Auditing Learned Associations in Deep Learning Approaches to Extract Race and Ethnicity from Clinical Text

Oliver J. Bear Don't Walk IV, PhD[1], Adrienne Pichon, MPH, MPhil, MA[2], Harry Reyes Nieva, MAS, MA[2,3], Tony Sun, MA[2], Jaan Altosaar, PhD[4], Karthik Natarajan, PhD[2], Adler Perotte, MD, MA[2], Peter Tarczy-Hornoch, MD[1], Dina Demner-Fushman, MD, PhD[5], Noémie Elhadad, PhD[2]

[1]University of Washington, Seattle, WA; [2]Columbia University, New York, New York; [3]Harvard Medical School, Boston, Massachusetts; [4]One Fact Foundation, Claymont, DE; [5]US National Library of Medicine, Bethesda, Maryland

Abstract

Complete and accurate race and ethnicity (RE) patient information is important for many areas of biomedical informatics research, such as defining and characterizing cohorts, performing quality assessments, and identifying health inequities. Patient-level RE data is often inaccurate or missing in structured sources, but can be supplemented through clinical notes and natural language processing (NLP). While NLP has made many improvements in recent years with large language models, bias remains an often-unaddressed concern, with research showing that harmful and negative language is more often used for certain racial/ethnic groups than others. We present an approach to audit the learned associations of models trained to identify RE information in clinical text by measuring the concordance between model-derived salient features and manually identified RE-related spans of text. We show that while models perform well on the surface, there exist concerning learned associations and potential for future harms from RE-identification models if left unaddressed.

Introduction

Complete and accurate race and ethnicity (RE) patient information is important for many areas of biomedical informatics research and clinical practice, such as defining and characterizing cohorts, performing quality assessments, and identifying health inequities. The electronic health record (EHR) provides a rich source of patient health data; but while RE can exist as structured data in the EHR, this information is often missing or inaccurate[1]. The clinical narrative provides an alternative, potentially more accurate source of RE information, and natural language processing (NLP) techniques have been proposed to extract RE from clinical text[1].

Methods such as imputation, multi-source data linkage, and NLP have been used to supplement missing or inaccurate RE data. For example, the Bayesian Improved Surname Geocoding[2] approach uses a patient's geocoded address and last name to compute the probability of a patient belonging to a given racial or ethnic group[3–5]. Other imputation approaches rely on surnames to identify Latino[6,7] and Asian/Pacific Islander[8] patients. Previous work has shown imputation approaches relying on surname/geocoding perform poorly for people from Native American and/or multi-racial backgrounds[9]. Data linkage approaches using multiple sources, such as claims data and cancer registries, have also been used to supplement missing/inaccurate RE data in the EHR[7,10,11]. Alternatively, clinical notes can be used to extract clinician-assessed RE information with NLP[1]. Sholle et al. leveraged a rule-based approach to extract RE information from clinical text, performing well for Black and Latino patients[1]. NLP provides a wide range of approaches to identify patient-associated features from clinical notes, such as social determinants of health[12–14].

While a rich source to recover RE data, the clinical narrative can contain negative or stigmatizing language (e.g., "aggressive" or "refuse") biased against marginalized racial groups[15–17]. Furthermore, NLP models have been shown to learn harmful language associations from clinical text[18,19]. These issues are compounded by the difficulty of auditing opaque deep learning models. While many methods exist to audit clinical NLP approaches[20], auditing learned associations is an equally important approach to interrogate bias, promote interpretability, and build trust. Without a deep understanding of the biases present in clinical text, it will be difficult to detect when clinical NLP models may inherit or exacerbate such biases. While previous work has interrogated learned association biases in clinical NLP models for RE identification[19], these approaches have not yet leveraged model-derived salient features to understand the potential for bias. In this work we examined the associations between inputs and outputs to interrogate potentially biased associations learned by models trained to identify RE in clinical text. We performed this bias audit by measuring the concordance between model-derived salient features and manually annotated spans of text that are potentially informative for RE identification (indicators). Understanding this concordance can improve RE identification research and provide insights into human biases that models could propagate.

Methods

We describe our approach to audit deep learning models for biases in their learned associations for the task of sentence-level RE labeling. We examine the concordance between spans of text known to contain explicit RE information and model-derived salient features. Our approach relies on sentences with two sets of gold-standard annotations for RE labels and information related to RE. We describe the dataset, classification task and model training, salient feature extraction, and bias audit approach.

Data

We sampled sentences from the Contextualized Race and Ethnicity Annotations for Clinical Text (C-REACT) dataset[21], built from MIMIC-III[22]. The sample consists of 5,834 sentences from clinical notes annotated for RE labels at the sentence level and RE indicators at the span level. RE labels consisted of the US census categories, as well as "No Information Indicated" to signal a sentence does not convey any RE information, and "Not Covered" to signal the presence of RE that falls outside of the census categories. Of importance to this work is the "No Information Indicated" label, which serves as the only negative label when a sentence lacks any information on a patient's race or ethnicity. Indicators included direct mentions of race or ethnicity (e.g., "Native American", "Black", "Latino", "non-Latino") as well as explicit discussions of country/nation of origin or geographic (country) and primary, preferred, or spoken language (language).

Sentences with the label white were the most common (n=3,318), followed by sentences with labels for Black and/or African American (Black/AA) (n=542), Latino (n=502), and Asian (n=398). Training and test sets were created for each label (Black/AA, white, Asian, and Latino) using a 75/25 split with approximately 50% positive and 50% negative labels (Table 1). For each RE category, all positively labeled sentences were randomly sampled and distributed among the training or test sets. Negative labels were defined as "No Information Indicated" labels or labels for other RE categories (e.g., Black/AA labels could be negative labels for the sentences labeled with Asian).

Table 1: Descriptive statistics for the training and test sets.

		Black/AA	Asian	Latino	White
Train					
	Sentences	815	599	753	4,977
	Positive labels	406	299	375	2,474
	Negative labels	409	300	378	2,503
	Patients	598	413	509	3,010
	Notes	689	467	598	3,678
Test					
	Sentences	273	199	253	1,678
	Positive labels	136	99	127	844
	Negative labels	137	100	126	834
	Patients	174	122	147	831
	Notes	214	153	194	1,155

Model Training and Evaluation

We trained one model for each of the four RE labels using binary classification. These RE "information models" were trained to identify sentences with information on a patient's race or ethnicity. The base of the models used a frozen BERT-small model that had been further pre-trained on PubMed[23] and then general clinical text data[24] to create contextual word embeddings. These embeddings were fed into a convolutional neural network followed by a single fully connected layer to classify each sentence. Given that the model pre-training step did not clean data by removing punctuation/numbers and performing lemmatization, no text cleaning beyond sentence splitting was performed. Models were trained for 200 epochs with a learning rate of 1e-05 and a batch size of 128. The best performing model was defined as the model with the best F1-score that improved over the previous best F1-score by at least 0.5% on the training set. We deployed a frozen BERT-base model given the relatively small training data. These models are collectively referred to as the BERT-based models.

We also developed baseline models following a previous approach by Sholle et al., where each label has its own set of regular expressions and a single mention constituted a positive prediction[1]. Regular expressions were used without any modification besides those needed to translate code to Python. These models are collectively referred to as the baseline models.

Salient Feature Extraction

We extracted salient features using integrated gradients[25] for all BERT-based models. This approach leverages gradients and input values to identify important features, while overcoming issues with insensitivity to small perturbation, common in well-trained models[25]. Following Sundararajan et al., we used embeddings of all zeros as a baseline for each input. The integrated gradients approach multiplies gradient scores by the input value to output saliency scores for each dimension in each token embedding. Following Ding and Koehn's approach[26], we calculate a token-level saliency score by summing over each dimension's saliency score.

Bias Audit

Learned association biases between input tokens and output classes were audited using two approaches. The first follows previous work[26] to rank all features in a sentence using saliency scores and to measure the median ranking of the most salient indicator and the percentage of sentences with at least one indicator in the top one, two, and three most salient features for all sentences ("ranking metrics"). Ranking metrics convey information on how often indicators are highly salient features for information models. The second approach ranks tokens by how frequently they are identified as one of the three most salient tokens in each sentence, over all sentences ("highly salient tokens"). Studying highly salient tokens uncovers exactly what kinds of tokens are salient, even if they're not indicators. Positive saliency scores indicate that a feature influenced the model towards a positive prediction, and so positive scores were used for all rankings concerning positive predictions. Similarly, for negative predictions, negative scores indicate that a feature influenced the model towards a negative prediction, and were thus used for rankings concerning negative predictions. In both cases, we refer to highly ranked features as the most salient features.

Finally, while RE indicators could provide a strong signal for identifying RE in clinical text, there could exist other weaker signals. To identify these weaker signals, we mask all indicators with the special BERT token "[MASK]" in the training and test sets previously described ("masked data") and assess classification performance and highly salient features. For example, the sentence "patient is an elderly Native American male" would become "patient is an elderly [MASK] male". The "[MASK]" token was used because during BERT's pre-training it indicated a missing token to be inferred by BERT from the surrounding context. While ours is a classification task, the underlying need to infer missing information given the context is still there. In this scenario, classification performance can flag the presence of signals outside of RE indicators and highly salient features could provide more specific details. Models were re-trained and re-evaluated on the masked data following the previously described approach.

Results

We first present results on model classification performance using F1, recall, and precision. Next, we present the bias audit results with ranking metrics and highly salient features. Finally, we interrogate what features are highly salient in an extreme scenario where RE indicator spans are removed from sentences.

Model Performance

All BERT-based models achieved a test set F1-score above 0.80, with the white information model performing the best (F1=0.90), followed by the Latino (F1=0.88), Black/AA (F1=0.84) and Asian (F1=0.83) information models (Table 2). Each baseline model outperformed its respective BERT-based counterpart, except for the white information model. The regular expressions for the white information model baseline extracted ambiguous terms like "white", which were often used in discussions not pertaining to race (e.g., "patient's white blood cell count is elevated"). We experimented with other architectures using a long short-term memory network or only a fully connected neural network on top of the frozen BERT-small, but found that they performed significantly worse than the current model.

Table 2: Test set performance metrics

	BERT-based Models				Baseline models			
Performance Metric	Black/AA	Asian	White	Latino	Black/AA	Asian	White	Latino
F1	0.84	0.83	0.90	0.88	0.97	0.99	0.87	0.96
Precision	0.85	0.84	0.87	0.86	0.98	1.0	0.78	0.93
Recall	0.82	0.82	0.94	0.89	0.97	0.99	1.0	1.0

Bias Audit: Ranking Metrics

For positive predictions, the median rank of the most salient indicator token was between 1 and 3 for each information model, with the Black/AA and Latino Information models ranking an indicator as the most salient feature a majority of the time (Table 3). The Latino information model had the highest percentage of top highly ranked indicators, ranging from 80.6% to 97.7% for the top one and top three metrics, followed by the Black/AA and white information models. The Asian information model ranked indicators much lower than the other models, ranging from 26.9% to 54.8% for the top one and top three metrics. Overall, all models contained at least one indicator in the top three most salient features a majority of the time: Latino (97.7%), Black/AA (89.3%), white (81.4%), and Asian (54.8%).

Ranking metrics for true positive cases were the same or better than the overall positive case. For false positive cases, all ranking metrics dropped significantly, especially for the Asian Information model. While all models had indicator tokens in the top three most salient features a majority of the time, the Asian information model was significantly lower with indicators in the top three most salient features only 7.7% of the time. Furthermore, the median rank of the most salient indicator is 28, compared to three or two for all other models.

Table 3: Indicator saliency score ranking metrics for positive, true positive, and false positive cases. Median is the median rank of the most salient indicator. Top n is the percentage of times an indicator is in the n most salient features.

	All Positive				True Positive				False Positive			
Model	Median	Top 1 (%)	Top 2 (%)	Top 3 (%)	Median	Top 1 (%)	Top 2 (%)	Top 3 (%)	Median	Top 1 (%)	Top 2 (%)	Top 3 (%)
Black/AA	1	56.5	84.0	89.3	1	58.0	87.5	93.8	2	47.4	63.2	63.2
Asian	3	26.9	38.7	54.8	3	31.2	43.8	62.5	28	0.0	7.7	7.7
White	2	42.1	68.9	81.4	2	43.4	71.7	84.0	3	28.6	40.3	54.5
Latino	1	80.6	96.9	97.7	1	88.3	100.0	100.0	2	33.3	77.8	83.3

Bias Audit: Highly Salient Tokens

While ranking metrics can tell us whether models are relying on indicators for predictions, interrogating exactly what kinds of features are highly salient can provide a more complete picture. Tables 4 and 5 contain the counts of salient features (features that are in the top three most salient for each sentence) across true positive, false positive (Table 4), true negative, and false negative (Table 5) cases. In true positive prediction cases, information models highly ranked tokens that coincided with indicator spans (Table 4). For example, positive predictions for the Black/AA information model ranks indicators related to race like “african” and “american” highly, while the white information model ranks “white” and “caucasian" highly. The Latino information model ranked language indicators or tokens related to language like “spanish” and “speaking”. Highly ranked tokens for the Asian information model included country and language indicators like “vietnamese” and “chinese", while also ranking special BERT tokens used during BERT’s pre-training phase such as “[CLS]” and “[SEP]” as highly salient features more often than other models. False positive cases often included features that are not directly related to a model’s specific race/ethnicity information task. For example, the most highly ranked salient feature for the Black/AA information model is “caucasian”, while the second most highly ranked salient feature for the Hispanic information model is “russian" (Table 4).

While there is not enough room to present highly salient features in overall positive cases, we discuss them here. For the Black/AA information model on positive cases, five out of 10 features are indicators, but one of these, “caucasian”, is not considered predictive for the Black/AA category. While “american” was never marked as an indicator by itself, it was a part of the span “african american” in 65 out of 66 instances with the remaining case a part of the span “haitian american”. In positive prediction cases for the Latino model, three of the top 10 features are indicators, however the indicator “russian” is not linked to information on a patient being Latino. In contrast, the white and Asian information models had three indicators out of 10 tokens, all of which were directly related to their respective race categories.

Table 4: Counts for highly ranked salient features across positive prediction cases. A feature is considered highly ranked if it is one of the top three most salient features in a sentence.

True Positive Cases			False Positive Cases		
Salient Features Across Models	Top Salient Feature Count n (% of top ranked features)		Salient Features Across Models	Top Salient Feature Count n (% of top ranked features)	
Black/AA					
"african"	70	(20.8)	"caucasian"	10	(17.5)
"american"	66	(19.6)	"female"	7	(12.3)
"[SEP]"	43	(12.8)	"[SEP]"	4	(7.0)
"female"	30	(8.9)	"physical"	2	(3.5)
"aa"	16	(4.8)	":"	2	(3.5)
"male"	16	(4.8)	"male"	2	(3.5)
"man"	12	(3.6)	"[CLS]"	2	(3.5)
"black"	9	(2.7)	","	2	(3.5)
"woman"	8	(2.4)	"in"	2	(3.5)
"haitian"	8	(2.4)	"##ic"	2	(3.5)
Asian					
"[CLS]"	53	(22.1)	"[CLS]"	13	(33.3)
"[SEP]"	24	(10.0)	"[SEP]"	6	(15.4)
"vietnamese"	14	(5.8)	"physician"	2	(5.1)
"chinese"	11	(4.6)	"-"	1	(2.6)
"old"	9	(3.8)	"physical"	1	(2.6)
"to"	8	(3.3)	","	1	(2.6)
"cambodian"	8	(3.3)	"with"	1	(2.6)
"-"	7	(2.9)	"p"	1	(2.6)
"male"	6	(2.5)	"81"	1	(2.6)
"korean"	5	(2.1)	"first"	1	(2.6)
White					
"white"	338	(14.2)	"[SEP]"	23	(10.0)
"caucasian"	235	(9.9)	"old"	21	(9.1)
"old"	229	(9.6)	"male"	11	(4.8)
"[SEP]"	222	(9.3)	"year"	10	(4.3)
"male"	175	(7.4)	"african"	9	(3.9)
"year"	146	(6.1)	":"	9	(3.9)
"female"	128	(5.4)	"spanish"	9	(3.9)
"race"	102	(4.3)	"female"	9	(3.9)
"russian"	88	(3.7)	"speaking"	8	(3.5)
"last"	51	(2.1)	"woman"	7	(3.0)
Latino					
"spanish"	95	(28.5)	"speaking"	14	(25.9)
"speaking"	74	(22.2)	"russian"	11	(20.4)
"[SEP]"	23	(6.9)	"female"	7	(13.0)
"male"	17	(5.1)	"all"	2	(3.7)
"hispanic"	15	(4.5)	"chinese"	2	(3.7)
"only"	13	(3.9)	"man"	2	(3.7)
"female"	10	(3.0)	"only"	1	(1.9)
"interpreter"	9	(2.7)	"vietnamese"	1	(1.9)
"man"	7	(2.1)	"year"	1	(1.9)
"old"	5	(1.5)	"male"	1	(1.9)

In true negative and false negative prediction cases, models ranked special BERT tokens “[SEP]” and “[CLS]” highly (Table 5). For true negative cases all models featured at least one indicator token as a highly ranked salient feature (e.g., “white” for the Black/AA, Asian, and Latino information models, and “vietnamese” for the white information model). The Asian information model ranked “american” highly in true negative cases (Table 5), all of which were part of the indicator span “african american” (“asian american” was only present twice in the entire Asian sentences

dataset). Highly ranked features for false negative (Table 5) cases did not contain any indicators in the top 10 most salient features and often contained special BERT tokens or language pertinent to discussing a patient's clinical presentation or history (e.g., "present", "general", "distress", and "history").

The Asian information model ranked special BERT tokens particularly high in terms of salient features. Removing these tokens from consideration during ranking raised each top n metric for overall positive cases by 5-15 percentage points compared to the results presented in Table 3 and brought the median rank of the most salient indicator to two.

Table 5: Counts for highly ranked salient features across negative prediction cases. A feature is considered highly ranked if it is one of the top three most salient features in a sentence.

True Negative Cases			False Negative Cases		
Salient Features Across Models	Top Salient Feature Count n (% of top ranked features)		Salient Features Across Models	Top Salient Feature Count n (% of top ranked features)	
Black/AA					
"[SEP]"	63	(23.9)	"[SEP]"	15	(20.8)
"white"	26	(9.8)	"[CLS]"	12	(16.7)
"[CLS]"	17	(6.4)	"present"	4	(5.6)
"history"	11	(4.2)	"was"	2	(2.8)
"."	8	(3.0)	"distress"	2	(2.8)
"old"	7	(2.7)	"major"	2	(2.8)
"distress"	6	(2.3)	"##ys"	2	(2.8)
"##uri"	6	(2.3)	"the"	2	(2.8)
"##shed"	6	(2.3)	"illness"	1	(1.4)
"dental"	4	(1.5)	"##par"	1	(1.4)
Asian					
"[SEP]"	35	(17.7)	"[SEP]"	9	(16.7)
"white"	27	(13.6)	"distress"	3	(5.6)
"male"	14	(7.1)	"general"	2	(3.7)
"female"	10	(5.1)	"physical"	2	(3.7)
"caucasian"	9	(4.5)	"[CLS]"	2	(3.7)
"distress"	5	(2.5)	"##lip"	2	(3.7)
"general"	5	(2.5)	"##mia"	2	(3.7)
"american"	4	(2.0)	"##les"	2	(3.7)
"cooperative"	4	(2.0)	"examination"	1	(1.9)
"corona"	4	(2.0)	"mel"	1	(1.9)
White					
"[SEP]"	17	(18.9)	"[SEP]"	29	(21.0)
"[CLS]"	11	(12.2)	"[CLS]"	25	(18.1)
"."	9	(10.0)	"."	12	(8.7)
"history"	8	(8.9)	"the"	6	(4.3)
"their"	3	(3.3)	"history"	5	(3.6)
"vietnamese"	2	(2.2)	"patient"	4	(2.9)
"to"	2	(2.2)	"with"	4	(2.9)
"]"	2	(2.2)	"through"	2	(1.4)
"past"	2	(2.2)	":"	2	(1.4)
")"	2	(2.2)	"on"	2	(1.4)
Latino					
"[SEP]"	68	(24.9)	"[CLS]"	11	(26.2)
"[CLS]"	55	(20.1)	"[SEP]"	10	(23.8)
"."	22	(8.1)	"."	3	(7.1)
"with"	14	(5.1)	","	3	(7.1)
"a"	11	(4.0)	"history"	2	(4.8)
":"	8	(2.9)	"a"	2	(4.8)
"white"	8	(2.9)	"in"	1	(2.4)
"in"	8	(2.9)	"past"	1	(2.4)

"male"	5	(1.8)	"n"	1	(2.4)
","	4	(1.5)	"##lip"	1	(2.4)

Masked data test performance dropped when compared to the original test data for all information models, with the white information model performing the best (F1=0.89), followed by the Latino (F1=0.77), Black/AA (F1=0.71), and Asian (F1=0.70) information models. Highly salient features (tables not shown due to space) for true positive cases often included demographic information such as “female”, “man”, “old”, etc. Tokens related to language like “interpreter” and “speaking” were highly salient in the Latino, Asian, and white information models, but not the Black/AA model. Finally, the Black/AA model found the token “obe” (as in “obese”) highly salient, which occurred more often in sentences labeled with Black/AA, but was never marked as an indicator. Salient features for false positive cases often included “obe” for the Black/AA information model and “speaking” for the white and Latino information models. Negative predictions were once again dominated by “[SEP]” and “[CLS]” and were difficult to interpret without any indicators in the most salient features.

Discussion

We presented audit results on the learned associations of a deep learning model trained to identify RE information in clinical text by measuring the concordance between model-derived salient features and manually annotated spans of text that are potentially informative for RE identification. We found that model performance in terms of F1-score did not necessarily translate to high reliance on explicit indicators for RE as seen in the ranking metrics and highly salient features. Importantly, three general patterns were noted in biased or incorrect learned associations: 1) benign artifacts; 2) helpful but not universally correct; and 3) helpful but ultimately biased and/or harmful if not addressed.

While all BERT-based information models performed well, they were outperformed by previously vetted rule-based, baseline models[1]. The rule-based models from Sholle et al., performed surprisingly well given the shift in geography (Boston vs New York), deidentification of MIMIC-III clinical notes, and the focus on the critical care unit. This high performance across datasets can be explained by the similarities in the RE categories used, drawing from federal standards[27].

Overall, model classification performance did not align with ranking metric performance. The best performing model was the white information model, followed by the Latino, Black/AA, and Asian information models, while the Latino model had the best ranking metrics followed by Black/AA, white, and Asian information models. Of note, the Asian information model’s classification performance was similar to other models, but performed drastically worse in terms of ranking metrics. While removing special BERT tokens (e.g., “[CLS”]) did improve the ranking metric for the Asian information model, a noticeable difference remained. This difference could be an artifact of pre-training given the importance of these tokens in pre-training tasks and the potential difficulty of unlearning these associations given the small training data for the RE identification task.

While ranking metrics can tell us how often indicators are highly salient features, exploring highly salient tokens provided insight into what kinds of associations the models learned. In addition to strong evidence for plausible learned associations for their respective models such as “african”, “caucasian”, “chinese”, and “spanish", we found evidence for three categories of biased or incorrect learned associations. The first category included benign artifacts such as “[CLS]”, “[SEP]”, and punctuation tokens that many models ranked highly. These are likely artifacts of the pre-training regimen for BERT models.

The second class of learned associations included helpful but not strictly correct associations such as language indicator-related tokens like “speaker”, “interpreter”, “spanish”, and “russian”. Indicator tokens in sentences labeled with Latino were largely made up of language indicators, hence the reliance on “spanish”. However, this reliance on language-related tokens is not always correct. In false positive cases, when “speaking” is a top three most salient feature it always occurs in the context of the language indicator spans like “russian", “chinese", or “vietnamese". A similar association was observed for the white information model, where “speaking” and “spanish” were highly ranked features potentially explained by the association of language indicators (e.g., “russian") with the white label in the training data. Overall, this could be explained by the way these tokens are represented in a similar vector space via their learned contextualized word embeddings[28].

The third class of learned associations included associating indicators for other RE categories with negative predictions which are potentially harmful if not properly addressed. For example, the Black/AA and Asian information models ranked “white” and “caucasian” as highly salient features when making negative predictions. In addition, the Asian

information model also ranked "american" highly for true negative cases, which were always a part of the span "african american". This behavior could be due to the fact that multiple race labels for the same sentence only occurred twice in the entire corpus. While this is not necessarily an issue in our models given that we trained one model per category, researchers training models in a multi-label setting would need to audit models that learn to identify all classes at once for this behavior. This behavior is concerning given that racial and ethnic categories are not necessarily mutually exclusive, and these associations could further perpetuate the erasure of multi-racial and multi-ethnic patients if applied without the proper skepticism.

While we did not find evidence for learned associations with negative or stigmatizing language found in previous work[15,16,19], the masked data results show that there exist signals even without explicit indicators present. Many highly salient features were shared by the original and masked saliency results, however a token relating to discussion of obesity ("obe") was the only token related to a medical condition and was only highly salient for the Black/AA information model. Information models performed surprisingly well even with explicit indicators removed. The white information model performance dropped by only a few points between the masked and original test sets. Our results are in line with previous work[19], and add to the evidence that removing explicit indicators of RE is not sufficient to prevent racial or ethnic biases in clinical NLP models. Unlike previous work focusing on the entire clinical note[19], our work focused on highly curated sentences with mentions of patients, demographic features, and explicit indicators. The highly curated nature of the C-REACT corpus could be a contributing factor to the lack of stigmatizing or negative language in the highly salient features.

The results of this work should be considered in light of three main limitations. The first is that our modeling process did not explore hyper-parameters that could lead to increased performance. It was not our objective to achieve state-of-the-art performance in automated RE labeling, but rather to audit models for potentially biased learned associations. The second limitation speaks to the generalizability of our results. The MIMIC-III dataset represents data from a single hospital, specific to the critical care setting, and the training data for RE categories was relatively small for deep learning approaches. While we used a frozen, clinically relevant BERT model to reduce the burden of training from scratch, it is likely that the training data was still too small for truly generalizable learning. Finally, it has been shown that different models and saliency measures can provide different ranking metric results[26] and so other saliency approaches might shed light on the full picture. Alternative approaches to measuring highly salient features might also take into account the random chance of a feature being ranked highly to determine highly salient and statistically significant features while addressing co-occurrence statistics.

Future research could improve our work by prioritizing performance, exploring different approaches to learned association audits and addressing learned association biases. Given the BERT-based models did not outperform the baseline, future directions could include hyper-parameter tuning and further optimization to boost performance above the baseline. Sholle et al., measured how many patients had RE data unique to clinical notes, which we touch on in a publication in progress with the gold-standard dataset used here[21]. Exploring further approaches to auditing learned associations could leverage additional saliency approaches such, as vanilla gradient methods[29], SmoothGrad[30], or rationale interrogation approaches[31]. Future work could also focus on addressing issues identified during auditing, by leveraging a training regimen that balances performance while relying on the appropriate information such as the explicit indicators used here, which can be accomplished by combining the RE indicator and the label gold-standard datasets[32]. Solutions to learned association biases must also take into account social and historical contexts that influence patient and provider understandings of what are the "appropriate" associations to learn and ultimately apply to patients and their data. Once bias audits have been performed and concerning learned associations are found, it will be important for future researchers to address these biases. Certain solutions may be more technical in nature, such as using indicators to explicitly guide models toward appropriate learned associations during training[32]. However, in the case of negative or stigmatizing language being used by models to infer race or ethnicity, it is important to acknowledge that no technical solution will solve the role that systemic and personal racism has on the data generation process.

Conclusion

In this work, we presented audit results on the learned associations of a deep learning model trained to identify RE information in clinical text by measuring the concordance between model-derived salient features and manually annotated spans of text that are potentially informative for RE identification. We found three general patterns in incorrect or biased learned associations: 1) benign artifacts; 2) helpful but not universally correct; and 3) helpful but ultimately biased and/or harmful if not addressed. Furthermore, models were still able to identify RE information contained in sentences with explicit mentions removed. Given the literature about negative patient descriptors in clinical text[15,16], and BERT-based models' ability to inherit biases in the training data[18], it is important to audit models

for unjust and/or biased associations. Auditing learned associations for bias is one approach to promoting interpretability, improving model trust, and ethically leveraging clinical machine learning models to recover patient-level attributes such as RE and social determinants of health. Moving forward, understanding the "appropriate" associations to leverage for RE identification and other clinical NLP tasks will be important for both humans and machines to understand.

Acknowledgements

This work was supported by grants from the National Library of Medicine (OBDW, HRN, TS, AP: T15LM007079, LR: R01 LM006910), Intramural Research Program of the National Library of Medicine and National Institutes of Health (OBDW, DDF), the Computational and Data Science Fellowship from the Association for Computing Machinery Special Interest Group in High Performance Computing (HRN), and the NIH-funded Artificial Intelligence and Machine Learning for the Advancement of Health Equity and Researcher Diversity (AIM-AHEAD program. (OBDW). OBDW completed much of this work at Columbia University, but finished writing and editing while at the University of Washington. The authors would like to acknowledge Dr. Andrea Hartzler and her lab for their insightful feedback and comments and Dr. Trevor Cohen for GPU access.

References

1. Sholle ET, Pinheiro LC, Adekkanattu P, Davila MA, Johnson SB, Pathak J, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. J Am Med Inform Assoc. 2019 Apr 26;26(8–9):722–9.
2. Elliott M, Fremont A, Morrison P, Pantoja P, Lurie N. A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity. Health Serv Res. 2008 Oct;43(5 Pt 1):1722–36.
3. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc. 2021 Dec 28;29(1):187–96.
4. Dembosky JW, Haviland AM, Haas A, Hambarsoomian K, Weech-Maldonado R, Wilson-Frederick SM, et al. Indirect Estimation of Race/Ethnicity for Survey Respondents Who Do Not Report Race/Ethnicity. Med Care. 2019 May;57(5):e28–33.
5. Wei II, Virnig BA, John DA, Morgan RO. Using a Spanish Surname Match to Improve Identification of Hispanic Women in Medicare Administrative Data. Health Serv Res. 2006 Aug;41(4 Pt 1):1469–81.
6. Morgan RO, Wei II, Virnig BA. Improving identification of Hispanic males in Medicare: use of surname matching. Med Care. 2004 Aug;42(8):810–6.
7. Pinheiro PS, Sherman R, Fleming LE, Gomez-Marin OW, Huang Y, Lee DJ, et al. Validation of ethnicity in cancer data: which Hispanics are we misclassifying? Journal of registry management. 2009 Jan 1;36(2):42–6.
8. Hsieh MC, Pareti LA, Chen VW. Using NAPIIA to improve the accuracy of Asian race codes in registry data. J Registry Manag. 2011 Jan 1;38(4):190–5.
9. LeRoy L, Wasserman M, Rezaee M, White A. Understanding Disparities in Persons with Multiple Chronic Conditions: Research Approaches and Datasets. Abt Associates [Internet]. 2013 [cited 2022 Mar 24]; Available from: https://aspe.hhs.gov/reports/understanding-disparities-persons-multiple-chronic-conditions-research-approaches-datasets-0
10. Bigback KM, Hoopes M, Dankovchik J, Knaster E, Warren-Mears V, Joshi S, et al. Using Record Linkage to Improve Race Data Quality for American Indians and Alaska Natives in Two Pacific Northwest State Hospital Discharge Databases. Health Services Research. 2015;50(S1):1390–402.
11. McClure LA, Koru-Sengul T, Hernandez MN, Mackinnon JA, Schaefer Solle N, Caban-Martinez AJ, et al. Availability and accuracy of occupation in cancer registry data among Florida firefighters. PLoS One. 2019;14(4):e0215867.
12. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. Journal of Biomedical Informatics. 2021 Jan 1;113:103631.
13. Yetisgen M, Vanderwende L. Automatic Identification of Substance Abuse from Social History in Clinical Text. In: Artificial Intelligence in Medicine [Internet]. Springer, Cham; 2017 [cited 2018 Jun 27]. p. 171–81. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-319-59758-4_18
14. Feller DJ, Zucker J, Bear Don't Walk IV O, Srikishan B, Martinez R, Evans H, et al. Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning. AMIA Annu Symp Proc. 2018 Dec 5;2018:422–9.

15. Sun M, Oliwa T, Peek ME, Tung EL. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. Health Affairs. 2022 Feb;41(2):203–11.
16. Himmelstein G, Bates D, Zhou L. Examination of Stigmatizing Language in the Electronic Health Record. JAMA Netw Open. 2022 Jan 27;5(1):e2144967.
17. Glassberg J, Tanabe P, Richardson L, DeBaun M. Among emergency physicians, use of the term "Sickler" is associated with negative attitudes toward people with sickle cell disease. Am J Hematol. 2013 Jun;88(6):532–3.
18. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM Conference on Health, Inference, and Learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2020 Aug 25]. p. 110–20. (CHIL '20). Available from: https://doi.org/10.1145/3368555.3384448
19. Adam H, Yang MY, Cato K, Baldini I, Senteio C, Celi LA, et al. 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 [Internet]. New York, NY, USA: Association for Computing Machinery; 2022 [cited 2023 Mar 16]. p. 7–21. (AIES '22). Available from: https://dl.acm.org/doi/10.1145/3514094.3534203
20. Bear Don't Walk OJ IV, Reyes Nieva H, Lee SSJ, Elhadad N. A scoping review of ethics considerations in clinical natural language processing. JAMIA Open. 2022 Jul 1;5(2):ooac039.
21. Bear Don't Walk OJ, Pichon A, Reyes Nieva H, Sun T, Altosaar J, Joseph J, et al. C-REACT: Contextualized Race and Ethnicity Annotations for Clinical Text [Internet]. physionet.org; 2023 [cited 2023 Mar 16]. Available from: https://doi.org/10.13026/***** [Accepted for Publication]
22. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.
23. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task [Internet]. Florence, Italy: Association for Computational Linguistics; 2019 [cited 2020 Jul 11]. p. 58–65. Available from: https://www.aclweb.org/anthology/W19-5006
24. Bear Don't Walk Iv OJ, Sun T, Perotte A, Elhadad N. Clinically relevant pretraining is all you need. J Am Med Inform Assoc. 2021 Aug 13;28(9):1970–6.
25. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney, NSW, Australia: JMLR.org; 2017. p. 3319–28. (ICML'17).
26. Ding S, Koehn P. Evaluating Saliency Methods for Neural Language Models. arXiv:210405824 [cs] [Internet]. 2021 Apr 12 [cited 2022 Jan 12]; Available from: http://arxiv.org/abs/2104.05824
27. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity [Internet]. The White House. [cited 2021 Dec 19]. Available from: https://obamawhitehouse.archives.gov/node/15626
28. Thompson L, Mimno D. Topic Modeling with Contextualized Word Representation Clusters. arXiv:201012626 [cs] [Internet]. 2020 Oct 23 [cited 2022 May 4]; Available from: http://arxiv.org/abs/2010.12626
29. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:13126034 [cs] [Internet]. 2014 Apr 19 [cited 2022 Jan 11]; Available from: http://arxiv.org/abs/1312.6034
30. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv:170603825 [cs, stat] [Internet]. 2017 Jun 12 [cited 2022 Apr 19]; Available from: http://arxiv.org/abs/1706.03825
31. Vafa K, Deng Y, Blei D, Rush A. Rationales for Sequential Predictions. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing [Internet]. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021 [cited 2022 May 2]. p. 10314–32. Available from: https://aclanthology.org/2021.emnlp-main.807
32. Ross AS, Hughes MC, Doshi-Velez F. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence [Internet]. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization; 2017 [cited 2021 Dec 29]. p. 2662–70. Available from: https://www.ijcai.org/proceedings/2017/371