Emerging Challenges in Personalized Medicine: Assessing Demographic Effects on Biomedical Question Answering Systems

Sagi Shaier,[▽] Kevin Bennett,[◇] Lawrence Hunter,[†] Katharina von der Wense[▽◇♠]
[▽]University of Colorado Boulder
[†]University of Colorado Denver
[◇]Memorial Healthcare System
[◇]Johannes Gutenberg University Mainz
[▽]E-mail: {sagi.shaier, katharina.kann}@colorado.edu
[†]E-mail: larry.hunter@cuanschutz.edu
[◇]E-mail: kevbennett@mhs.net

Abstract

State-of-the-art question answering (QA) models exhibit a variety of social biases (e.g., with respect to sex or race), generally explained by similar issues in their training data. However, what has been overlooked so far is that in the critical domain of biomedicine, *any* unjustified change in model output due to patient demographics is problematic: it results in the unfair treatment of patients. Selecting only questions on biomedical topics whose answers do *not* depend on ethnicity, sex, or sexual orientation, we ask the following research questions: (RQ1) Do the answers of QA models change when being provided with irrelevant demographic information? (RQ2) Does the answer of RQ1 differ between knowledge graph (KG)-grounded and text-based QA systems? We find that irrelevant demographic information change up to 15% of the answers of a KG-grounded system and up to 23% of the answers of a text-based system, including changes that affect accuracy. We conclude that unjustified answer changes caused by patient demographics are a frequent phenomenon, which raises fairness concerns and should be paid more attention to. Code and data can be found here: https://github.com/Shaier/personalized_medicine_challenges.

1 Introduction

Natural language processing (NLP) has long been used in health care and life sciences. However, NLP systems exhibit surprising behaviors that can be difficult to predict or control: problems with general-purpose NLP systems reflecting stereotyping and stigmatizing biases have been apparent since the Microsoft Taybot debacle in 2016 and remain a major issue to this day (Nadeem et al., 2021; Rudinger et al., 2017; Blodgett et al., 2020; Savoldi et al., 2021; Zarrieß et al., 2022).

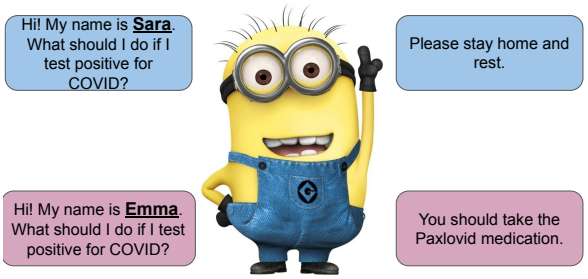

Figure 1: An undesired behavior from a biomedical QA system: the model changes its answers when provided with different biomedically irrelevant information (e.g., names).

The World Health Organization states that social determinants of health, including the experience of racism, sexism, and other forms of discrimination, "can be more important than health care or lifestyle choices in influencing health."[♠] Thus, for biomedical NLP systems it is of particular importance to not be affected by factors irrelevant to biology and medicine, and for researchers to ensure they serve their users fairly irrespective of irrelevant attributes, such as names, as shown in Figure 1. Here, we test the effect irrelevant demographic information has on biomedical question answering (QA) systems.

As a test-bed, we choose a subset of questions from the US Medical Licensing Exam level 1 (USMLE1; Jin et al., 2021) **whose answers, according to two medical professionals, are independent of the patient's demographics.** Although the questions are multiple-choice, correct answers require broad medical knowledge, including diagnosis and treatment of all common diseases, as well

[♠]Formerly: Katharina Kann

[♠]https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 540–550
November 1–4, 2023.

as an understanding of the underlying molecular and physiological mechanisms, potential drug side effects, probabilistic reasoning, and more.

We add irrelevant demographic information in a controlled way to the USMLE questions in order to answer the following research questions: (RQ1) Do the models' answers change when being provided with irrelevant demographic information? (RQ2) Is the answer to RQ1 different for knowledge graph (KG)-grounded and text-based QA systems? We experiment with two biomedical QA systems: BioLinkBERT (Yasunaga et al., 2022), a text-based model, and QAGNN (Yasunaga et al., 2021), which is the highest performing KG-based model on USMLE.

There are good reasons to believe that neither system should be affected by irrelevant patient information: both are trained solely on biomedical text, which is most often independent of irrelevant demographic information, and QAGNN is additionally grounded by a KG that does not contain any demographic representations. Unfortunately, we find that both systems change many of their answers when provided with irrelevant patient demographic information. We also observe that the two systems differ in which demographic information affects them. Finally, we compare biomedical to generic systems (i.e., trained on generic English text) and find that, as expected, the generic system changes even more of its answers in most cases (up to 17% for gender). However, for some demographics, such as sexual orientation, the biomedical system changes up to 23% of its answers. We hope that shedding light on this problematic behavior will motivate future work to further investigate its impact as well as possible solutions.

2 Related Work

Medical QA Many medical QA datasets are drawn from a variety of medical settings. MLEC-QA (Li et al., 2021) for example, is based on the National Medical Licensing Examination in China, while emrQA (Pampari et al., 2018) is based on clinical notes. HEAD-QA (Vilares and Gómez-Rodríguez, 2019) uses exams from the Spanish healthcare system, and MedQuAD (Abacha and Demner-Fushman, 2019) is based on 12 NIH websites and has questions on drugs, diseases, and other medical entities. DiSCQ (Lehman et al., 2022) has questions from MIMIC-III discharge summaries that were generated by medical experts, and the Q-Pain dataset (Logé et al., 2021) focuses on pain management. MedQA (Jin et al., 2021) has questions from the professional medical board exams and covers three languages. Recent datasets focus on specific challenges identified from previous efforts (Niu et al., 2003; Kell et al., 2021). We selected English language questions from MedQA for this study, based on the breadth and depth of medical knowledge required and the fact that students must pass an exam with similar questions to become a physician in the US.

Biases in NLP Models Social biases have been reported in widely divergent NLP training sets and models, ranging from gender bias in machine translation (Cho et al., 2019) to racial bias in opioid misuse prediction (Thompson et al., 2021). Social biases in dialog systems were examined through the use of demographically indicative names (Smith and Williams, 2021). Several studies of natural language generation systems, transformers, and related models have shown outputs influenced by a variety of demographic characteristics in prompts, e.g., (Sheng et al., 2019). Methods to measure stereotype bias in language models (LMs) have been proposed, such as StereoSet (Nadeem et al., 2021) and the CrowS-Pairs dataset (Nangia et al., 2020) which contains information on nine types of demographics, such as age and race.

Bias in Medical NLP Some also focus particularly on evaluating biases in the medical domain. One work (Borgese et al., 2021) analyzes unhealthy alcohol use risk bias between classifiers on electronic health records in trauma patients, while another (Logé et al., 2021) examines gender and ethnicity biases in a pain management setting between GPT-2 and GPT-3. However, getting unbiased data to investigate model bias in a controlled way is difficult for pain management, where there is extreme societal bias. Hence, we use other data in our study. Racial biases in clinical settings were also examined (Huang et al., 2022). Some also focus on using NLP to evaluate whether medical licensure exams contain language patterns that exhibit biased or stereotypical language (Padhee et al., 2021).

Lastly, there is also work on evaluating pretrained transformer models and examining whether they contain biased information towards different demographics (Zhang et al., 2020). In contrast to prior work, we 1) examine the effect irrelevant demographic information has on QA systems in a clinical setting, using questions which require

Dimensionless	A 23-year-old **patient** presents to a psychiatrist for evaluation of situational anxiety. The **patient** reports that **they** recently started a new job and is very stressed.
Ethnicity	A 23-year-old **Black patient** presents to a psychiatrist for evaluation of situational anxiety. The **patient** reports that **they** recently started a new job and is very stressed.
Gender	A 23-year-old **female** presents to a psychiatrist for evaluation of situational anxiety. **She** reports that **she** recently started a new job and is very stressed.
Names	A 23-year-old **patient named Tom** presents to a psychiatrist for evaluation of situational anxiety. The **patient** reports that **they** recently started a new job and is very stressed.
SOr	A 23-year-old **bisexual patient** presents to a psychiatrist for evaluation of situational anxiety. The **patient** reports that **they** recently started a new job and is very stressed.
SOr+Gender	A 23-year-old **bisexual female** presents to a psychiatrist for evaluation of situational anxiety. **She** reports that **she** recently started a new job and is very stressed.
Ethnicity+Gender	A 23-year-old **Asian male** presents to a psychiatrist for evaluation of situational anxiety. **He** reports that **he** recently started a new job and is very stressed.
Ethnicity+Gender +Names	A 23-year-old **Hispanic female named Guadalupe** presents to a psychiatrist for evaluation of situational anxiety. **She** reports that **she** recently started a new job and is very stressed.

Table 1: Dimensions example. Given a question, for each dimension, we demographically-enhance the question by adding relevant words (e.g., Black, bisexual, named X) and changing its gender tokens in order to create multiple datasets for the specific dimension. SOr=sexual orientation.

broad medical knowledge and are used by US medical students; 2) focus on models that are trained on biomedical text; 3) compare the effect of KG grounding on biases in a transformer-based model; and 4) compare biomedically-trained systems to a generic one, trained on English text.

3 Experimental Setup

3.1 Motivation

Biomedical QA systems can be beneficial for both healthcare providers and patients for many reasons: 1) With traditional search engines, finding reliable medical information can take time and effort due to the vast amount of unfiltered content available online, while QA systems allow users to quickly find answers; 2) such systems can serve as powerful learning tools for students and residents seeking to deepen their understanding of complex medical topics; 3) in low-resource settings there may be limited access to qualified healthcare professionals, which leads to delayed or incorrect diagnoses that may worsen health outcomes over time. Fortunately, biomedical QA systems can bridge this gap and extend the reach of health services to vulnerable populations worldwide.

However, in order for such systems to be safely deployed, ensuring that they provide fair behavior towards patients is critical. For example, imagine that a White and an African-American patient present themselves with similar symptoms at a hospital and that none of their symptoms indicate a problem related to their ethnicity. If one was treated with the correct medication while the other received an incorrect one, this would be highly problematic. Thus, it is important to understand if current biomedical QA systems could result in such an outcome.

3.2 MedQA-USMLE

The MedQA-USMLE dataset (Jin et al., 2021) is an open-domain QA dataset, which covers three languages: English, traditional Chinese, and simplified Chinese. MedQA has medical questions which represent real-world scenarios and evaluate physicians on their clinical decision making skills. The questions are varied and require a significant understanding of medical concepts. Here, we choose to only use the English version, which is composed of 12,723 multiple-choice prompts taken from the professional medical board exams. Each prompt consists of *context* and *question*, e.g., *"An 18-year-old male presents to the emergency room smelling quite heavily of alcohol and is unconscious. A blood test reveals severe hypoglycemic and ketoacidemia. A previous medical history states that he does not have diabetes. The metabolism of ethanol in this patient's hepatocytes resulted in an increase of the [NADH]/[NAD+] ratio. Which of the following reaction is favored under this condition?"*. Each question comes with four answer choices. The options for the above example are: *Pyruvate to acetyl-CoA, Citrate to isocitrate, Oxaloacetate to malate*, and *Oxaloacetate to phosphoenolpyruvate*.

3.3 Question Selection

Some phenomena are more prevalent in certain populations, such as pregnancy (Sogancioglu et al., 2022) or prostate cancer. For other diagnoses, patient demographic information is irrelevant and

	Random	Gender		Ethnicity					SOr			Gender+Ethnicity									Gender+SOr					
		M	F	W	A-A	B	H	As	Hetero	Bi	Homo	M+W	M+A-A	M+B	M+H	M+As	F+W	F+A-A	F+B	F+H	F+As	M+Hetero	M+Bi	M+Homo	F+Hetero	F+Bi	F+Homo
QAGNN	2	6	**7**	6	**9**	7	6	6	**9**	**7**	**15**	**8**	**8**	**9**	**10**	**8**	**9**	**10**	9	**9**	**9**	8	6	11	**10**	8	9
BioLinkBert	2	6	6	6	8	7	**7**	**11**	6	6	14	6	7	5	7	6	8	8	9	8	8	**9**	**9**	**23**	8	**13**	**23**

Table 2: Percentage of questions with changed answers as compared to a question with no demographic information about the patient. *M*=male; *F*=female; *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian; *SOr*=sexual orientation; Random=Random change as described in Section 3.5.

		Gender		Ethnicity					SOr		
		M	F	W	A-A	B	H	As	Hetero	Bi	Homo
Correct → Incorrect	QAGNN	1	**3**	**4**	**4**	**4**	3	3	**4**	**3**	**6**
	BioLinkBert	1	2	2	3	3	3	**6**	2	1	5
Incorrect→Incorrect	QAGNN	2	1	2	2	**3**	**3**	3	**3**	3	6
	BioLinkBert	**4**	**2**	2	**4**	2	2	3	2	3	6
Incorrect→Correct	QAGNN	**3**	**3**	0	**3**	0	0	0	2	1	3
	BioLinkBert	1	2	**2**	1	**2**	**2**	**2**	2	**2**	3

Table 3: Percentage of answers that changed from from correct to incorrect, incorrect to incorrect, and incorrect to correct for each model. *M*=male; *F*=female; *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian; *SOr*=sexual orientation.

should accordingly not be taken into account. For our experiments we build a dataset consisting of **only questions whose answers do not depend on sex, ethnicity, or sexual orientation.** We do so by following Logé et al. (2021)'s approach and extract 100 vignettes, which are designed to allow for the inclusion of diverse ethnics and gender "profiles" in order to assess potential biases. These vignettes are verified by two medical experts to be demographics-independent, and after the demographics-enhancing process, which will be discussed in the next section, result in **16,700 questions overall**, which are used to evaluate the effect irrelevant demographic information has on QA systems.

3.4 Demographics-enhanced Dataset Creation

We experiment with the following types of modified questions: dimensionless (i.e., no demographic information), ethnicity, gender, names, sexual orientation, gender+ethnicity, gender+sexual orientation, and gender+ethnicity+names.

The reasoning for each chosen dimension are as follows: dimensionless shows no demographic information, and hence will be used as a baseline to compare how many of the answers change when we add irrelevant demographic information. Ethnicity, sexual orientation, and gender, while not always shown in medical text, are sometimes mentioned when the demographic information is relevant. Hence, we want to see if the models associate any medical conditions with them. We use two genders, but expect that our results will generalize to additional genders. As for names, these are clearly not medically relevant ever and are rarely shown in medical text. Hence, we choose them to see if there are unexpected differences in answers change.

Ethnicities include White, Black, African-American, Hispanic, and Asian. Genders include male and female. Names include the 10 names for each ethnicity from the Q-Pain dataset, which originated from the Harvard Dataverse's *Demographic aspects of first names* dataset (Tzioumis, 2018). And while "Black" and "African American" are largely synonymous, we want to see if they are different from the models' perspective. Notably, to medically-untrained users, all of these may seem relevant and hence potentially be added to queries when such users request medical assistance.

We follow a similar process as the creators of the Q-Pain dataset and make each context, question, and answer (CQA) as neutral as possible. Given a CQA, such as "A 23-year-old female presents to a psychiatrist...", we first automatically mask any word that indicates gender (e.g., male, female, he, she, wife, boyfriend): "A 23-year-old

	*O**	O	D	Gen.		Ethnicity					SOr			Gender+Ethnicity									Gender+SOr					
				M	F	W	A-A	B	H	As	Hetero	Bi	Homo	M+W	M+A-A	M+B	M+H	M+As	F+W	F+A-A	F+B	F+H	F+As	M+Hetero	M+Bi	M+Homo	F+Hetero	F+Bi	F+Homo
1	*38*	**40**	40	**42**	40	36	**39**	36	37	**37**	38	38	37	38	**38**	36	35	37	35	34	35	35	35	36	39	**38**	36	36	**37**
2	***40***	39	40	40	40	**40**	38	**39**	**39**	36	**40**	**41**	**38**	**39**	37	**41**	**40**	**40**	**40**	**36**	**40**	**41**	**40**	**41**	**41**	36	**41**	**40**	36

Table 4: Accuracy (in percentages) of the two models on our demographically enhanced datasets. *M*=male; *F*=female; *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian; *SOr*=sexual orientation; O*=original test dataset; O=the original, unmodified 100 vignettes; D=No demographic information; Gen=Gender; 1=QAGNN; 2=BioLinkBERT.

[GENDER_MASK] presents to a psychiatrist...". Then, given a dimension (e.g., gender), we automatically replace each unique masking with their corresponding token replacement (e.g., replacing "[GENDER_MASK]" with "male").

Overall, each of these dimensions and their variations augment each of the 100 vignettes and result in overall 16,700 questions. See Table 1 for examples. And while we only use the English version of the dataset, this process can be easily applied to other languages. The data will be publicly available and have an MIT License.

3.5 Random Change

We use a version of the questions with no demographic information, and, in each prompt's first sentence, replace the word "patient" with "person". With this we examine the effect of a small but insignificant textual variation on each model. We choose this change over others (e.g., adding random words, irrelevant demographics, or fictitious cities) as this reduces the possibility of models changing their answers due to the context such random words had in the training data (e.g., Africa is more prevalent to the sleeping sickness disease than the US). Moreover, neither "person" nor "patient" reveal information about the human.

4 Models

We compare two existing algorithms: QAGNN (Yasunaga et al., 2021) and BioLinkBert (Yasunaga et al., 2022). While better models exist for the USMLE dataset, many of them have billions of parameters and we are unable to test them for computational reasons. That being said, BioLinkBert is currently among the state of the art on the USMLE dataset, and QAGNN is the top (and, to the best of our knowledge, only) KG-grounded model. We use existing implementations and models and describe both systems in the following.

4.1 QAGNN

The main component of QAGNN is its KG, which is based on the Disease Database portion of the Unified Medical Language System (UMLS) and DrugBank. The graph contains about 10k nodes and 44k edges, where the embeddings for each node are initialized using the biomedically trained language model SapBERT (Liu et al., 2020). SapBERT was trained using the UMLS vocabulary set 2020AA version, which contains biomedical synonyms from more than 150 controlled vocabularies, such as Gene Ontology and MeSH. QAGNN has 360M parameters.

For each answer choice of a given question, QAGNN first retrieves a subgraph from its KG using entity linking. That is, it finds entity mentions in the question and retrieves any entity in the main KG that appears in any 2-hop paths between pairs of found entities. Then, it concatenates the answer choice and question, followed by encoding using a LM. Next, it connects the encoded representation to the graph as a node. It then performs relevance scoring on each node in the created subgraph by concatenating it to the encoded representation node and calculating the likelihood using a LM. Lastly, using an attention-based graph neural network (GNN) module, it reasons over the graph to get a score for the answer choice. During the training procedure, it optimizes both the LM and its GNN end-to-end using cross-entropy loss. On the MedQA-USMLE dataset, SapBERT-based QAGNN achieves 38% accuracy.

4.2 BioLinkBert

The defining features of BioLinkBert are its pretraining method that incorporates document links and its LM which has similar hyperparameters to PubmedBERT (Gu et al., 2020) and is trained from

	Names				Gender+Ethnicity+Names									
	W	**A-A/B**	**H**	**As**	**M+W**	**M+A-A**	**M+B**	**M+H**	**M+As**	**F+W**	**F+A-A**	**F+B**	**F+H**	**F+As**
QAGNN	**10.5**	**10.5**	**12.6**	**10.5**	**9.3**	**14.2**	**11.5**	**9.8**	8.5	**9.3**	**15.0**	**11.5**	**12.5**	7.9
BioLinkBERT	7.4	6.0	8.5	6.0	8.8	11.9	9.8	8.1	**9.6**	8.5	11.5	10.3	10.0	**9.0**

Table 5: Percentage of questions with changed answers as compared to a question with no demographic information about the patient. *M*=male; *F*=female; *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian; *SOr*=sexual orientation.

Model	**Names**									
	W		**B**		**A-A**		**H**		**AS**	
QAGNN	**38.6**		**39.5**		**39.5**		**39.3**		**38.5**	
BioLinkBert	38.2		37.3		37.3		37.6		37.6	
	M	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
QAGNN	38.6	38.1	**39.5**	**39.4**	**39.7**	**39.1**	**39.0**	**39.2**	**39.2**	**37.9**
BioLinkBert	**39.4**	**38.5**	38.6	37.0	35.1	38.7	36.8	37.2	37.3	35.4

Table 6: Accuracy when including names (rows 1 and 2) or names together with gender and ethnicity information (rows 3 and 4) for each model. *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian;

scratch on the PubMed abstracts PubmedBERT is trained on. BioLinkBert has 340M parameters.

Given a corpus of text, BioLinkBert views it as a graph: it uses Pubmed Parser to extract citation links between documents and views the hyperlinks as edges. Then, to use the links in its LM pretraining procedure it places two documents which share a link in the same context, in addition to placing two random documents in the same context or a single document (contiguous). Next, it uses two self-supervised objectives. The first, masked language modeling, is common in many of the large LMs such as BERT (Devlin et al., 2019). In the second, document relation prediction, it classifies the link between the two documents as random, linked, or contiguous. On the MedQA-USMLE dataset, the base version of BioLinkBert achieves 40% accuracy while the large version achieves 44.6%. Here, we work with the base version because of its lower compute requirements.

5 Results

We look at two different effects of providing the model with irrelevant demographic information: 1) the percentage of questions for each model that change and 2) the accuracy change for each model. Note that these are not necessarily correlated: for example, accuracy does not change when initially incorrect answers change to other incorrect answers, or if the same numbers of answers change from incorrect to correct. It is also worth mentioning that **any change in model's answers is problematic, as these questions were verified to be independent of demographics.**

5.1 Changed Answers

Table 2 shows the percentage of questions for each model that change between each dimension's attribute and the dimensionless variation (e.g., between male and genderless).

The first column of Table 2, "Random", shows the result of our random change (Sec. 3.5). While the other values in the table are larger, and while the words "patient" and "person" may have different connotations for each model based on its training data, this suggests that, to some extent, random noise plays a role in the amount of change each model exhibits. Notably for gender, ethnicity, and sexual orientation, both models change around the same number of answers, except that BioLinkBert has a much higher number for Asian. Additionally, both models have almost double the amount of changed answers for homosexual than bisexual or heterosexual. For gender+ethnicity, QAGNN has an equivalent amount or more than BioLinkBert, though for gender+sexual orientation, BioLinkBert has more than double the amount for homosexuals, with a massive percentage of 23. We also examine the amount of answers for each model for gender, ethnicity, and sexual orientation, that change from

	Random	Gender		Ethnicity					SOr			Gender+Ethnicity										Gender+SOr					
		M	F	W	A-A	B	H	As	Hetero	Bi	Homo	M+W	M+A-A	M+B	M+H	M+As	F+W	F+A-A	F+B	F+H	F+As	M+Hetero	M+Bi	M+Homo	F+Hetero	F+Bi	F+Homo
Generic	2	**17**	**16**	6	**14**	7	**9**	7	**9**	**11**	11	**11**	**11**	**12**	**13**	**8**	**13**	**13**	**13**	**12**	**12**	8	**10**	6	**12**	**15**	11
Biomedical	2	6	6	6	8	7	7	**11**	6	6	**14**	6	7	5	7	6	8	8	9	8	8	**9**	9	**23**	8	13	**23**

Table 7: Percentage of questions with changed answers between the biomedical and generic model as compared to a question with no demographic information about the patient. *M*=male; *F*=female; *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian; *SOr*=sexual orientation.

	O*	**O**	**D**	**Gender**		**Ethnicity**					**SOr**			**Gender+Ethnicity**										**Gender+SOr**					
				M	F	W	A-A	B	H	As	Hetero	Bi	Homo	M+W	M+A-A	M+B	M+H	M+As	F+W	F+A-A	F+B	F+H	F+As	M+Hetero	M+Bi	M+Homo	F+Hetero	F+Bi	F+Homo
1	*28.9*	26	25	29	27	27	26	26	27	27	28	27	26	27	27	27	26	25	27	27	27	26	24	27	27	28	25	26	25
2	***40***	**39**	**40**	**40**	**40**	**40**	**38**	**39**	**39**	**36**	**38**	**38**	**37**	**39**	**37**	**41**	**40**	**40**	**40**	**36**	**40**	**41**	**40**	**41**	**41**	**36**	**41**	**40**	**36**

Table 8: Accuracy (in percentages) of the biomedical and generic models on our demographically enhanced datasets. *M*=male; *F*=female; *W*=White; *B*=Black; *A-A*=African-American; *H*=Hispanic; *As*=Asian; *SOr*=sexual orientation; O*=original test dataset; O=the original, unmodified 100 questions; D=No demographic information; 1=Generic; 2=Biomedical.

being correct to incorrect, from incorrect to correct, and from incorrect to incorrect (Table 3). We can see that a model can have an increase in performance (see QAGNN males column which results in a 2% increase) while having the same number of answers change as a demographics which result in a decrease in performance (see QAGNN White column which result in a 4% decrease). This implies that accuracy alone is not sufficient to understand the effect irrelevant demographic information has on models' answer, and that further examination of the answers can contribute. For example, we see that adding most ethnicities results in 0 answers changing from incorrect to correct for QAGNN.

5.2 Changed Accuracy

While the reported accuracy on the original test dataset is 38% for QAGNN and 40% for BioLinkBert, the accuracy on our 100 randomly selected demographic-independent questions use to construct the vignettes is 40% for QAGNN and 39% for BioLinkBert. Table 4 shows our accuracy results for each dimension for each algorithm.

As noted, accuracy change does not always correlate with answer change. For example from Table 2, while both models have about the same number of changed answers for gender, only QAGNN's accuracy for males is affected (increased by 2%). For ethnicity, both models' accuracy drops, with BioLinkBert's accuracy by 3% for Asian and QAGNN's accuracy by 4% for Black. Sexual orientation improves BioLinkBert performance on bisexual and decreases QAGNN's on every variation. Gender+ethnicity decreases QAGNN performance the most (up to 6%), while gender+sexual orientation improves BioLinkBert's performance on any variation except for homosexual.

6 Analysis: Names

Similarly to the above experiments, we also evaluate the effect names have on the two types of models. For names by themselves, for each ethnicity (Black, White, Hispanic, Asian) we use the corresponding 20 names (10 for males and 10 for females). For names+ethnicity+gender, we split the names into their ethnicity and gender.

Table 5 and 6 show our results: Tables 5 displays the number of changed answers, while Table 6 shows accuracy changes. We can see that names alone have a moderate effect on the performance of both models, decreasing the performance in any variation by up to 1.65%. From our baseline experiment this may be due to random noise. However, by looking at the number of changed answers, we can see that both models have the most change for Hispanics, with QAGNN change of up to 12.6% and BioLinkBert by up to 8.5%. Interestingly, QAGNN has the same number of changed answers for White,

Black, and Asian, but a different number for Hispanic. More results can be seen in the combination of gender, ethnicity, and names, in which the performance can decrease by up to 3.9% for BioLinkBert in African American males, and by up to 2.1% for QAGNN in Asian females. However, the amount of changed answers is up to 15% in QAGNN for African American females and up to 11.9% for BioLinkBert in African American males. This implies that even though both models were trained on PubMed data, irrelevant information like names affect them, which is highly problematic.

7 Medical vs. Generic LMs

In addition to our main results, we also compare how the performance of a biomedically-trained transformer differs from that of a generic one. In particular, we use the same code for the BioLinkBert QA system, but instead of using the medically-trained base (trained from scratch on PubMed abstracts), we use a transformer which is trained on generic English text.

Similar to our analysis between QAGNN and BioLinkBert above, our analysis between the biomedical and generic models can be split into the amount of answers and accuracy that changes when the dimensions change. From Table 7 it is visible that the generic transformer has more than double the amount of answers change for each gender. It also has an equivalent amount or more for almost any ethnicity, except for Asians. Notably, for sexual orientation, the generic transformer has almost double the amount of answers change for bisexuals, while the biomedical transformer has more for homosexuals. The generic transformer has significantly larger values than the biomedical transformer in any gender+ethnicity combination, while for gender+sexual orientation, the biomedical system has significantly larger values for homosexuals. From Table 8 it is clear that BioLinkBert significantly outperforms its generic LM variation. From the change in accuracy we can see that, while the biomedical transformer's accuracy increases when gender is removed ("no info"), the generic transformer's accuracy decreases. We can also see that the biomedical transformer's accuracy changes more for ethnicity and sexual orientation, while the generic model changes more for gender.

8 Future Work

Finally, we discuss three potential approaches to alleviate the aforementioned effects, including model architectures, data, and regularization.

Model Architecture While both the KG-grounded LM and the text-based one are susceptible to irrelevant demographic information, our initial assumption that the KG-based LM would be less susceptible still holds. In particular, KGs are a condensed representation of knowledge, which rarely holds such irrelevant information. Hence, models that use such representations have a significant potential to be less affected. That being said, a potential reason that the tested KG-based LM is still susceptible may be due to the fact that it grounds the text using the KG, and does not only uses the KG. Hence, the demographically irrelevant information may still leak into the final representation, which the model uses to answer the question.

Data Generally, large LMs are trained using a massive corpus. This is problematic as it is almost impossible to ensure that every piece of data is demographically independent. To try to alleviate this issue, we select biomedical models that are trained only on biomedical data, which often does not contain demographically irrelevant information. However, we still find that these models are susceptible to such information. Hence, future work should examine methods to reduce such issues in the training data, especially for models intended to critical settings.

Regularization While developing models with different architectures or ensuring that every piece of data is demographically independent is time consuming, a potentially simple method to alleviate such problem is to regularize the input itself. For example, by masking demographically-significant words. And while relatively simple to implement (e.g., using keywords search), in medicine it is sometimes the case where such demographically-significant words are in fact significant. Hence, simple masking might reduce bias, but will also reduce performance. Future work should examine potential masking approaches that consider times where such words are actually needed.

9 Conclusion

We examine the effect of irrelevant demographic information on purely text-based and KG-grounded

biomedical QA systems as well as a generic QA system, using a subset of the USMLE questions whose answers do not depend on the patient's demographics. Our results show that irrelevant demographic information results in changed answers for all systems. We also find that, while all systems are affected by irrelevant demographic information, they differ with regards to how different types of demographic information influence them. These results provide evidence that more work is needed in order to ensure fair treatment of all patients by biomedical QA systems.

Limitations

As expert annotation is expensive, we only use 100 unique vignettes to create the 16,700 questions. However, this is almost twice as many as other published datasets, such as Logé et al. (2021). Additionally, we only analyze one KG-grounded and one purely text-based system. While our main point, that there are problems one should be aware of, can be made based on experiments with two models, evaluating more systems can potentially lead to more fine-grained insights.

Ethics Statement

The main reason for this paper is to point out potential problems regarding fair treatment of all patients by biomedical QA systems. Future work should improve existing biomedical QA systems to ensure equal and just patient care. Moreover, such systems can be problematic for both patients and health experts. For example, a patient could follow the recommendations of such a QA model at home without expert supervision and a system could recommend an incorrect treatment because of their name, or physicians could use such systems to improve their quality of care, but the system could cloud their judgment and direct them to an incorrect answer.

Acknowledgments

We thank Dr. Peter Pressman for his help with reviewing the data and the reviewers for their feedback. The authors acknowledge financial support from NIH grants OT2TR003422 and R01LM013400.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1).

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*, pages 5454–5476, Online. Association for Computational Linguistics.

Marissa Borgese, Cara Joyce, Emily E Anderson, Matthew M Churpek, and Majid Afshar. 2021. Bias assessment and correction in machine learning algorithms: A use-case in a natural language processing algorithm to identify hospitalized patients with unhealthy alcohol use. *AMIA Annu. Symp. Proc.*, 2021:247–254.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Jonathan Huang, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. 2022. Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Med Inform*, 10(5):e36388.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Gregory Kell, Iain Marshall, Byron Wallace, and Andre Jaun. 2021. What would it take to get biomedical QA systems into practice? In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 28–41, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eric Lehman, Vladislav Lialin, Katelyn Edelwina Legaspi, Anne Janelle Sy, Patricia Therese Pile, Nicole Rose Alberto, Richard Raymund Ragasa,

Corinna Victoria Puyat, Marianne Katharina Taliño, Isabelle Rose Alberto, Pia Gabrielle Alfonso, Dana Moukheiber, Byron Wallace, Anna Rumshisky, Jennifer Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 74–86, Seattle, WA. Association for Computational Linguistics.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations.

Cécile Logé, Emily Ross, David Yaw Amoah Dadey, Saahil Jain, Adriel Saporta, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Q-pain: A question answering dataset to measure social bias in pain management.

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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. 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)*, pages 1953–1967, Online. Association for Computational Linguistics.

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clinical questions with role identification. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 73–80, Sapporo, Japan. Association for Computational Linguistics.

Swati Padhee, Kimberly Swygert, and Ian Micir. 2021. Exploring language patterns in a medical licensure exam item bank.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Eric Michael Smith and Adina Williams. 2021. Hi, my name is martha: Using names to measure and mitigate bias in generative dialogue models.

Gizem Sogancioglu, Fabian Mijsters, Amar van Uden, and Jelle Peperzak. 2022. Bias in (non)-contextual clinical word embeddings.

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

Konstantinos Tzioumis. 2018. Data for: Demographic aspects of first names.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Sina Zarrieß, Hannes Groener, Torgrim Solstad, and Oliver Bott. 2022. This isn't the bias you're looking for: Implicit causality, names and gender in German language models. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 129–134, Potsdam, Germany. KONVENS 2022 Organizers.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: Quantifying biases in clinical contextual word embeddings.