In the Name of Fairness: Assessing the Bias in Clinical Record De-identification

Yuxin Xiao*
yuxin102@mit.edu
Massachusetts Institute of Technology
USA

Shulammite Lim*
shulim@mit.edu
Massachusetts Institute of Technology
USA

Tom Joseph Pollard
tpollard@mit.edu
Massachusetts Institute of Technology
USA

Marzyeh Ghassemi
mghassem@mit.edu
Massachusetts Institute of Technology
USA

ABSTRACT

Data sharing is crucial for open science and reproducible research, but the legal sharing of clinical data requires the removal of protected health information from electronic health records. This process, known as de-identification, is often achieved through the use of machine learning algorithms by many commercial and open-source systems. While these systems have shown compelling results on average, the variation in their performance across different demographic groups has not been thoroughly examined. In this work, we investigate the bias of de-identification systems on names in clinical notes via a large-scale empirical analysis. To achieve this, we create 16 name sets that vary along four demographic dimensions: gender, race, name popularity, and the decade of popularity. We insert these names into 100 manually curated clinical templates and evaluate the performance of nine public and private de-identification methods. Our findings reveal that there are statistically significant performance gaps along a majority of the demographic dimensions in most methods. We further illustrate that de-identification quality is affected by polysemy in names, gender context, and clinical note characteristics. To mitigate the identified gaps, we propose a simple and method-agnostic solution by fine-tuning de-identification methods with clinical context and diverse names. Overall, it is imperative to address the bias in existing methods immediately so that downstream stakeholders can build high-quality systems to serve all demographic parties fairly.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Fairness**; • **Social and professional topics** → **Patient privacy**.

KEYWORDS

Fairness, Named Entity Recognition, Clinical De-identification

*Equal contribution.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3593982

ACM Reference Format:
Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3593013.3593982

1 INTRODUCTION

The increased availability of clinical datasets [49, 71, 72] plays a significant role in the recent advancements in machine learning (ML)-aided healthcare systems [13, 37, 104, 115]. In order to share clinical trial data legally, stakeholders must adhere to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provisions by masking 18 types of protected health information (PHI). If done appropriately, clinical data sharing adds significant value to scientific reproducibility [90] at low risk to patient privacy [77, 112]. In this regard, various open-source software [73, 93] and commercial companies provide services to de-identify electronic health records (EHRs). Named entity recognition (NER) tools [78, 117, 134] in natural language processing (NLP) libraries [15, 64, 86, 105] are commonly used in this space.

Despite the compelling performance of many ML-empowered healthcare systems [123], models have been shown to underperform in minorities and minoritized populations, and naive applications can extend and increase existing biases [37, 55, 55, 56, 113, 129, 132]. Disparities in performance between demographic groups can lead to real harm [59, 88, 137]. For instance, a state-of-the-art early warning model for acute kidney injury [122] failed to extend to female patients due to its male-dominated training data [34]. In de-identification specifically, failing to remove the PHI of certain demographic groups would violate the Safe Harbor regulations. This failure could exacerbate the known misuse of data from minorities [30, 47, 54] and expose these groups to targeted attacks such as identity theft [17, 18].

In this paper, we audit the performance of de-identification methods on a specific PHI category—names—from clinical notes. We focus on names because they are correlated with demographic features and are disproportionately identifiable amongst the defined PHI categories. To date, existing studies [87, 91, 94] have compared a limited number of baselines on short sentence templates that are much simpler than the real clinical notes. In contrast, we conduct a large-scale empirical evaluation of nine commercial and open-source de-identification methods based on 16 name sets that vary

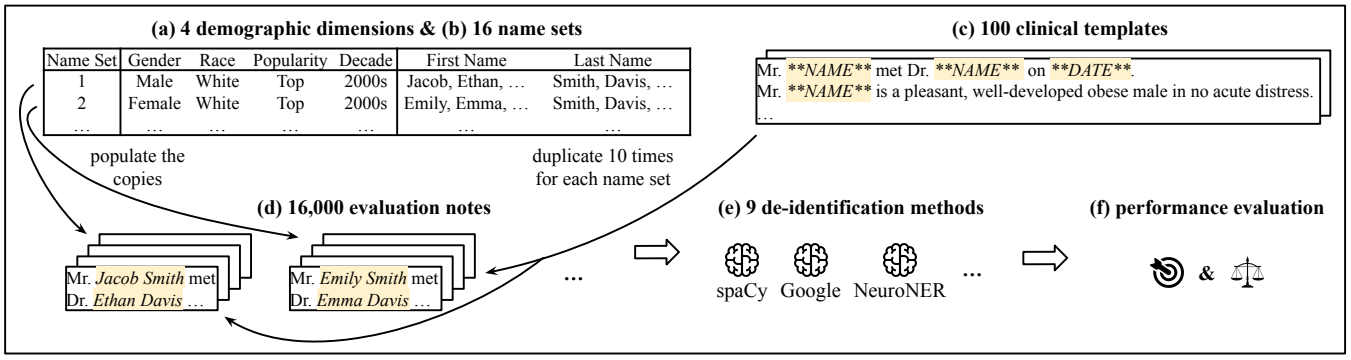

Figure 1: Workflow of our empirical study. We identify (a) four demographic dimensions and prepare (b) 16 name sets with diverse settings. For each name set, we duplicate each of the (c) 100 clinical templates ten times and populate the copies with randomly generated names. We then use these (d) 16,000 evaluation notes to assess (e) nine de-identification methods.

along four demographic dimensions—gender, race, name popularity, and the decade of popularity—and 100 note templates [80] curated from real-world clinical records. We adopt the gender and racial categories in the U.S. Social Security [6] and Census [3] datasets and calculate popularity from name frequency over three selected decades. While we acknowledge the inherent limitation of using standardized racial categorization and binary gender groups, our work is a first step toward the evaluation of de-identifying names in EHRs, capturing the real harm that gaps could incur.

First, we investigate whether demographic bias exists in clinical de-identification methods. While some methods attain an overall competitive recall, a majority of the examined methods exhibit statistically significant performance gaps along most demographic dimensions. For instance, we note that these methods are, on average, significantly better at recognizing "rare" names in White people than "popular" names in Asian people.

Second, we assess potential factors contributing to the observed underperformance. We find that names with polysemy—other meanings in English—are disproportionately unrecognized, regardless of the associated races. Most methods suffer when the gender inferred from a name disagrees with the gender suggested by the semantic context. Certain note characteristics, such as length and the number of unique names included, also reduce performance.

Third, we perform fine-tuning on two of the open-source de-identification methods (`spaCy` [64] and `NeuroNER` [43, 44]) with clinical context and diverse names. We find that this significantly improves the methods' overall performance and reduces their demographic bias, especially along the dimensions of race and popularity. We advise that this simple, method-agnostic solution should be a minimal first step for practitioners in the de-identification space.

We contribute a comprehensive analysis of the bias in de-identifying names from clinical notes, with insights into the existence of the bias and the cause of the underperformance, and provide a simple mitigation option. We emphasize that asymmetric de-identification by existing methods could violate legal regulations and is a serious socio-technical ethical issue. We encourage future work to build upon our results, balancing both de-identification performance and demographic fairness.

2 RELATED WORK

De-identification. The HIPAA Safe Harbor regulations require clinical trial data to be properly anonymized before being shared for various purposes [83, 125]. Toward this goal, the de-identification of EHRs has drawn long-lasting attention from both clinical practitioners and the NLP community [73, 93]. Traditional de-identification methods use rule-based pattern matching [20, 50, 97, 121] or ML algorithms [10, 44, 128, 135] for sequence tagging and attain competitive results in the i2b2 (Informatics for Integrating Biology and the Bedside) de-identification challenges [119, 127]. Several companies, like Google and Amazon, also provide commercial services to detect and obscure PHI data in plain text. Along a related line of research, many NLP systems [15, 64, 86, 105] can fulfill a similar goal by treating de-identification as an NER problem [78, 117, 134].

Bias in NLP Systems. Existing work reports the prevalence of systematic bias in NLP frameworks [26, 114]. Unfairness in text representations [28, 33, 75, 100, 138] or language models [95, 96] can be escalated in downstream applications such as sentiment analysis [24, 74], machine translation [110, 118], and coreference resolution [107, 139]. Gender [36, 89, 120] and racial [27, 41] bias in NLP systems may bring about catastrophic social consequences [68, 109]. In response, researchers have proposed metrics [29, 40, 70] and methods [65, 103, 116] to mitigate bias in NLP models.

Bias in Healthcare and Other High-Stakes Applications. Demographic bias exists in healthcare systems [129, 132], typically in an implicit and unconscious way [59, 88, 137]. For instance, when medical assistance leverages biased artificial intelligence [67, 92], the unfairness is usually carried forward to subsequent healthcare practice [53, 57]. Hence, addressing the bias here demands joint efforts from both ML researchers [23, 38] and healthcare professionals [32, 99]. Bias could also occur in other high-stakes domains such as job applications [22, 42, 60] and law enforcement [31, 45, 46]. We leave a detailed discussion of the bias in those areas to future work.

Bias in Clinical De-identification. In light of the discussion above, it is crucial to carefully examine the bias in de-identification methods, given the pivotal role of de-identification in healthcare pipelines. Previous work [87, 91, 94] has only compared a small set of baselines based on template sentences that are much simpler than realistic

Name Set	Gender	Race	Popularity	Decade	First Name Examples	Last Name Examples
1	Male	White	Top	2000s	Jacob, Ethan, Tyler, …	Smith, Davis, Brown, …
2	Female	White	Top	2000s	Emily, Emma, Olivia, …	Smith, Davis, Brown, …
3	Male	White	Medium	2000s	Wade, Ted, Brien, …	Waldon, Clapp, Bogle, …
4	Female	White	Medium	2000s	Mabel, Liz, Terressa, …	Waldon, Clapp, Bogle, …
5	Male	White	Bottom	2000s	Nicki, Leslee, Marti, …	Lofft, Lyna, Tamaro, …
6	Female	White	Bottom	2000s	Glenn, Lyle, Heath, …	Lofft, Lyna, Tamaro, …
7	Male	Black	Medium	2000s	Cedric, Marlon, Ollie, …	Booker, Grier, Spikes, …
8	Female	Black	Medium	2000s	Aisha, Ebony, Jamila, …	Booker, Grier, Spikes, …
9	Male	Asian	Medium	2000s	Zhi, Nguyen, Rajeev, …	Ngo, Mao, Ahmed, …
10	Female	Asian	Medium	2000s	Neha, Priya, Xin, …	Ngo, Mao, Ahmed, …
11	Male	Hispanic	Medium	2000s	Leonel, Camilo, Cruz, …	Ceja, Amaro, Recinos, …
12	Female	Hispanic	Medium	2000s	Celina, Rebeca, Luisa, …	Ceja, Amaro, Recinos, …
13	Male	White	Top	1970s	Patrick, Brian, Eric, …	Smith, Davis, Brown, …
14	Female	White	Top	1970s	Amy, Lisa, Laura, …	Smith, Davis, Brown, …
15	Male	White	Top	1940s	Jerry, George, Frank, …	Smith, Davis, Brown, …
16	Female	White	Top	1940s	Linda, Carol, Nancy, …	Smith, Davis, Brown, …

Table 1: 16 name sets of diverse demographic backgrounds and examples of first and last names for each set. Name Sets $1 \sim 6$ **are names with top, medium, and bottom popularity in the 2000s that are also exclusive to the White racial group. Name Sets** $7 \sim 12$ **are names with medium popularity in the 2000s that are also exclusive to the Black, Asian, and Hispanic racial groups. Name Sets** $13 \sim 16$ **are names with top popularity in the 1970s and 1940s that are also exclusive to the White racial group.**

clinical de-identification challenges. There lacks a holistic analysis that explores the bias in de-identification methods of different categories, the factors leading to the methods' underperformance, and the solution to alleviate the bias. Therefore, our paper aspires to fill this gap via extensive empirical studies based on 16 name sets with diverse demographic backgrounds, 100 real-world clinical note templates, and nine public and private de-identification methods.

3 EXPERIMENT SETUP

In this paper, we focus on assessing the bias in de-identifying a specific type of PHI data—people's names—from clinical records. We choose names amongst the defined PHI types because they are commonly associated with specific demographic features and are particularly identifiable.

As illustrated in Figure 1, we first identify (a) four demographic dimensions (i.e., gender, race, name popularity, and the decade of popularity) and prepare (b) 16 name sets with diverse demographic settings in Table 1. Each name set consists of 20 first and 20 last names, which can be paired to produce 400 full names in total. We then curate (c) 100 clinical templates from hospital discharge records [80]. For each name set, we duplicate each of the 100 templates ten times and fill in full names randomly generated from that name set. This creates a total of (d) 16,000 notes with 116,160 name mentions for evaluation. We use these notes to conduct a large-scale empirical analysis of (e) nine de-identification baseline methods to inspect the bias along the four demographic dimensions.[1]

3.1 Definition of Demographic Dimensions

To measure the demographic information associated with a name, we define the following four demographic dimensions.

- The **gender** of a name refers to the sex assigned at birth to someone with that name, because the phonological property of a name suggests the associated gender [35]. We examine two groups for gender: male and female.
- The **race** of a name refers to the expected racial or ethnic identity of someone with that name, reflecting the variation in prevalence that exists between different self-reported racial or ethnic groups [62]. We consider four racial or ethnic groups: White, Black, Asian, and Hispanic. Other groups are skipped due to prohibitively small community sizes.
- The **popularity** of a name refers to the size of the population of a gender within a decade having that name. We compare three groups here: top, medium, and bottom popularity.
- The **decade** of popularity refers to the decade in which a name is popular in the U.S. in terms of babies being given the name, as name trends change over time [58]. We assess three decade groups: 2000s, 1970s, and 1940s.

Limitations of Standardized Demographic Categories. We acknowledge the limitation of using standardized self-reported racial categorization and binary gender groups when composing the name sets. More fine-grained racial categorizations are possible in future work, and there could be variety in the linguistic norms and naming traditions even within each racial group we consider. Transgender and non-binary gender groups are also important to consider in future work, as these groups may use gender-neutral names or have variations in name usage between records.

We use standardized self-reported racial categorization and binary gender groups because it is important to evaluate the performance of de-identification methods on data that is routinely collected in EHRs [21]. We emphasize that we do not perform any demographic inference as part of a classification system or training set in this work. We do not believe that these categories should be viewed as scientific truth and recognize the larger critical interrogation surrounding whether gender and ethnicity should be discerned from names in such systems [84]. Instead, we use these categories

[1] Our code is available at https://github.com/xiaoyuxin1002/bias_in_deid.

in the spirit in which they were created by the U.S. Office of Management and Budget to "monitor and redress social inequality" [25]. The examination of the impact of more fluid categorizations of gender, race, and religion is important for future work in this space.

3.2 Construction of Name Sets

In this study, we compute the popularity of first names for each gender based on the U.S. Social Security dataset [6] across the entire population, rather than for each racial group. We then select names that are primarily associated with a self-identified racial group with a margin over 10% based on the mortgage application dataset in [126]. We note that this is different from picking the most popular names for each racial group independently.

In the U.S. setting, all top popularity names, as evaluated by absolute frequency ranking, are identified with the White racial group. For this reason, we consider names associated with the Black, Asian, or Hispanic groups that are of medium popularity. First names of medium popularity for each race and gender (i.e., Name Sets 3, 4, 7, 8, 9, 10, 11, and 12) are randomly sampled from those with a frequency ranking between 400 and 8,000 in the entire population in the 2000s. First names of bottom popularity for the White group (i.e., Name Sets 5 and 6) are randomly sampled from those occurring exactly five times in the 2000s. We set each name set to 20 names since based on the procedure described above, there are only 20 names that are of medium popularity in the 2000s and primarily used by Black males. We also ensure that first names of top popularity within each gender and decade are mutually exclusive (i.e., no shared first names in Name Sets 1, 2, 13, 14, 15, and 16).

We prepare last names in a similar fashion based on the 2000 Census dataset alone [3], because we assume that the last name popularity is relatively fixed. Specifically, this means that the most popular last names for the White racial group in the 1970s and 1940s are assigned to be the same as those in the 2000s.

Limitations of the Datasets. We acknowledge that our datasets are limited to the U.S., and therefore, our findings need to be reproduced in other contexts with distinct name distributions. Furthermore, our use of the mortgage application dataset for self-reported racial matching is limited to those who have the financial security to apply for a loan. As we do not have access to other sources of names and self-reported races, we use the available data to demonstrate that—even in this presumably more privileged subset of the population—there are de-identification gaps.

3.3 Group Pooling for Demographic Performance Comparisons

To evaluate model performance along each demographic dimension, we design experiments that control for other dimensions as follows.

- We assess the impact of **gender** by pooling and comparing the results of male (i.e., 1, 3, 5, 7, 9, 11, 13, and 15) and female Name Sets (i.e., 2, 4, 6, 8, 10, 12, and 16). Race, popularity, and decade of popularity all vary within these two groups.
- We compare performance along **race** by pooling Name Sets 3 and 4 for the White group, Name Sets 7 and 8 for the Black group, Name Sets 9 and 10 for the Asian group, and Name Sets 11 and 12 for the Hispanic group. These are the male and female names of medium popularity in the 2000s across the four racial groups.
- We examine the influence of **popularity** by forming and comparing names with varying levels of popularity within the White group, where top popularity is based on Name Sets 1 and 2, medium popularity is based on Name Sets 3 and 4, and bottom popularity is based on Name Sets 5 and 6.
- We evaluate the difference in performance among the three **decade** groups by comparing the male and female names of top popularity for the White group in each decade: Name Sets 1 and 2 for the 2000s, Name Sets 13 and 14 for the 1970s, and Name Sets 15 and 16 for the 1940s.

3.4 Preparation of Clinical Templates

We manually curate 100 clinical note templates based on hospital discharge records from Beth Israel Lahey Health between 2017 and 2019. We follow the HIPAA Safe Harbor provisions by marking the occurrence of names in the templates and replacing other PHI classes with realistic, synthetic values. We note that our templates [80] are more complex than those used in existing benchmark datasets [87, 91, 94], with an average of 12,893 characters and 3.5 unique names per template and each unique name appearing an average of 2.1 times per template. This design is more analogous to real-world de-identification challenges and more likely to expose flaws in less effective methods.

3.5 De-identification Baseline Methods

In our large-scale empirical analysis, we examine nine popular de-identification methods of three different categories. For packages that offer multiple model options, we report the option with the highest performance in our experiments.[2]

Three off-the-shelf NLP libraries with the NER function:

- `spaCy` [64] (25.9k GitHub Stars) is widely adopted for industrial information extraction. We choose RoBERTa-base [82], which is pre-trained on a massive general-purpose corpus, as the backbone of its NER pipeline.
- `Stanza` [105] (6.6k GitHub Stars) is a natural language analysis package. We apply its 18-class NER model variant based on the contextual string representations [16] and pre-trained on the OntoNotes corpus [130].
- `flair` [15] (12.7k GitHub Stars) is a powerful NLP framework. We employ its large four-class NER model variant built on XLM-R embeddings [39] and document-level features [111] and pre-trained on the CoNLL03 corpus [108].

Three commercial services for PHI detection:

- `Amazon` Comprehend Medical DetectPHI Operation [4] is a HIPAA-eligible NLP service. We segment input notes into pieces shorter than 20,000 characters, the maximum allowed input length, when making the API calls.
- `Microsoft` Azure Cognitive Service for Language PHI Detection [9] de-identifies PHI information in unstructured texts. We divide notes into slices shorter than 5,120 characters to obey the input length threshold.
- `Google` Cloud Data Loss Prevention De-identification API [1] inspects and redacts sensitive data intelligently. We select the outputs for the class *PERSON_NAME* and remove the titles before the recognized full names.

[2]The number of GitHub Stars and citations listed below are accessed on April 24, 2023.

Method	Overall Performance (↑)			Bias along Dimensions (↓)			
	Precision	Recall	F1	Gender	Race	Popularity	Decade
spaCy	0.917±0.001	0.629±0.001	0.746±0.001	0.002*±0.001	0.013*±0.002	0.028*±0.002	0.007*±0.002
Stanza	0.678±0.001	0.881±0.001	0.766±0.001	0.002*±0.001	0.016*±0.002	0.011*±0.001	0.005*±0.001
flair	0.920±0.001	**0.974±0.000**	**0.946±0.000**	0.003*±0.000	**0.006*±0.001**	**0.008*±0.001**	0.002*±0.000
Amazon	**0.923±0.001**	0.925±0.001	0.924±0.001	0.005*±0.001	0.022*±0.001	0.032*±0.001	**0.001±0.001**
Microsoft	0.664±0.001	**0.960±0.001**	0.785±0.001	0.003*±0.001	0.023*±0.001	0.010*±0.001	0.006*±0.001
Google	0.609±0.001	0.869±0.001	0.716±0.001	0.009*±0.001	0.025*±0.001	0.014*±0.002	0.010*±0.001
NeuroNER	**0.946±0.001**	0.944±0.001	**0.945±0.000**	**0.001±0.001**	0.045*±0.001	0.026*±0.001	0.002±0.001
Philter	0.227±0.001	0.794±0.001	0.353±0.001	**0.000±0.001**	**0.000±0.001**	**0.003*±0.002**	**0.000±0.001**
MIST	0.474±0.001	0.751±0.001	0.581±0.001	0.013*±0.001	0.022*±0.002	0.017*±0.002	0.003*±0.002

Table 2: Overall performance (higher is better), bias along demographic dimensions (lower is better), and the associated bootstrapped standard error of the examined de-identification methods. We measure the bias with recall equality difference and bold the best two scores in each column. In particular, flair achieves the highest recall and F1 and the lowest bias for race and popularity. Moreover, the asterisk next to a bias score indicates a statistically significant difference in performance at an adjusted significance level (5% for gender, 0.833% for race, 1.667% for popularity and decade). A majority of the examined methods exhibit statistically significant performance gaps along most demographic dimensions.

We note that both Amazon Comprehend Medical DetectPHI Operation and Microsoft Azure Cognitive Service for Language PHI Detection are intended to be used for our specific case of free-text medical note de-identification. Google Cloud Data Loss Prevention De-identification is intended for the general text. We use this service because other medically-focused services operated by Google do not operate on free-text notes. Specifically, Google Cloud Healthcare API for de-identification [2] only operates on FHIR JSON embeddings and DICOM images, and Google Cloud Healthcare Natural Language API [8] only recognizes medical knowledge categories.

Three open-source de-identification toolkits:

- NeuroNER [43, 44] (212 citations) is an NER tool based on the long short-term memory model [63]. We use the model pre-trained on the 2014 i2b2 de-identification corpus [119] with GloVe word embeddings [102] and collect the outputs for *PATIENT* and *DOCTOR* as the set of recognized names.
- Philter (Protected Health Information filter) [97] (31 citations) leverages the Python NLTK module and regular expressions for rule-based de-identification.
- MIST (MITRE Identification Scrubber Toolkit) [10] (156 citations) is a suite of tools for identifying and redacting PHI in free-text medical records. We pre-train the model supplied by the Carafe engine, a conditional random field-based [76] sequence tagger, on the 2006 i2b2 de-identification corpus [127] as instructed and view the outputs for the classes *PATIENT* and *DOCTOR* as the set of recognized names.

3.6 Evaluation of Bias

To quantify the bias of each method along each dimension, we follow [87] by introducing the recall equality difference: the average absolute difference between the recall of each demographic group and that of all the groups along the corresponding demographic dimension. More specifically, for dimension D and its entailed set of demographic groups $\mathcal{G}^D = \{\mathcal{G}_1^D, \mathcal{G}_2^D, \dots\}$, recall equality difference $= \frac{1}{|\mathcal{G}^D|} \sum_{\mathcal{G}_i^D \in \mathcal{G}^D} |\text{Recall}(\mathcal{G}_i^D) - \text{Recall}(\mathcal{G}^D)|$. We use the recall equality difference as the fairness metric since it demonstrates the difference in recall each demographic group would experience while expecting the reported average performance. We also explore another fairness metric—recall maximum difference—and report the results in Appendix A.3.

We carry out the Wilcoxon signed-rank test [133] for the dimension of gender and the Friedman test [51] for the dimensions of race, popularity, and decade to assess the null hypothesis that a de-identification method treats all the groups equally well along a demographic dimension. After applying the Bonferroni correction [131], the adjusted significance levels for gender, race, popularity, and decade are 5%, 0.833%, 1.667%, and 1.667%, respectively.

4 Q1: IS THERE DEMOGRAPHIC BIAS?

Toward the first question of whether demographic bias exists in de-identification methods, we obtain two key takeaways. First, the tested de-identification methods perform differently, with some achieving a relatively high recall. Second, a majority of the methods exhibit statistically significant performance gaps along most demographic dimensions. Such disparities call for urgent review and action to address bias in existing de-identification methods.

4.1 Overall Performance Varies

We present the overall performance of the nine de-identification methods in Table 2. The performance varies across the methods with some methods obtaining a relatively high recall. In particular, flair performs rather well, especially in recall and F1, probably due to its use of large pre-trained language models and document-level features. NeuroNER also achieves competitive scores, especially in precision and F1, possibly because it is pre-trained on clinical corpora. In contrast, spaCy gives the lowest recall, which suggests a high risk of information leakage, albeit its popularity in the NLP community (it has the most GitHub Stars among the three NLP libraries we consider). Interestingly, Google dramatically underperforms compared to the other two commercial platforms (i.e., Amazon and Microsoft). As a rule-based method, Philter outputs highly imprecise predictions in the complicated clinical context.

4.2 Significant Demographic Gaps in De-identification Performance

We find that a majority of the examined methods demonstrate statistically significant differences in performance along most of the

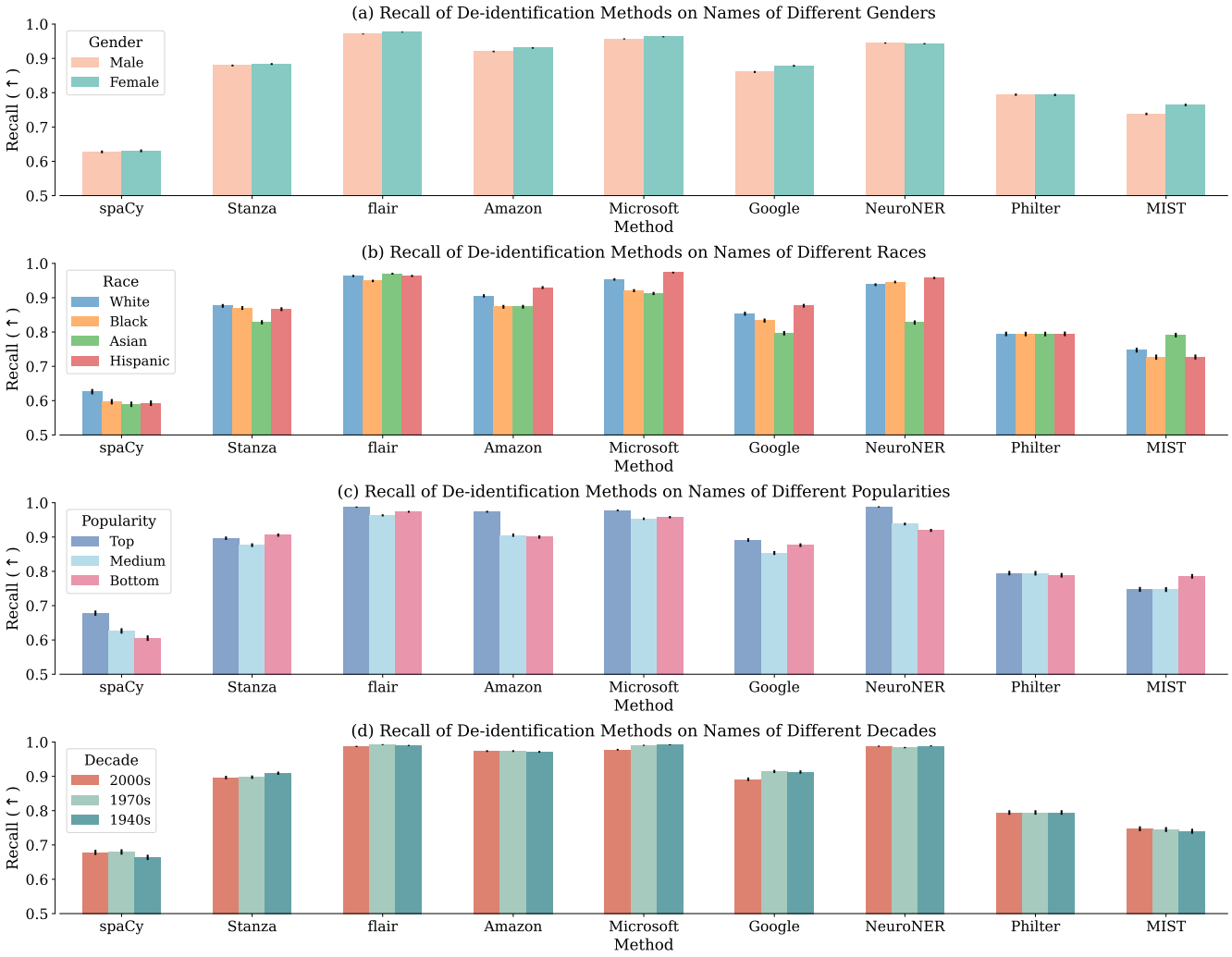

Figure 2: Recall and 95% bootstrapped confidence interval of the demographic groups along each dimension by each examined de-identification method. Disparities in performance between different groups are more obvious along the dimensions of race and popularity than along the dimensions of gender and decade.

four demographic dimensions. Table 2 exhibits the recall equality difference and the hypothesis test results, where an asterisk next to a score indicates a statistically significant difference at the corresponding significance level. In particular, `Amazon` and `Google` give the highest recall equality difference for name popularity and the decade of popularity, respectively, which should be a call for action for these commercial services. Although `NeuroNER` delivers an overall competitive de-identification performance, its recall equality difference is rather high, especially along the dimensions of race and popularity. We note that the rule-based `Philter` has very low bias and that `flair` achieves not only the highest recall but also relatively low recall equality differences along all four dimensions.

At a fine-grained level, we plot in Figure 2 the recall of the demographic groups along each dimension by each method. Along the dimension of gender, all the methods score better or equally well for female names than male names. Nevertheless, these methods act very differently in the four racial groups. More specifically, `Stanza` and `NeuroNER` attain very low recall in the Asian racial group, while `MIST` scores much higher. The three commercial services—`Amazon`, `Microsoft`, and `Google`—all perform better in the White and Hispanic racial groups than in the Black and Asian racial groups. Moreover, the performance of most methods deteriorates with the popularity of names, with the exceptions of `Stanza` and `MIST`. Finally, the disparity in recall among the three groups with different decades of popularity is more moderate. `Stanza`, `Microsoft`, and `Google` are more capable of recognizing popular names from more recent decades, while `spaCy` behaves oppositely.

We visualize in Figure 3 the recall of the 16 name sets averaged across the examined methods to further examine performance disparities. We observe that the average recall of the name sets with top popularity (i.e., Name Sets 1, 2, 13, 14, 15, and 16) outperforms the other sets. In addition, we note that the least popular names associated with the White racial group (i.e., Name Sets 5 and 6) score higher on average recall than the more popular names associated with the Asian racial group (i.e., Name Sets 9 and 10).

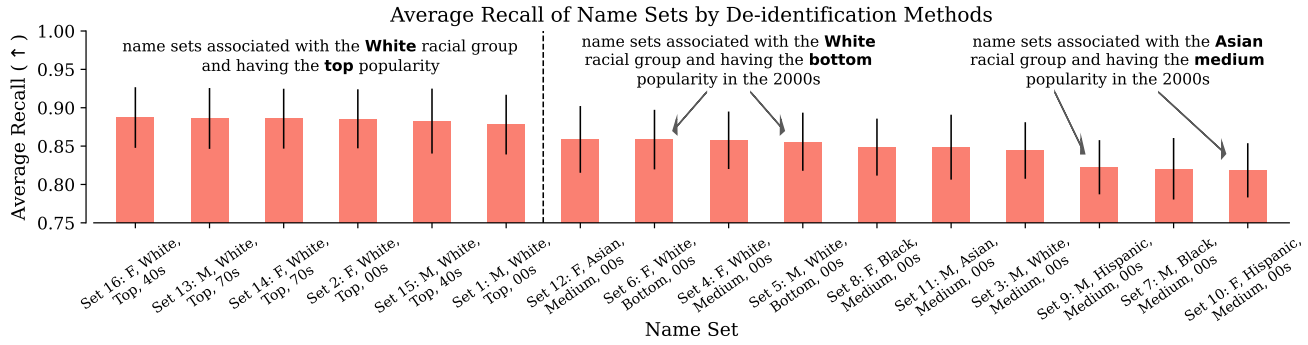

Figure 3: Average recall and standard error of each name set by the examined de-identification methods, ordered by decreasing recall. The average recall on name sets with top popularity exceeds the other sets by a clear margin. Moreover, the methods are, on average, more capable of recognizing less popular names associated with the White racial group compared to more popular names associated with the Asian racial group.

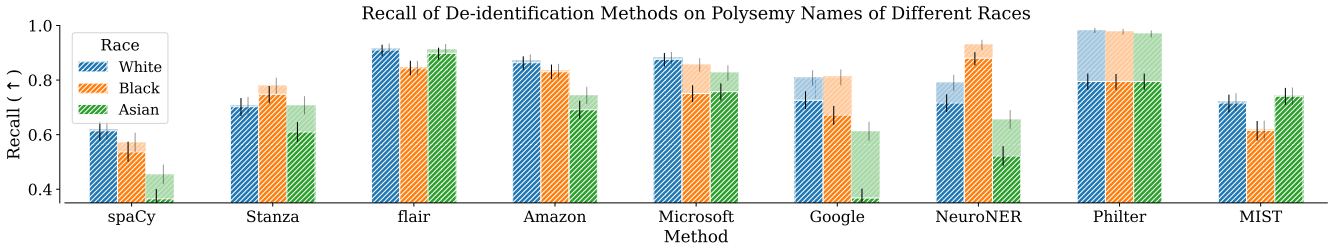

Figure 4: Recall and 95% bootstrapped confidence interval on polysemy first names associated with three racial groups by each examined de-identification method. The recall ranking among the three groups remains relatively consistent for most methods as that based on the original setting in Figure 2 (b). The increase in recall illustrated by the lighter color bar refers to the partially correct de-identification of non-polysemy last names.

5 Q2: WHAT CAUSES DE-IDENTIFICATION UNDERPERFORMANCE?

For the second question of what factors contribute to the underperformance, we draw three critical findings.

- Polysemy names account for methods' underperformance but not necessarily their demographic bias.
- Most methods are better at recognizing names in agreement with the gender suggested by the local context.
- Longer templates with more unique names and medication injection histories make de-identification harder.

5.1 Polysemy Names Lead to Underperformance

To understand what names are the hardest to recognize, we calculate the recall of each sampled name. We observe that names with the lowest recall usually have other meanings in English (i.e., polysemy). For instance, "*An*" in *An Dizon* and *An Son*—the two names with the lowest recall—is both a prevalent determiner in English and a first name of medium popularity associated with the Asian female group. "*Cleveland*" in *Cleveland Spikes*—the fifth hardest name by recall—is both a large city in the U.S. and a first name of medium popularity for the Black male group.

Therefore, we prepare five polysemy first names for each of the White, Black, and Asian racial groups as follows:

- White: Sydney, Faith, Forest, Cliff, June
- Black: Quincy, Cleveland, Kenya, Prince, Ivory
- Asian: Asian, Thai, King, Long, Young, Can

These sets share the same gender, name popularity, and the decade of popularity and only differ in race. Since we can only find five polysemy first names from the Black racial group and not enough polysemy first names from the Hispanic group that meet this requirement, we limit all sets to five names and omit the Hispanic group here. We then follow the procedure in Sec 3 and evaluate the methods on the polysemy first names listed above.

As shown in Figure 4, although we utilize polysemy first names for all three racial groups, the variation in performance persists. In addition, for all the methods except `Stanza`, the recall ranking of the three racial groups assessed on polysemy first names remains relatively consistent as that based on the original setting in Figure 2 (b). We also consider the scenario when a method can correctly recognize the non-polysemy last names and plot the increased recall above the original bar in lighter colors in Figure 4. In this case, most methods can see a significant increase in recall, especially for `Google`, `NeuroNER`, and `Philter`. Hence, names with overlapping meanings in English only explain the underperformance of the de-identification methods, but not necessarily their bias across demographic groups.

5.2 Methods Improve when De-identifying Context-Consistent Names

NER systems usually capture the contextual dependencies for tag decoding [78], and the semantic context often indicates the gender associated with a name. For example, titles (e.g., Mr. and Mrs.) can

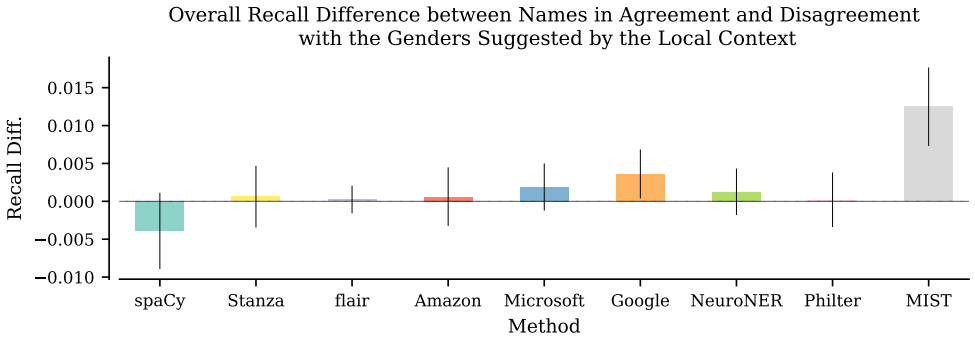

Figure 5: Difference in recall and 95% bootstrapped confidence interval between names that are consistent and inconsistent with the genders suggested by the local context. A positive recall difference means that performance was best when there was gender consistency, while a negative recall difference means that performance was best when there was gender inconsistency. Methods leveraging the gender context for name recognition are expected to see a positive recall difference.

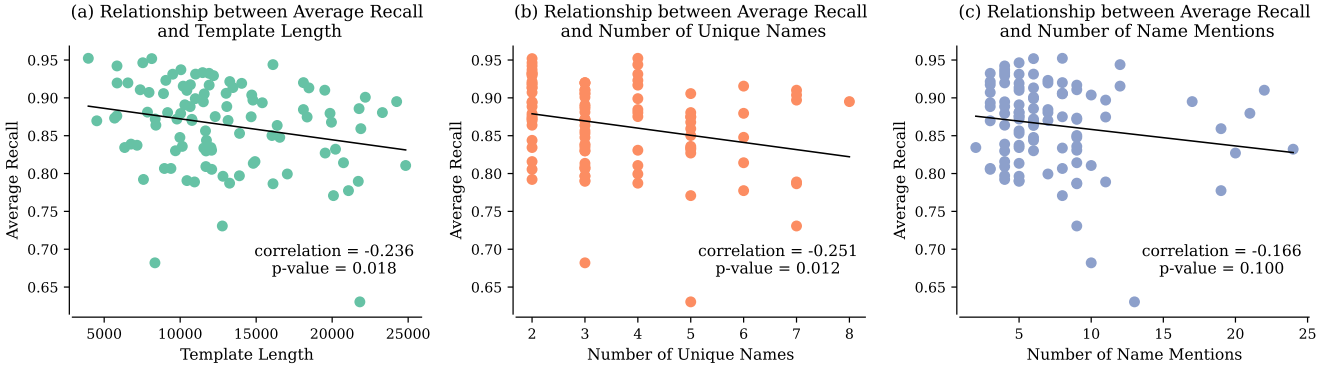

Figure 6: Relationship between template characteristics and template recall averaged across the examined methods. With statistically significant p-values, a template's average recall decreases with its length and the number of unique names included.

Template 86: (Average Recall = 0.625)

Patient was given:
DATE 08:29 PO/NG Atenolol 25 mg ***NAME***
DATE 08:29 PO/NG Aspirin 81mg ***NAME***
DATE 08:29 PO Cetirizine 10 mg ***NAME***

Template 67: (Average Recall = 0.675)

Medications on Admission:
amlodipine 10 mg tablet, 1 tablet(s) by mouth once a day
(***NAME*** ***DATE*** 20:17)
chlorthalidone 25 mg tablet, 1 tablet(s) by mouth once a day
(***NAME*** ***DATE*** 20:17)

Template 50: (Average Recall = 0.952)

Mr. ***NAME*** arrived from ***HOSPITAL***.
Mr. ***NAME*** is a pleasant, well-developed obese male in no acute distress.
It is very important to keep your appointment with your new PCP, Dr. ***NAME***.

Figure 7: Average recall and snippets of three templates. Unlike usual templates (e.g., Template 50), templates with a low average recall (e.g., Templates 86 and 67) usually include medication injection histories that offer little semantic context for name recognition.

appear before full names, and appositions (e.g., son and daughter) can describe relationships. We expect methods leveraging such context for name recognition to have higher recall on names where there is local context agreement with the gender as compared to those with disagreement. To assess this, we identify in our note templates where name gender can be easily inferred from the local context to determine if the consistency between the names and the inferred genders impacts de-identification quality.

Figure 5 plots the recall difference between context-consistent and -inconsistent names by the examined methods. Albeit with relatively large confidence intervals, We find that most methods perform better on names aligned with the implied gender. spaCy is the only exception, perhaps shedding light on its lowest overall recall (see Table 2).

Limitations of Gender-Inconsistent Evaluation in Experiment Setup. We acknowledge that replacing gender-inconsistent pronouns in notes prior to evaluation would be an easier test for models. However, we note that not all clinical records will contain gender-confirming pronouns, especially for transgender and non-binary individuals [84], and argue that de-identification methods should be able to operate properly in these gender-inconsistent situations. We also note that if we limit our analysis to only using male-originating notes with male name sets and female-originating notes with female name sets, our results still hold (see Appendix A.1). We note that in this setting, we do not explicitly assess the gender gap since male- and female-originating notes do not overlap.

Method	Fine-tuning Context	Fine-tuning Name	Overall Performance (↑) Precision	Recall	F1	Bias along Dimensions (↓) Gender	Race	Popularity	Decade
spaCy	out-of-the-box		0.916	0.623	0.741	**0.003**	0.027	0.025	0.005
	clinical	diverse	0.990±0.007	**0.950±0.006**	**0.969±0.002**	0.012±0.004	**0.024±0.005**	**0.005±0.002**	0.006±0.001
	clinical	popular	**0.998±0.004**	0.737±0.072	0.846±0.046	0.012±0.007	0.094±0.029	0.127±0.035	**0.003±0.004**
	general	diverse	0.915±0.072	0.830±0.083	0.864±0.035	0.036±0.005	0.071±0.011	0.049±0.042	0.008±0.005
	general	popular	0.873±0.110	0.492±0.069	0.629±0.083	0.010±0.003	0.059±0.032	0.326±0.060	0.007±0.003
NeuroNER	out-of-the-box		0.955	0.953	0.954	0.005	0.044	0.030	**0.001**
	clinical	diverse	0.978±0.014	**0.978±0.009**	**0.978±0.005**	0.007±0.001	**0.019±0.006**	**0.012±0.008**	0.002±0.001
	clinical	popular	**0.989±0.003**	0.865±0.021	0.923±0.013	0.008±0.004	0.065±0.007	0.118±0.010	**0.001±0.001**
	general	diverse	0.958±0.022	0.943±0.029	0.950±0.010	0.016±0.007	0.041±0.010	0.031±0.014	0.007±0.006
	general	popular	0.924±0.022	0.777±0.018	0.844±0.019	**0.003±0.001**	0.062±0.005	0.324±0.021	0.004±0.003

Table 3: Overall performance (higher is better) and bias along demographic dimensions (lower is better) of two de-identification methods fine-tuned with different setups. We measure the bias with recall equality difference, report the mean scores and standard errors based on five trials with different seeds, and bold the best score in each column for each method. For both methods, using clinical context and diverse names for fine-tuning improves the overall performance and reduces the demographic bias along most dimensions, especially race and popularity.

5.3 Performance Decays with Template Length and Name Quantity

Other properties of a note template may also affect the de-identification performance. We consider three characteristics—template length, number of unique names, and number of name mentions in a template—and visualize their relationships with a template's average recall in Figure 6. Our findings suggest that recall deteriorates with both the length of a note and the number of unique names that it contains.

We identify two of the worst-performing templates in terms of recall: Templates 86 and 67. These templates appear six and four times, respectively, in the five templates with the lowest recall by a method. As shown in Figure 7, unlike other templates (e.g., Template 50), Templates 86 and 67 are notable for having large blocks of medication history that provide little indication for the names that intersperse them. This unique characteristic of clinical records calls for special attention in future de-identification systems. We further investigate the performance of the examined methods on these hard templates in Appendix A.2 and find that their performance follows the overall pattern in Table 2.

6 Q3: CAN BIAS BE MITIGATED?

To answer the third question of how to mitigate the bias in de-identification methods, we propose a simple and method-agnostic solution of fine-tuning the methods with clinical context and diverse names. This setup not only improves the overall recall but also reduces the bias significantly along most demographic dimensions.

6.1 Fine-tuning De-identification Methods

We prepare the fine-tuning de-identification datasets by considering two types of context and two types of names. We treat the longitudinal clinical narratives in the 2014 i2b2 de-identification challenge [119] as the clinical context and the Wikipedia articles in the DocRED dataset [136] as the general context. We generate 160 diverse names by randomly sampling ten names from each of the 16 name sets in Table 1 and 160 popular names based on the most popular names over the three chosen decades that do not appear in the 16 name sets. For each type of context, we randomly sample 1,000 templates for training and 100 for validation. These templates are then populated with the names of each type (i.e., diverse names and popular names) separately. In this way, we create four fine-tuning setups in total by pairing the two types of context with the two types of names.

To compare the effectiveness of these setups, we fine-tune two de-identification methods—spaCy [64] and NeuroNER [43, 44]—with distinct out-of-the-box performance. spaCy is a widely-adopted NLP library that delivers a low de-identification recall and a moderate demographic bias in Table 2. In contrast, NeuroNER is pre-trained on the original 2014 i2b2 de-identification corpus, which yields a competitive recall with high bias along the dimensions of race and popularity. After fine-tuning with their respective default hyperparameters, these methods are evaluated on 1,600 test notes. These test notes are constructed by filling in the 100 templates in Sec 3.4 with the remaining ten names (not selected for the 160 diverse names during fine-tuning) from each of the 16 name sets separately. Here, the test notes are disjoint with the fine-tuning context/names.

6.2 Clinical Context and Diverse Names Improve Performance

Table 3 displays the overall performance and the demographic bias (i.e., the recall equality difference) of the two methods after fine-tuning. We repeat the fine-tuning five times with different seeds and report the mean scores and standard errors. Impressively, despite the distinct out-of-the-box performance of the two fine-tuned methods, the setup composed of clinical context and diverse names largely enhances the overall performance of both methods and diminishes their unfairness, especially along the dimensions of race and popularity.

In particular, although most of the four fine-tuning setups improve spaCy's overall performance, fine-tuning with clinical context and diverse names sees the largest boost in spaCy's recall by over 0.3. On the other hand, since NeuroNER is pre-trained on clinical corpora, most of the four fine-tuning setups are ineffective in enhancing NeuroNER's strong out-of-the-box performance. However, fine-tuning with clinical context and diverse names is the only exception here, which increases the precision, recall, and F1 of

NeuroNER by around 0.02 each. Moreover, along the dimensions of race and popularity, where the degree of unfairness is rather high, this setup can significantly reduce the bias of both methods.

We suggest that fine-tuning de-identification methods with clinical context and diverse names should be done as an immediate fix to improve fairness before applying the methods to clinical tasks. The method-agnostic effectiveness and simplicity of this setup highlight the importance of training data diversity to model fairness [85].

7 DISCUSSION

Demographic Associations of Names. Names can be associated with certain demographic features [52, 81]. For instance, in our U.S. Social Security [6] and Census [3] data sources, there is variation in name popularity between self-reported ethnic groups. In human decision-making, such associations have been shown to correlate with discriminative hiring [22, 60] and loan granting [61] practices. Other work has explored the biases learned by large language models when the demographic context is varied directly in input [79] or using names as a proxy for demographic [87, 91, 94]. For example, NLP models link the female gender to specific stereotypical occupations [28] and tend to generate violent or negative-toned text when given "Muslim" as a demographic descriptor for input [11]. We emphasize that the biases inherently learned by NLP models may perpetuate biases and, therefore, require careful audits. We acknowledge that our analysis based on de-identifying names may not necessarily generalize to other PHI types and leave this further investigation to future work.

Bias in Healthcare. Bias in healthcare can occur in both systematic and implicit ways based on demographic factors such as race, ethnicity, gender, sexual orientation, or socio-economic status [48, 59, 88, 137]. These biases can then be unintentionally learned by ML models [14, 57, 101]. For instance, NLP models trained on race-redacted clinical notes have been shown to capture self-reported race through other proxy information [12] and mimic the existing biases in text completions for clinical treatment decisions [138]. Our study demonstrates that existing clinical de-identification methods discriminate based on the demographic associations of names. The bias in these methods could further escalate the unfairness in downstream healthcare systems.

Importance of De-identified Data for Reproducibility. ML models rely on large amounts of data for training [19], but in the case of health data, there are privacy concerns. By removing PHI, researchers can protect stakeholders' privacy with de-identified data [112] and avoid biasing their models through more representative datasets [37]. To this end, clinical de-identification has attracted long-lasting attention from the research community [73, 93] and large amounts of resources from the industrial world (e.g, [5, 7]). We highlight the importance of equitable de-identification because legal and ethical data sharing should be encouraged [112] to improve the reproducibility of clinical findings and the credibility of healthcare systems [90, 124].

Harm of Minority Exclusion. We stress that it is not acceptable to exclude some populations from de-identified data sharing. When demographic groups are absent in data, models trained on that data will perform poorly on the missing groups [98]. This can result in misdiagnoses, inadequate treatments, and a failure to address health disparities [56]. Hence, it is crucial to ensure that data for model training is diverse and representative of the populations they will serve [37]. Future work should consider proactive measures to collect and include data from underrepresented populations and address systemic biases during data collection and analysis.

Ramifications of Poorer Privacy for Marginalized Groups. General disparities in de-identification performance can lead to poorer privacy for marginalized groups and engender crimes such as identity theft [17, 18]. This adds to the existing difficulties with data collection and monitoring faced by marginalized communities [30, 47]. Even when data sharing is consented, the data can be used outside of the given context, leading to representational harm for groups that are already targeted [54]. In future work, we advocate for data collection and de-identification practices that promote trust and do not discourage minorities from seeking medical care and participating in clinical data sharing.

Importance of Audits to Create Change. Audits in healthcare help to identify areas of improvement [69], assess compliance with regulations and standards [66], and hold organizations accountable for their actions [106]. Past work on ML audits has demonstrated the ability to make meaningful changes and reduce performance gaps in deployed systems with biases. For example, a recent audit on the bias in automated facial analysis algorithms [31] stimulated the targeted companies to reduce accuracy disparities between demographic groups [106], However, companies that provided similar algorithms and were not included in the original audit did not make corresponding changes [106]. We encourage clinical practitioners to build upon our de-identification audit to provide high-quality, equitable de-identification services to all demographic groups.

8 CONCLUSION

In this paper, we contribute a large-scale empirical analysis of de-identifying names from clinical records and present findings that demonstrate systemic bias in performance. Our results should sound the alarm for clinical and ML stakeholders, as bias in clinical de-identification not only raises legal concerns but also make certain demographic groups more prone to privacy leakage. Hence, we call for an urgent review of existing de-identification methods and actions (e.g., fine-tuning with our recommended setup) to improve the fairness and accountability of healthcare systems.

Despite the comprehensiveness of our study, we acknowledge the limitation of using coarse racial and gender categorizations when constructing our name sets. In addition, while our analysis is readily applicable to many widely-adopted de-identification methods, we did not evaluate its generalization to approaches focusing on other PHI classes. We leave to future work the investigation of bias in de-identifying other PHI classes based on more fluid racial and gender categorizations.

ACKNOWLEDGMENTS

This project is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362. We would like to acknowledge the contributions of Dana Moukheiber, Lama Moukheiber, and Mira Moukheiber in annotating the clinical note templates used in our experiments.

REFERENCES

[1] [n. d.]. De-identifying sensitive data | Data Loss Prevention Documentation | Google Cloud — cloud.google.com. https://cloud.google.com/dlp/docs/deidentify-sensitive-data. [Accessed 24-November-2022].

[2] [n. d.]. De-identifying sensitive data | cloud healthcare API | google cloud. https://cloud.google.com/healthcare-api/docs/how-tos/deidentify. [Accessed 24-November-2022].

[3] [n. d.]. Decennial Census Surname Files (2010, 2000) — census.gov. https://www.census.gov/data/developers/data-sets/surnames.html. [Accessed 30-June-2022].

[4] [n. d.]. Detect PHI - Amazon Comprehend Medical — docs.aws.amazon.com. https://docs.aws.amazon.com/comprehend-medical/latest/dev/textanalysis-phi.html. [Accessed 24-November-2022].

[5] [n. d.]. HealthVerity Census – Real-Time Patient Identity Resolution Technology — healthverity.com. https://healthverity.com/solutions/healthverity-census/. [Accessed 06-Feb-2023].

[6] [n. d.]. Popular Baby Names — ssa.gov. https://www.ssa.gov/oact/babynames/limits.html. [Accessed 30-June-2022].

[7] [n. d.]. Privacy Analytics - Software to Anonymize Text — privacy-analytics.com. https://privacy-analytics.com/health-data-privacy/health-data-software/software-to-anonymize-text/. [Accessed 06-Feb-2023].

[8] [n. d.]. Using the healthcare natural language API | cloud healthcare API | google cloud. https://cloud.google.com/healthcare-api/docs/how-tos/nlp. [Accessed 24-November-2022].

[9] [n. d.]. What is the Personally Identifying Information (PII) detection feature in Azure Cognitive Service for Language? - Azure Cognitive Services — learn.microsoft.com. https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/personally-identifiable-information/overview. [Accessed 24-November-2022].

[10] John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. 2010. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *International journal of medical informatics* (2010).

[11] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

[12] Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

[13] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*.

[14] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. 2020. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[15] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. 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)*.

[16] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*.

[17] Keith B Anderson. 2005. Identity theft: Does the risk vary with demographics? *Federal Trade Commission, Bureau of Economics Working Paper* (2005).

[18] Keith B Anderson. 2006. Who are the victims of identity theft? The effect of demographics. *Journal of Public Policy & Marketing* (2006).

[19] Andrew L Beam and Isaac S Kohane. 2018. Big data and machine learning in health care. *Jama* (2018).

[20] Bruce A Beckwith, Rajeshwarri Mahaadevan, Ulysses J Balis, and Frank Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC medical informatics and decision making* (2006).

[21] Anna Bergdall, Stephen Asche, Nicole Schneider, Tessa Kerby, Karen Margolis, JoAnn Sperl-Hillen, Jaime Sekenski, Rachel Pritchard, Michael Maciosek, and Patrick O'Connor. 2012. CB3-01: comparison of ethnicity and race categorization in electronic medical records and by Self-report. *Clinical Medicine & Research* (2012).

[22] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* (2004).

[23] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

[24] Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

[25] Catherine Bliss. 2012. *Race decoded: The genomic fight for social justice*. Stanford University Press.

[26] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[27] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).

[28] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* (2016).

[29] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*.

[30] Simone Browne. 2015. *Dark matters: On the surveillance of blackness*. Duke University Press.

[31] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*.

[32] Aidan Byrne and Alessandra Tanesini. 2015. Instilling new habits: addressing implicit bias in healthcare professionals. *Advances in Health Sciences Education* (2015).

[33] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* (2017).

[34] Jie Cao, Xiaosong Zhang, Vahakn Shahinian, Huiying Yin, Diane Steffick, Rajiv Saran, Susan Crowley, Michael Mathis, Girish N Nadkarni, Michael Heung, et al. 2022. Generalizability of an acute kidney injury prediction model across health systems. *Nature Machine Intelligence* (2022).

[35] Kimberly Wright Cassidy, Michael H Kelly, and Lee'at J Sharoni. 1999. Inferring gender from name phonology. *Journal of Experimental Psychology: General* (1999).

[36] Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

[37] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical machine learning in healthcare. *Annual review of biomedical data science* (2021).

[38] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* (2020).

[39] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[40] 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* (2021).

[41] 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*.

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

[43] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

[44] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of Patient Notes with Recurrent Neural Networks. *Journal of the American Medical Informatics Association (JAMIA)* (2016).

[45] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society* (2020).

[46] Russell Eisenman. 1995. Is there bias in US law enforcement? *The Journal of Social, Political, and Economic Studies* (1995).

[47] Jeffrey Fagan, Anthony A Braga, Rod K Brunson, and April Pattavina. 2016. Stops and stares: Street stops, surveillance, and race in the new policing. *Fordham Urb. LJ* (2016).

[48] Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics* (2017).

[49] Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. 2014. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association* (2014).

[50] F Jeff Friedlin and Clement J McDonald. 2008. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association* (2008).

[51] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* (1937).

[52] S Michael Gaddis. 2017. How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science* (2017).

[53] Melanie Ganz, Sune H Holm, and Aasa Feragen. 2021. Assessing bias in medical ai. In *Workshop on Interpretable ML in Healthcare at International Connference on Machine Learning (ICML)*.

[54] Marzyeh Ghassemi and Shakir Mohamed. 2022. Machine learning and health need better values. *npj Digital Medicine* (2022).

[55] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. 2020. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits on Translational Science Proceedings* (2020).

[56] Marzyeh Ghassemi and Elaine Okanyene Nsoesie. 2022. In medicine, how do we machine learn anything real? *Patterns* (2022).

[57] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* (2018).

[58] Matthew W Hahn and R Alexander Bentley. 2003. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London. Series B: Biological Sciences* (2003).

[59] William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley. 2015. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American journal of public health* (2015).

[60] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*.

[61] Andrew Hanson, Zackary Hawley, Hal Martin, and Bo Liu. 2016. Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics* (2016).

[62] J Andrew Harris. 2015. What's in a name? A method for extracting information about ethnicity from names. *Political Analysis* (2015).

[63] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[64] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020).

[65] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

[66] Lisanne Hut-Mossel, Kees Ahaus, Gera Welker, and Rijk Gans. 2021. Understanding how and why audits work in improving the quality of hospital care: A systematic realist review. *PloS one* (2021).

[67] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*.

[68] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[69] Noah Ivers, Gro Jamtvedt, Signe Flottorp, Jane M Young, Jan Odgaard-Jensen, Simon D French, Mary Ann O'Brien, Marit Johansen, Jeremy Grimshaw, and Andrew D Oxman. 2012. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane database of systematic reviews* (2012).

[70] Abigail Z Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[71] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* (2023).

[72] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* (2016).

[73] Mehmet Kayaalp. 2017. Modes of De-identification. In *AMIA Annual Symposium Proceedings*.

[74] 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*.

[75] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

[76] John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

[77] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[78] 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* (2020).

[79] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*.

[80] Shulammite Lim, Alistair Johnson, Yuxin Xiao, Dana Moukheiber, Lama Moukheiber, Mira Moukheiber, Marzyeh Ghassemi, and Tom Pollard. 2023. Annotated MIMIC-IV discharge summaries for a study on deidentification of names (version 1.0). *PhysioNet* (2023). https://doi.org/10.13026/ngc0-0f54.

[81] Wendy Liu and Derek Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*.

[82] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[83] Bernard Lo. 2015. Sharing clinical trial data: maximizing benefits, minimizing risk. *Jama* (2015).

[84] Jeffrey W Lockhart, Molly M King, and Christin Munsch. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour* (2023).

[85] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. 2022. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence* (2022).

[86] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*.

[87] Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. 2022. Behind the Mask: Demographic bias in name detection for PII masking. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*.

[88] Jasmine R Marcelin, Dawd S Siraj, Robert Victor, Shaila Kotadia, and Yvonne A Maldonado. 2019. The impact of unconscious bias in healthcare: how to recognize and mitigate it. *The Journal of infectious diseases* (2019).

[89] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. 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)*.

[90] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. 2021. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* (2021).

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

[92] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* (2021).

[93] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology* (2010).

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

[95] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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)*.

[96] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[97] Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. 2020. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine* (2020).

[98] Natalia Norori, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. 2021. Addressing bias in big data and AI for health care: A call for open science. *Patterns* (2021).

[99] Jessica H Ochs. 2022. Addressing health disparities by addressing structural racism and implicit bias in nursing education. *Nurse Education Today* (2022).

[100] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*.

[101] Ravi B Parikh, Stephanie Teeple, and Amol S Navathe. 2019. Addressing bias in artificial intelligence in health care. *Jama* (2019).

[102] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

[103] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

[104] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2020. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering* (2020).

[105] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[106] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

[107] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

[108] Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

[109] 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 annual meeting of the association for computational linguistics*.

[110] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics* (2021).

[111] Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993* (2020).

[112] Kenneth P Seastedt, Patrick Schwab, Zach O'Brien, Edith Wakida, Karen Herrera, Portia Grace F Marcelo, Louis Agha-Mir-Salim, Xavier Borrat Frigola, Emily Boardman Ndulue, Alvin Marcelo, et al. 2022. Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digital Health* (2022).

[113] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Chen, and Marzyeh Ghassemi. 2021. Medical imaging algorithms exacerbate biases in underdiagnosis. (2021).

[114] Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[115] K Shailaja, B Seetharamulu, and MA Jabbar. 2018. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*.

[116] Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing Gender Bias in Word Embeddings with Latent Disentanglement and Counterfactual Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

[117] Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. 2021. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics* (2021).

[118] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[119] Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics* (2015).

[120] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[121] Sean M Thomas, Burke Mamlin, Gunther Schadow, and Clement McDonald. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method.. In *Proceedings of the AMIA Symposium*.

[122] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* (2019).

[123] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* (2019).

[124] K TSIMA. 2023. The reproducibility issues that haunt health-care AI. *Nature* (2023).

[125] Katherine Tucker, Janice Branson, Maria Dilleen, Sally Hollis, Paul Loughlin, Mark J Nixon, and Zoë Williams. 2016. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC medical research methodology* (2016).

[126] Konstantinos Tzioumis. 2018. Demographic aspects of first names. *Scientific data* (2018).

[127] Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association* (2007).

[128] Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine* (2008).

[129] Craig S Webster, Saana Taylor, Courtney Thomas, and Jennifer M Weller. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA education* (2022).

[130] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA* (2013).

[131] Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/* (2004).

[132] David R Williams and Ronald Wyatt. 2015. Racial bias in health care and health: challenges and opportunities. *JAMA* (2015).

[133] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007).

[134] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*.

[135] Hui Yang and Jonathan M Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics* (2015).

[136] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of ACL 2019*.

[137] Colin A Zestcott, Irene V Blair, and Jeff Stone. 2016. Examining the presence, consequences, and reduction of implicit bias in health care: a narrative review. *Group Processes & Intergroup Relations* (2016).

[138] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*.

[139] 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)*.

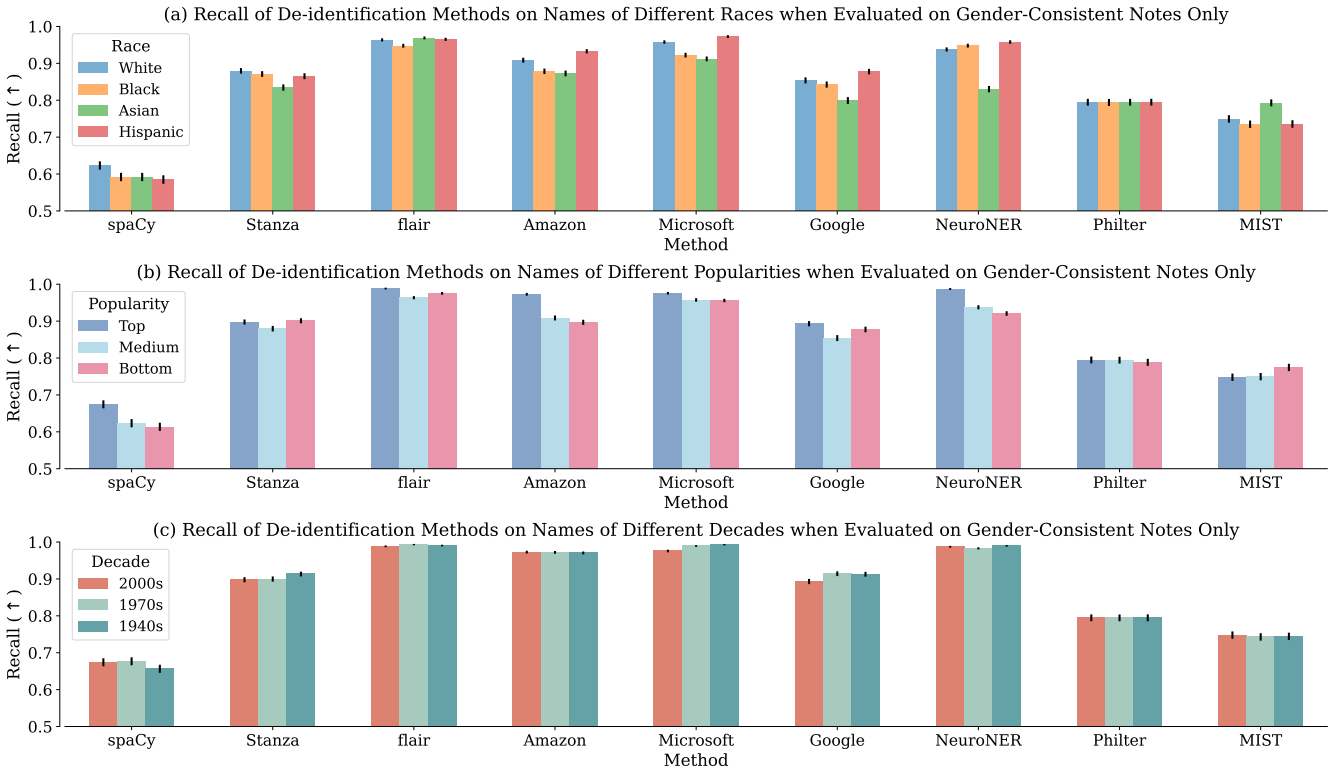

Figure 8: Recall and 95% bootstrapped confidence interval of the demographic groups along the dimensions of race, name popularity, and the decade of popularity by each examined de-identification method under gender-consistent evaluation. These methods behave similarly compared to the original setup in Figure 2.

A APPENDIX

In the appendix, we include additional analysis exploring the robustness of our results in gender-consistent note population, in the subset of notes with the poorest overall performance, and using another fairness metric of recall maximum difference.

A.1 Gender-Consistent Note Population

To examine the influence of gender-inconsistent pronouns used in our note template population, we run a robustness check on our results where we only consider male-originating clinical notes populated with male name sets and female-originating notes populated with female name sets. We note that in this setting, we do not conduct a direct comparison of the gender gap since the male- and female-originating notes are disjoint. Otherwise, the experiment follows the procedure in Sec 3.

Figure 8 illustrates the recall of the demographic groups along the dimensions of race, name popularity, and the decade of popularity by each de-identification method under this gender-confirming evaluation setup. The Wilcoxon signed-rank test with p-value = 0.082 indicates that these methods behave consistently to the original setup in Sec 3, and our observations about the race, popularity, and decade disparities based on Figure 2 still hold.

Method	Overall Performance (↑)			Bias along Dimensions (↓)			
	Precision	Recall	F1	Gender	Race	Popularity	Decade
spaCy	0.874±0.003	0.504±0.003	0.640±0.003	0.004±0.003	0.022*±0.004	0.037*±0.005	0.005±0.004
Stanza	0.615±0.003	0.791±0.003	0.692±0.002	**0.001±0.002**	**0.007±0.003**	0.028*±0.004	0.011*±0.003
flair	0.878±0.002	**0.945±0.001**	**0.910±0.001**	0.005*±0.001	0.014*±0.002	0.016*±0.002	0.004*±0.001
Amazon	**0.882±0.002**	0.883±0.002	0.883±0.002	0.009*±0.002	0.025*±0.003	0.047*±0.003	**0.003*±0.002**
Microsoft	0.619±0.003	**0.936±0.002**	0.745±0.002	0.003*±0.001	0.033*±0.003	0.013*±0.002	0.009*±0.002
Google	0.558±0.003	0.856±0.002	0.676±0.002	0.011*±0.002	0.034*±0.003	**0.011*±0.003**	0.008*±0.003
NeuroNER	**0.929±0.002**	0.899±0.002	**0.914±0.001**	0.005*±0.002	0.044*±0.003	0.052*±0.003	0.005±0.002
Philter	0.134±0.001	0.562±0.003	0.216±0.002	**0.000±0.002**	**0.000±0.003**	**0.003*±0.004**	**0.000±0.004**
MIST	0.306±0.002	0.532±0.003	0.388±0.002	0.020*±0.003	0.040*±0.004	0.019*±0.005	0.009*±0.004

Table 4: Overall performance (higher is better), bias along demographic dimensions (lower is better), and the associated bootstrap standard error of the examined de-identification methods on the hardest 20 templates. We measure the bias with recall equality difference and bold the best two scores in each column. These methods' overall performance follows the general pattern when evaluated on the full set of 100 templates in Table 2. Some methods exhibit lower bias here, possibly due to equally poor performance across demographic groups in harder templates.

Method	Recall Maximum Difference (↓)			
	Gender	Race	Popularity	Decade
spaCy	0.002±0.002	0.025±0.004	0.042±0.004	0.010±0.004
Stanza	0.002±0.001	0.032±0.003	0.017±0.003	0.008±0.002
flair	0.003±0.001	**0.013±0.002**	**0.013±0.001**	0.003±0.001
Amazon	0.005±0.001	0.034±0.002	0.047±0.002	**0.001±0.001**
Microsoft	0.003±0.001	0.033±0.002	0.015±0.001	0.009±0.001
Google	0.009±0.001	0.044±0.003	0.020±0.003	0.015±0.003
NeuroNER	**0.001±0.001**	0.089±0.003	0.040±0.001	0.003±0.001
Philter	**0.000±0.001**	**0.000±0.002**	**0.004±0.003**	**0.000±0.002**
MIST	0.013±0.002	0.043±0.004	0.026±0.004	0.004±0.003

Table 5: Recall maximum difference (lower is better) and the associated bootstrapped standard error of the examined de-identification methods. We bold the best two scores in each column. The bias in these methods measured by recall maximum difference along each dimension is similar to the pattern measured by recall equality difference in Table 2.

A.2 Evaluation of Difficult Note Templates

Here we identify the set of 20 templates that receive the lowest average recall by the examined de-identification methods and investigate how the performance of these methods changes in Table 4 when evaluated on these harder templates. Although the scores of their overall performance drop compared to Table 2, the best-performing methods based on the original full set of 100 templates still perform well on these hardest 20 templates. However, some of the examined methods, such as Stanza and Google, exhibit lower bias now, potentially due to equally poor performance across demographic groups in harder templates.

A.3 Recall Maximum Difference

Besides recall equality difference, we consider an additional fairness metric—recall maximum difference, which illustrates the largest gap in recall any demographic group would experience while anticipating the reported average performance. For dimension D and its entailed set of demographic groups $\mathcal{G}^D = \{\mathcal{G}_1^D, \mathcal{G}_2^D, \dots\}$, recall maximum difference = $\max_{\mathcal{G}_i^D \in \mathcal{G}^D} |\text{Recall}(\mathcal{G}_i^D) - \text{Recall}(\mathcal{G}^D)|$.

Table 5 displays the recall maximum difference of each examined de-identification method along each dimension. These methods' behaviors here are similar to their bias measured by recall equality difference in Table 2. Methods that attain the lowest recall equality difference still perform well in terms of recall maximum difference.