Overall, we have several major findings. First, again, we find substantial female-related bias in the Chemcial NER system. Here, the bias is based on self-identified posts, not names. For instance, the CDR+BERT model has a recall for the Female posts nearly 4% lower (i.e., .0367) than the Male posts. However, what does this mean in real-world terms? Considering a sample of 1,000,000 chemical mentions across male and female posts (a relatively small number in social media), a 4% recall difference results in an additional 40,000 false negatives for the female group. For example, there are well-known health disparities between men and women for depression, with absolute differences of less than 3% (Salk et al., 2017). Hence, a 4% recall difference can substantially impact findings if applied researchers or practitioners use out-of-domain models to understand medications for this disease. Such a considerable gap can markedly affect the utility and trustworthiness of these predictive outcomes in practical scenarios.

In summary, these results underline the necessity to acknowledge potential gender bias in information extraction tasks within biomedical applications. The performance disparity across genders calls for applying bias mitigation techniques to ensure equitable system performance. Further, the influences of the chosen NLP model and training corpus on this bias underline the importance of careful model selection and data curation in creating unbiased NLP systems.

Second, the models with the most bias on the synthetic data do not correlate with the findings on real data. For example, the models trained on CDR corpus have the largest bias on the synthetic data but have the smallest bias when evaluating real data. Likewise, the synthetic data suffered from precision bias, while much of the bias on real data is related to recall. These findings are important given the reliance on synthetic data in bias analysis papers. In comparison, synthetic data allows us to target specific types of biases, **synthetic data alone does not provide an accurate estimate of bias in practice**.

Third, a pivotal question often raised is, "Does increasing model accuracy inherently lead to decreased bias?" From our observations on the AskDoc and synthetic datasets, there is no direct correlation. Intriguingly, Word Embedding-based models, which were the least accurate among the models tested, exhibited the least bias. This underscores the idea that accuracy and fairness are both essential axes of model evaluation. Relying solely on improving accuracy will not automatically address bias or fairness concerns. Hence, the results suggest that performance disparities need to be directly addressed; simply developing more "accurate" models will not suffice. Finally, note that the biases can range higher than .0367. See the supplementary material for median, max, and min scores. Furthermore, it's important to mention that these biases can exceed a measure of .0367. For more detailed results including median, maximum, and minimum scores, refer to the supplementary material.

6 Conclusion

In this paper, we evaluate the gender bias of Chemical NER systems. Moreover, we compare bias measurements from synthetic data with real-world self-identified data. We make two major findings. First, Chemical NER systems are biased with regard to gender for synthetic data. Specifically, our study found that **female name-like patterns feature prominently in chemical naming conventions**. This characteristic leads to a notable bias in NER systems, where female-related names are disproportionately identified as chemicals, inadvertently escalating the gender bias in these systems.

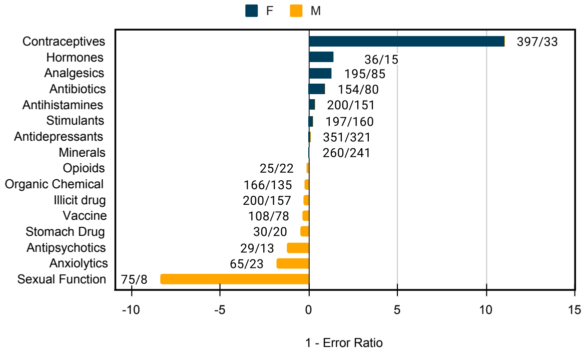

Figure 1: Ratio of false negatives for various drug categories. The ratio is represented next to each bar. For female-leaning errors, the female false negative count (FN_f^k) is in the numerator. For male-leaning errors, the male false negative count (FN_m^k) is in the numerator.

Furthermore, we explored the performance of these models in real-world scenarios and found that most models perform better on male-related data than female-related data. **A striking revelation was the system's poor performance when identifying chemicals frequently found in female-related data, such as mentions of contraceptives.** This result further compounds the concern of bias, bringing attention to the potential real-world implications of such inaccuracies.

Additionally, our analysis exposed the limitations of synthetic data in estimating gender bias. **While synthetic data serves as a useful tool for identifying specific types of biases, it fails to provide a comprehensive reflection of the bias in real-world applications.** This discovery underscores the need for real-world bias analyses alongside synthetic data investigations.

Our study also drew attention to the non-correlation between model accuracy and bias. We discovered that **the least accurate models, based on Word Embeddings, exhibited the least bias.** This finding reiterates that enhancing model accuracy alone will not suffice in addressing these biases; instead, it is necessary to explicitly tackle the disparities in performance.

In conclusion, the results of our study emphasise the urgent need for deliberate bias mitigation strategies in Chemical NER systems. Our findings spotlight the necessity for incorporating both synthetic and real-world data considerations to develop models that are both fair and reliable.

References

Pankaj Agarwal and David B. Searls. 2008. Literature mining in support of drug discovery. *Briefings in bioinformatics*, 9 6:479–92.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019a. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019b. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Elisa Chilet-Rosell. 2014. Gender bias in clinical research, pharmaceutical marketing, and the prescription of drugs. *Global Health Action*, 7(1):25484.

Yu-Wen Chiu, Wen-Chao Yeh, Sheng-Jie Lin, and Yung-Chun Chang. 2021. Recognizing chemical entity in biomedical literature using a bert-based ensemble learning methods for the biocreative 2021 nlm-chem track. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, et al. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1):81.

Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. 2019. Pmc text mining subset in bioc: about three million full-text articles and growing. *Bioinformatics*, 35(18):3533–3535.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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)*, pages 4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Lizzy Farrugia and Charlie Abela. 2020. Mining drug-drug interactions for healthcare professionals. *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*.

Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, et al. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Evan Hepler-Smith. 2015. “just as the structural formula does”: Names, diagrams, and the structure of organic chemistry at the 1892 geneva nomenclature congress. *Ambix*, 62(1):1–28.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Theo Lieven, Bianca Grohmann, Andreas Herrmann, Jan R Landwehr, and Miriam Van Tilburg. 2015. The effect of brand design on brand gender perceptions and brand preference. *European Journal of Marketing*.

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2021. Combining word embeddings to extract chemical and drug entities in biomedical literature. *BMC bioinformatics*, 22(1):1–18.

Brandon Lwowski and Anthony Rios. 2021. The risk of racial bias while tracking influenza-related content on social media using machine learning. *Journal of the American Medical Informatics Association*, 28(4):839–849.

Maria Mammì, Rita Citraro, Giovanni Torcasio, Gennaro Cusato, Caterina Palleria, and Eugenio Donato di Paola. 2013. Pharmacovigilance in pharmaceutical companies: An overview. *Journal of Pharmacology & Pharmacotherapeutics*, 4:S33 – S37.

Lisa A Martin, Harold W Neighbors, and Derek M Griffith. 2013. The experience of symptoms of depression in men vs women: analysis of the national comorbidity survey replication. *JAMA psychiatry*, 70(10):1100–1106.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and A. G. Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104.

World Health Organization and Key Centre for Women’s Health in Society. 2009. Mental health aspects of women’s reproductive health: a global review of the literature.

Karen O’Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O’Reilly Media, Inc.".

S Pyysalo, F Ginter, H Moen, T Salakoski, and S Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44.

Joseph L Riley III, Michael E Robinson, Emily A Wise, Cynthia D Myers, and Roger B Fillingim. 1998. Sex differences in the perception of noxious experimental stimuli: a meta-analysis. *Pain*, 74(2-3):181–187.

Anthony Rios. 2020. Fuzze: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 881–889.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Rachel H Salk, Janet S Hyde, and Lyn Y Abramson. 2017. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological bulletin*, 143(8):783.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1668–1678.

Mirsada Serdarevic, Catherine W Striley, and Linda B Cottler. 2017. Gender differences in prescription opioid use. *Current opinion in psychiatry*, 30(4):238.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412.

David RM Smith, F Christiaan K Dolk, Timo Smieszek, Julie V Robotham, and Koen B Pouwels. 2018. Understanding the gender gap in antibiotic prescribing: a cross-sectional analysis of english primary care. *BMJ open*, 8(2):e020203.

Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2021. Deep learning with language models improves named entity recognition for pharmaconer. *BMC bioinformatics*, 22(1):1–16.

Lena Thunander Sundbom, Kerstin Bingefors, Kerstin Hedborg, and Dag Isacson. 2017. Are men undertreated and women over-treated with antidepressants? findings from a cross-sectional survey in sweden. *BJPsych bulletin*, 41(3):145–150.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Bonan Min, and Thien Huu Nguyen. 2022. Generating complement data for aspect term extraction with gpt-2. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 203–213.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongxuan Zhang, Hongfei Lin, Yuqi Ren, Zhihao Yang, Jian Wang, Xiaodong Duan, and Bo Xu. 2021. Identifying adverse drug reaction entities from social media with adversarial transfer learning model. *Neurocomputing*, 453:254–262.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

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.

A Appendix

A.1 Error Analysis and Discussion

Our experiments show that Chemical NER systems are biased. However, what specifically is causing the errors? For the synthetic data, the answer is gender-related names. To understand the errors in the AskDoc data, we analyzed the errors made by the best NER models trained on the out-of-domain corpus (CHEMDNER and CDR) and tested the male and female splits of the AskDocs corpus. In Figure 1, we report the ratio of false negatives for different categories of drugs/chemicals. Specifically, for every false negative made by the top models of each dataset-model combination, we manually categorized them into a general chemical class (e.g., Contraceptives, Analgesics/Pain Killers, and Stimulants). Formally, let FN_m^k represent the total number of false negatives for chemical types k and male data m. Let FN_f^k represent the female false negatives. If FN_m^k is larger than FN_f^k, we define the ratio as $-(1 - FN_m^k/FN_f^k)$. Likewise, if FN_f^k is greater than FN_m, then we define the ratio as $1 - (FN_f^k/FN_m^k)$. Hence, when male ratios are higher, the score is negative; otherwise, it is positive.

Overall, we make several important findings. First, we find that the models make slightly more false negatives on the chemicals categories Contraceptives (e.g., birth control and Plan B One-Step), Hormones (e.g., Megace used to treat the symptoms of loss of appetite and wasting syndrome in people with illnesses such as breast cancer), Analgesics (i.e., Pain Killers such as Tylenol) and Antibiotics on the female dataset. In contrast, the models make slightly more errors in the chemical categories Anxiolytics (e.g., drugs used to treat anxiety), Antipsychotics (e.g., chemicals used to manage psychosis, principally in schizophrenia), and sexual function drugs (e.g., Viagra). Furthermore, while the ratio for the most male- and female-related errors (Contraceptives and Sexual Function) are similar, the absolute magnitudes are substantially different. For instance, there are 397 Contraceptive FNs in the female dataset, but only 75 Sexual Function FNs appear in the male dataset. This provides an explanation for the large differences in recall on the AskDoc corpus between the male and female datasets.

Interestingly, we find that the prevalence of chemicals across gender-related posts matches the prevalence found in traditional biomedical studies. Previous research report that women have been prescribed analgesics (e.g., pain killers such as opioids) twice as often as men (Chilet-Rosell, 2014; Serdarevic et al., 2017). While there is still limited understanding about whether men are under-prescribed or women are over-prescribed, the disparities in prescriptions are evident. Thus, the finding in Figure 1 that we receive twice as many analgesics FNs for female data is important. Depending on the downstream application of the Chemical NER system, these performance disparities may potentially increase harm to women. For example, if more varieties of drugs are prescribed to women, but our system does not detect them, then an ADR detection system will not be able to detect important harms.

We also find differences in Antibiotic FNs in Figure 1. There have also been medical studies showing gender differences in Antibiotic prescriptions. For example, a recent meta-analysis of primary care found that women received more antibiotics than men, especially women aged 16–54, receiving 36%–40% more than males of the same age (Smith et al., 2018). Again, if we do not detect many of the antibiotics prescribed to women, this can cause potential health disparities in downstream ADR (and other) systems.

Next, in Table 6, we report the false negative rate (FNR) for each category along with the general frequency of each category. Using the Pearson correlation coefficient, we relate the frequency of each category with the false negative rate for the male and female groups, respectively. Intuitively, we would expect the false negative rate to go down as the frequency increases, which matches our findings. However, we find that the correlation is much stronger for the male group than the female group.

In Table 7, we report the FNR for the female and male groups, respectively. We also introduce a new metric, weighted FNR, which assigns importance scores for each of the FNRs shown in to create a macro-averaged metric. Intuitively, the distribution of categories is different for both the male and female groups. So, we want to test whether the FNR scores are distributed uniformly across all categories, irrespective of, or if the errors are more concentrated for gender-specific categories. More errors in gender-specific categories can adversely impact a group that is not captured with the global FNR metric. Formally, we define wFNR for the female group as

$$wFNR^f = \sum_i^N w_i^f FNR_i^f \qquad 1093$$

where FNR_i^f represents the female false negative rate for category i. Likewise, w_i^f is defined as

$$w_i^f = \frac{1}{\sum_i w_i^f} \cdot \frac{N_i^f/N^f}{N_i^m/N^m} \qquad 1096$$

where N_i^f and N_f^m represent the total number of times a category i appears for the female and male groups, respectively. Intuitively, we are diving the ratio of each category for female and male groups. So, if a category appears more often for females than males, proportionally, then the score will be

	Total Male	**FNR Male**	**Total Female**	**FNR Female**
Contraceptives	33	1.0000	408	.9730
Hormones	170	.0882	230	.1565
Analgesics	571	.1489	952	.2048
Antibiotics	326	.2454	347	.4438
Antihistamines	270	.5593	295	.6780
Stimulants	522	.3065	390	.5051
Antidepressants	781	.4110	1043	.3365
Minerals	605	.3983	785	.3312
Opioids	43	.5814	95	.2316
Organic Chemical	441	.3764	346	.3902
Illicit drug	353	.5666	311	.5048
Vaccine	108	1.0000	78	1.0000
Stomach Drug	55	.5455	44	.4545
Antipsychotics	47	.6170	95	.1368
Anxiolytics	126	.5603	100	.2300
Sexual Function Drug	78	.9615	8	1.0000
PCC between Total and FNR	**-.58**		-.26	

Table 6: False negatives rate (FNR) for female and male-related AskDoc datasets. The pearson correlation coefficient (PCC) between the frequency of each chemical type and the FNR for teach group is marked in the last row.

	FNR	**wFNR**
Male	.3948	.6875
Female	**.4064**	**.8088**
Gap	.0116	.1213
Ratio	1.0294	1.1764

Table 7: FNR and weighted FNR (wFNR) results.

higher. We normalize these scores for each group so they sum to one. Overall, we find an absolute gap of more than 1% (3% relative difference) between the FNR for male and female groups. But, even worse, there is a much larger gap (.1213 vs .0116) when using wFNR. This result suggests that many of the false negatives are concentrated for gender-specific categories (e.g., contraceptives) for the female group more than the male group.

A.2 Limitation

There were several limitations to our study. First, the adjudication of disagreeing items was dependent on the judgment of a single graduate student, potentially introducing human error and bias compared to a multi-adjudicator approach. Second, the vast volume of data from the active r/AskDoc subreddit community makes the feasibility of one person's comprehensive review debatable. Although our annotation method is in line with standard practices, a more multi-faceted approach involving numerous annotators and adjudicators might offer improved accuracy and consistency in future datasets.