Report for Assignment 2

Weijia Sun

Section 1

2(a) What is the distribution of the cohort (in percentages) for gender, ethnicity, language, and insurance?

--Distribution of gender--
M 56.9%
F 43.1%

--Distribution of ethnicity--
white 70.1%
other 16.9%
black 7.6%
hispanic 3.2%
asian 2.3%

--Distribution of language--
English 51.2%
Missing 40.5%
Other 8.3%

--Distribution of insurance--
Medicare 54.0%
Private 33.7%
Medicaid 8.3%
Government 2.9%
Self Pay 1.1%

2(b) What is the distribution of the cohort (in percentages) for the intersection of gender and ethnicity?

M_white 40.2%
F_white 29.9%
M_other 10.0%
F_other 6.9%
F_black 4.2%

M_black 3.4%
M_hispanic 1.9%
M_asian 1.4%
F_hispanic 1.3%
F_asian 0.9%

2(c) Explain, using one sentence each, what the following terms mean in an insurance context: Medicare, private, Medicaid, self-pay.

Medicare: It primarily provides health insurance for Americans aged 65 and older, but also for some younger people with disability status as determined by the Social Security Administration, as well as people with end stage renal disease and amyotrophic lateral sclerosis (ALS or Lou Gehrig's disease).

Private: Any health insurance plan that is not run by the federal or state government. Private insurance can be purchased from a variety of sources: your employer, a state or federal marketplace, or a private marketplace.

Medicaid: Medicaid in the United States is a federal and state program that helps with medical costs for some people with limited income and resources. Medicaid also offers benefits not normally covered by Medicare, including nursing home care and personal care services.

Self-pay: Those balances due from patients/guarantors for hospital and physician services as a result of having no insurance or having a balance due even after insurance pays, due to coinsurance, deductibles, or noncovered services.

3(b)
Explain why the parity gap is almost never a useful metric in a healthcare context.

Parity gap is more suitable for group fairness instead of individual fairness, since it equalizes outcomes across protected and non-protected groups.

What assumption would have to hold for the parity gap to become useful?

The protected and non-protected groups have similar property. (i.e., any two individuals who are similar with respect to *a particular task* should be classified similarly.)

3(c) Explain why individual fairness might be tricky to define in a healthcare setting.

Because the personal difference could be huge. The assumption of many healthcare work is “Treating similar individuals similarly”, but in reality this assumption is too ideal.

4(b)
Report the average length (in characters) of the notes for men versus women.

- The average length (in characters) of the notes is (for all women): 8823.981253938247
- The average length (in characters) of the notes is (for all men): 8662.751617076327

Use an appropriate statistical test to determine if there is a significant difference.

T-test is used here: A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.
p-value=0.2613002584853309
Traditionally, the cutoff of p-value is 5% (i.e., 0.05), so there is no significant difference.

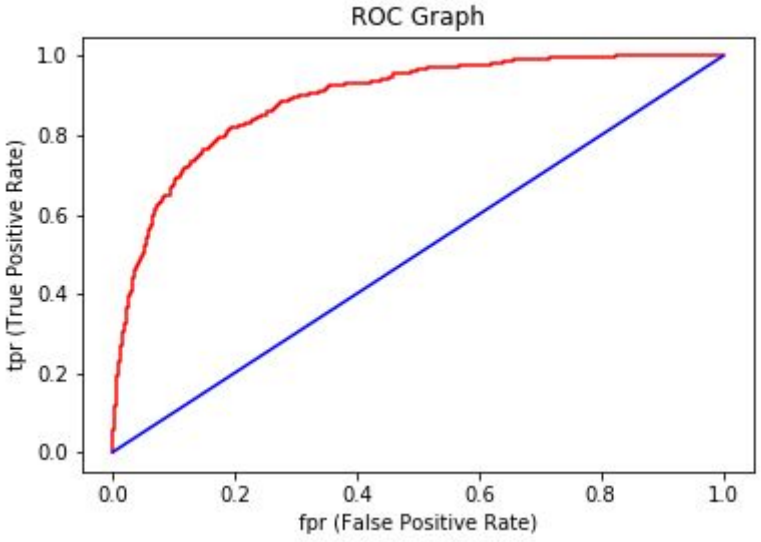
4(c) Report the prevalence of mort icu for men versus women.

Total number of women is: 6348
Total number of mortality in ICU for women is: 633
The prevalence of mort icu for women is: 0.0997164461247637 (i.e., 9.97%)

Total number of men is: 8503
Total number of mortality in ICU for men is: 772
The prevalence of mort icu for men is: 0.09079148535810891(i.e., 9.08%)

4(d) Report your model AUROC and AUPRC on the test set.

Area Under PR Curve(i.e., AUPRC): 0.5094244560801303
AUC score (i.e., AUROC): 0.8905556059095441



4(e)

Report each of the following (with 95% CIs): parity gap, specificity gap, recall gap.

For the six gaps, the mean value as follows:
{'mean of parity_M': -0.0022550272182548805,
'mean of parity_F': 0.0022550272182548805,
'mean of specificity_M': 0.00489765562416902,
'mean of specificity_F': -0.00489765562416902,
'mean of recall_M': 0.029900987977574473,
'mean of recall_F': -0.029900987977574473}

Standard Deviation of ls_parity_M is 0.007533554092300888
Standard Deviation of ls_parity_F is 0.007533554092300888
Standard Deviation of ls_specificity_M is 0.005735528599101285
Standard Deviation of ls_specificity_F is 0.005735528599101285
Standard Deviation of ls_recall_M is 0.05018580464716909
Standard Deviation of ls_recall_F is 0.05018580464716909

For recall gap, t_val, p_val: 26.64524049472698 4.222312706924797e-134
For parity gap, t_val, p_val: -13.386494845197644 3.3596400529847585e-39
For specificity gap, t_val, p_val: 38.18825315505221 4.522522407492755e-240

Use the convention that a positive gap denotes better performance in males.Looking at the recall gap, which gender does it favor? Is this gap statistically significant?

t_val, p_val: 26.64524049472698 4.222312706924797e-134
Male. According to p-value (less than 0.05), there is a statistically significant.

Section2

2(b) Come up with a set of clinically relevant sentences where ClinicalBERT displays biases towards a protected group.

races = ['caucasian', 'asian', 'hispanic', 'black']
text = '[CLS] this is an 81-year-old male [RACE] with a history of _ (not on home o2), who presents with three days of shortness of breath. [SEP]'

[CLS] this is an 81-year-old male caucasian with a history of diabetes (not on home o2), who presents with three days of shortness of breath. [SEP]
[CLS] this is an 81-year-old male asian with a history of falls (not on home o2), who presents with three days of shortness of breath. [SEP]
[CLS] this is an 81-year-old male hispanic with a history of falls (not on home o2), who presents with three days of shortness of breath. [SEP]
[CLS] this is an 81-year-old male black with a history of diabetes (not on home o2), who presents with three days of shortness of breath. [SEP]

3(b) Complete the log cats variable by filling in clinical terms for the category related to mental health. Run the provided script to generate a summary table. Include this table in your report. Briefly interpret the results.

	Category	Male log-prob	Female log-prob	p-value
0	cvd	1.079810	0.962445	0.000002
1	hypertension	0.310041	0.263864	0.094717
2	diabetes	0.740889	0.658606	0.037215
3	mental health	0.320344	0.405167	0.196981

If male log-prob is larger, it means male has more probability to associate with the masked word (.i.e., the model prefers to make associations with males instead of females).

3(c) Does having a significant difference between the genders necessarily imply the presence of “bad” bias? Explain your answer.

Not really. Some reasons for not bad bias:

- Sampling errors could occur

- The number of two groups differs a lot, so the selected samples can't represent the general population.

A statistic is biased if it is calculated in such a way that it is systematically different from the population parameter being estimated. P-value can tell us the probability of this data being observed given that the null hypothesis is true (the null hypothesis here being "no bias at all"), and we're allowed to conclude that the null hypothesis is incorrect if the probability of the data given the null hypothesis is less than 5%.

4(a) Given that machine learning classifiers are most likely to be used as diagnostic systems, explain why the TPR gap might be a more useful fairness metric than the specificity gap.

From the defination, we know that:

- **Specificity** is the correctly labeled by the model to all who are healthy in reality. Specificity answers the following question: Of all the people who are healthy, how many of those did we correctly predict? (Specificity = TN/(TN+FP))
- **Recall** (aka Sensitivity and TPR) is the ratio of the correctly labeled by our model to all who are not healthy in reality. Recall answers the following question: Of all the people who are not healthy, how many of those we correctly predict? (Recall = TP/(TP+FN))

The point of using machine learning emphasises more on those who are not healthy (i.e. we want to predict the (potential) diseases for people). In this case, TPR gap is more useful.

5(a) Report the value requested in the notebook to gain the point for this question.

>>> print(inputs[35, 10, 1234])
>>> tensor(-0.0743)

5(b) After building the model, run the included code to generate a table summarizing your model performance per task. Include this table in your report.

Mean AUC: 0.769

	Task	AUROC	logloss	AUPRC
0	Acute Renal	0.804	0.457	0.566
1	Cerebrovascular	0.889	0.220	0.572
2	Myocardial	0.821	0.320	0.439
3	Dysrhythmias	0.707	0.592	0.550
4	Chronic Kidney	0.777	0.296	0.290
5	COPD	0.695	0.361	0.246
6	Comp. Surgical	0.721	0.552	0.521
7	Conduction	0.743	0.210	0.165
8	Heart Failure	0.794	0.481	0.614
9	Atherosclerosis	0.853	0.446	0.755
10	Diabetes Comp	0.719	0.267	0.186
11	Diabetes No Comp	0.640	0.471	0.277
12	Lipid Metabolism	0.737	0.518	0.514
13	Hypertension	0.679	0.640	0.600
14	Fluid Disorder	0.729	0.556	0.550
15	GI Hemorrhage	0.773	0.235	0.248
16	Hypertension Comp	0.765	0.298	0.279
17	Other Liver	0.804	0.295	0.391
18	Lower Resp	0.708	0.227	0.152
19	Upper Resp	0.751	0.156	0.180
20	Pleurisy	0.667	0.358	0.217
21	Pneumonia	0.796	0.405	0.451
22	Resp Failure	0.864	0.433	0.760
23	Septicemia	0.831	0.362	0.518
24	Shock	0.824	0.283	0.380
25	Any Acute	0.837	0.287	0.971
26	Any Chronic	0.826	0.386	0.948
27	Mean	0.769	0.375	0.457

5(c) Report each of the five tables generated.

Note: I met the problem of running the given cell. I didn't change the code but the table is so weird (i.e., %Pop make sense, but other items are strange).

	# Significant	# Favored	% Favored	% Pop
Medicaid	0	0	nan%	8.34%
Medicare	0	0	nan%	53.97%
Government	0	0	nan%	2.90%
Private	0	0	nan%	33.66%
Self Pay	0	0	nan%	1.13%

ethnicity

	# Significant	# Favored	% Favored	% Pop
white	0	0	nan%	70.05%
hispanic	0	0	nan%	3.15%
other	0	0	nan%	16.88%
black	0	0	nan%	7.61%
asian	0	0	nan%	2.30%

gender

	# Significant	# Favored	% Favored	% Pop
F	0	0	nan%	43.14%
M	0	0	nan%	56.86%

language

	# Significant	# Favored	% Favored	% Pop
English	0	0	nan%	51.22%
Other	0	0	nan%	8.31%
Missing	0	0	nan%	40.47%

intersection

	# Significant	# Favored	% Favored	% Pop
F asian	1	0	0.00%	0.26%
F white	0	0	nan%	9.10%
M hispanic	0	0	nan%	0.56%
M other	0	0	nan%	2.89%
M white	0	0	nan%	12.03%
M black	0	0	nan%	1.09%
F other	0	0	nan%	1.92%
M asian	0	0	nan%	0.44%
F hispanic	0	0	nan%	0.37%
F black	0	0	nan%	1.16%

5(d) For each protected variable, interpret your results from the previous table

Which group(s) does the classifier favor?
Insurance: private, medicaid, medicare
Ethnicity: other, white, black
Gender: male
Language: English
Intersection: M white

Which groups(s) does the classifier disfavor?
Insurance: self pay, government
Ethnicity: asian, hispanic
Gender: female
Language: other
Intersection: F asian, M asian, F hipanic, F hispanic

5(e) For one of the groups that are disfavored, give a conjecture based on epidemiology or societal norms for why they might be disfavored.

For gender:

- Doctor's menstrual cycles and fluctuating hormones: scientists feared may limit the reliability and reproducibility of findings obtained in female subjects.
- Cost: Repeating the experiments in both sexes would require twice the resources.

For these reasons, experiments were often conducted in males, with the assumption that the findings would apply to females as well.

For asian:

The total number of asian in the United States is less than no matter white or black. Social status is not as important as other races.

5(f) Describe two sources of bias that might responsible for these performance gaps.

- Gender blindness and stereotyped preconceptions about men and women are identified as key causes to gender bias.
- However, exaggeration of observed sex and gender differences can also lead to bias.
- The medical training for doctors which is based on the average patient being a male may also be a source.
- Many research studies into diseases and treatments are skewed with a higher number of male participants.
- It's important that medical teams understand how certain health issues can present differently in men and women.

5(g) Describe a method that you could use to "debias" the classifier we just created (i.e. bring the TPR gaps closer to zero). If this is a published method, provide a reference.

Define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding.

Resource: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings ↩ (i.e., https://arxiv.org/abs/1607.06520)