Label Creation of Dataset

Missing Label

In the paper “COVID-19 vaccine hesitancy: Vaccination intention and attitudes of community health volunteers in Kenya”, they mentioned that the data collected for the research was based on a “WHO SAGE vaccine hesitancy matrix” form which is in 3 parts and this form can be easily downloaded using the hyperlink of the pdf paper. In the paper they mention statistics of COVID-19 vaccination intention among the volunteers, and this information can be seen in the third part of the questions that was asked to understand vaccine hesitancy, but when inspecting the dataset, we couldn’t find this data in the dataset. So, the labels are missing.

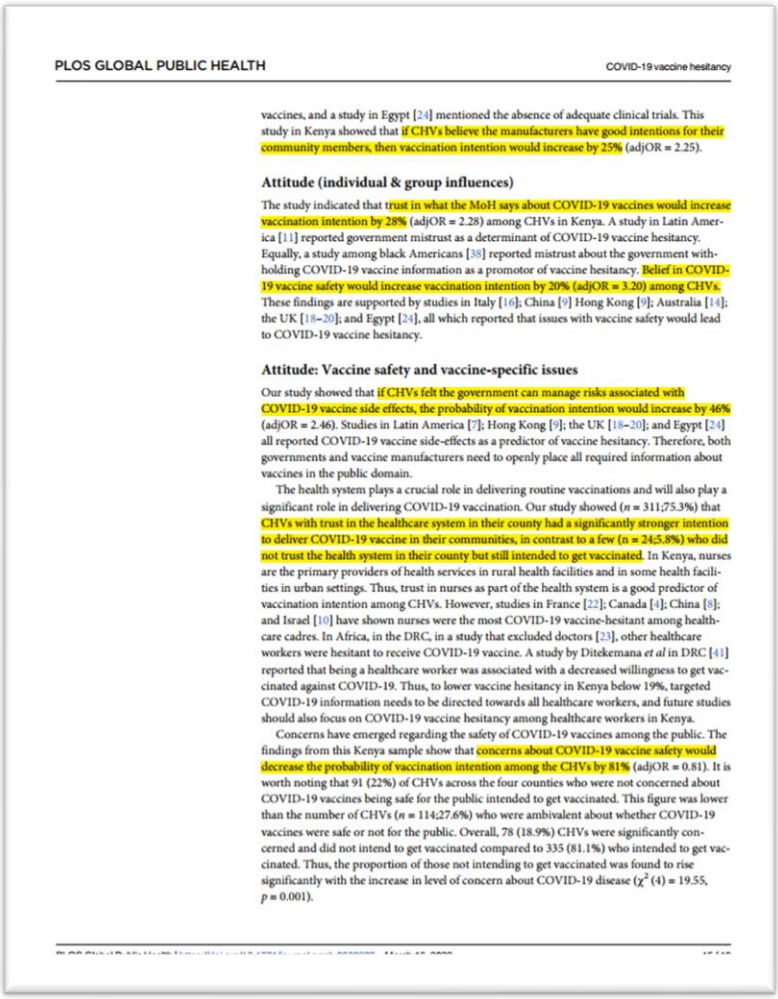
Missing values from dataset that could’ve been used as labels:

Vaccination Intention Questions

1. I am planning to get the COVID-19 vaccination once it’s available in the country
   Strongly Disagree (1): Disagree (2) : Neutral (3): Agree (4): Strongly Agree (5)
2. I intend to get the COVID-19 vaccination once it’s available in the country
   Strongly Disagree (1): Disagree (2) : Neutral (3): Agree (4): Strongly Agree (5)
3. I will try to get the COVID-19 vaccination once it’s available in the country
   Strongly Agree (1): Agree (2) : Neutral (3): Disagree (4): Strongly Disagree (5)

Useful Data from the Dataset

In the paper they mentioned a few information that corresponds to few columns/features in the dataset, that can influence vaccination intention significantly.



vaccines, and a study in Egypt [24] mentioned the absence of adequate clinical trials. This study in Kenya showed that if CHVs believe the manufacturers have good intentions for their community members, then vaccination intention would increase by 25% (adjOR = 2.25).

Attitude (individual & group influences)

The study indicated that trust in what the MoH says about COVID-19 vaccines would increase vaccination intention by 28% (adjOR = 2.28) among CHVs in Kenya. A study in Latin America [11] reported government mistrust as a determinant of COVID-19 vaccine hesitancy. Equally, a study among black Americans [38] reported mistrust about the government withholding COVID-19 vaccine information as a promotor of vaccine hesitancy. Belief in COVID-19 vaccine safety would increase vaccination intention by 20% (adjOR = 3.20) among CHVs. These findings are supported by studies in Italy [16]; China [9] Hong Kong [9]; Australia [14]; the UK [18–20]; and Egypt [24], all which reported that issues with vaccine safety would lead to COVID-19 vaccine hesitancy.

Attitude: Vaccine safety and vaccine-specific issues

Our study showed that if CHVs felt the government can manage risks associated with COVID-19 vaccine side effects, the probability of vaccination intention would increase by 46% (adjOR = 2.46). Studies in Latin America [7]; Hong Kong [9]; the UK [18–20]; and Egypt [24] all reported COVID-19 vaccine side-effects as a predictor of vaccine hesitancy. Therefore, both governments and vaccine manufacturers need to openly place all required information about vaccines in the public domain.

The health system plays a crucial role in delivering routine vaccinations and will also play a significant role in delivering COVID-19 vaccination. Our study showed (n = 311;75.3%) that CHVs with trust in the healthcare system in their county had a significantly stronger intention to deliver COVID-19 vaccine in their communities, in contrast to a few (n = 24;5.8%) who did not trust the health system in their county but still intended to get vaccinated. In Kenya, nurses are the primary providers of health services in rural health facilities and in some health facilities in urban settings. Thus, trust in nurses as part of the health system is a good predictor of vaccination intention among CHVs. However, studies in France [22]; Canada [4]; China [8]; and Israel [10] have shown nurses were the most COVID-19 vaccine-hesitant among healthcare cadres. In Africa, in the DRC, in a study that excluded doctors [23], other healthcare workers were hesitant to receive COVID-19 vaccine. A study by Ditekemana *et al* in DRC [41] reported that being a healthcare worker was associated with a decreased willingness to get vaccinated against COVID-19. Thus, to lower vaccine hesitancy in Kenya below 19%, targeted COVID-19 information needs to be directed towards all healthcare workers, and future studies should also focus on COVID-19 vaccine hesitancy among healthcare workers in Kenya.

Concerns have emerged regarding the safety of COVID-19 vaccines among the public. The findings from this Kenya sample show that concerns about COVID-19 vaccine safety would decrease the probability of vaccination intention among the CHVs by 81% (adjOR = 0.81). It is worth noting that 91 (22%) of CHVs across the four counties who were not concerned about COVID-19 vaccines being safe for the public intended to get vaccinated. This figure was lower than the number of CHVs (n = 114;27.6%) who were ambivalent about whether COVID-19 vaccines were safe or not for the public. Overall, 78 (18.9%) CHVs were significantly concerned and did not intend to get vaccinated compared to 335 (81.1%) who intended to get vaccinated. Thus, the proportion of those not intending to get vaccinated was found to rise significantly with the increase in level of concern about COVID-19 disease (χ^2 (4) = 19.55, p = 0.001).

Columns in dataset with the questions asked for the data in the column:

- acq6 - In your view is the MoH making the right decisions on COVID-19 vaccination?
- acq7 - Do you think vaccine manufacturers have good intentions for you and people in your community?
- aiq1 - Are you aware of any bad reactions in people who have had COVID-19 vaccination?
- aiq4 - Do you trust what the MoH says about COVID-19 vaccination?
- aiq5 - In your view is the COVID-19 vaccine safe enough for people to be injected?
- avq1 - Do you feel our country can manage risks associated with COVID-19 vaccine side effects?
- avq6 - In general, how safe do you think COVID-19 vaccine is for the general population?

These columns have few distinct values/answers, making it easy to encode and use.

Encoding

To use the values from the columns we need to encode them to numerical data. So we use sklearn's LabelEncoder.

Inspecting values in the selected columns:

```
Yes    370
No      43
Name: ACQ6: Attitude on Contextual Influences, dtype: int64
Yes    369
No      44
Name: ACQ7: Attitude on Contextual Influences, dtype: int64
No     370
Yes     43
Name: AIQ1: Attitude on Individual and Group Influences, dtype: int64
Yes    369
No      44
Name: AIQ4: Attitude on Individual and Group Influences, dtype: int64
Yes    342
No      71
Name: AIQ5: Attitude on Individual and Group Influences, dtype: int64
Yes    278
No     135
Name: AVSQ1: Attitude on Vaccine safety and vaccination specific issues, dtype: int64
Somewhat        176
Very much       101
Don't know       65
Not too much     58
Not at all       13
Name: AVSQ6: Attitude on Vaccine safety and vaccination specific issues, dtype: int64
```

Inspecting the values after encoding:

```
1    370
0     43
Name: ACQ6: Attitude on Contextual Influences, dtype: int64
1    369
0     44
Name: ACQ7: Attitude on Contextual Influences, dtype: int64
0    370
1     43
Name: AIQ1: Attitude on Individual and Group Influences, dtype: int64
1    369
0     44
Name: AIQ4: Attitude on Individual and Group Influences, dtype: int64
1    342
0     71
Name: AIQ5: Attitude on Individual and Group Influences, dtype: int64
1    278
0    135
Name: AVSQ1: Attitude on Vaccine safety and vaccination specific issues, dtype: int64
3    176
4    101
0     65
2     58
1     13
Name: AVSQ6: Attitude on Vaccine safety and vaccination specific issues, dtype: int64
```

After inspection we can match the encoded values to be either responding positively to vaccination or negatively (this is decided based on response to the questions recorded in the dataset)

```
acq6 - In your view is the MoH making the right decisions on COVID-19 vaccination?
positive = 1, negative = 0

acq7 - Do you think vaccine manufacturers have good intentions for you and people in your community
positive = 1, negative = 0

aiq1 - Are you aware of any bad reactions in people who have had COVID-19 vaccination?
positive = 0, negative = 1

aiq4 - Do you trust what the MoH says about COVID-19 vaccination
positive = 1, negative = 0

aiq5 - In your view is the COVID-19 vaccine safe enough for people to be injected
positive = 1, negative = 0

avq1 - Do you feel our country can manage risks associated with COVID-19 vaccine side effects?
positive = 1, negative = 0

avq6 - In general, how safe do you think COVID-19 vaccine is for the general population
values 0-4, positive=3,4, negative=1,2, neutral=0
```

Label Creation

From previous inspection we can see in 6/7 cases positive reaction towards vaccination is encoded as a higher value, and negative reaction is a lower value, so we can add these when finding average. The one value where positive is lower value and negative is higher value, we can subtract it when finding average. So, if we add them and create an average, we can understand the overall vaccine hesitation from the volunteers.

```
a = encoded_dataset["ACQ6: Attitude on Contextual Influences"]
b = encoded_dataset["ACQ7: Attitude on Contextual Influences"]
c = encoded_dataset["AIQ1: Attitude on Individual and Group Influences"]
d = encoded_dataset["AIQ4: Attitude on Individual and Group Influences"]
e = encoded_dataset["AIQ5: Attitude on Individual and Group Influences"]
f = encoded_dataset["AVSQ1: Attitude on Vaccine safety and vaccination specific issues"]
g = encoded_dataset["AVSQ6: Attitude on Vaccine safety and vaccination specific issues"]

# logic behind this formula: higher result value means more positive values in the row (based on the encodings), meaning less hesitancy
res = (a+b-c+d+e+f+g)/7
print(res.value_counts())
```

```
1.142857    129
1.285714     69
1.000000     65
0.714286     38
0.857143     38
0.571429     30
0.428571     24
0.285714     12
0.142857      4
0.000000      4
```

When doing a value count, we have a few distinct values, so now we need to find the threshold value that can best represent the percentage of volunteers who intent to get vaccinated.

```
tmp = []
for i in res:
    if i < 0.7:
        tmp.append(1)  # 1 means hesitant
    else:
        tmp.append(0)  # 0 means not hesitant

df_label = pd.DataFrame(tmp)

new_data["Label"] = df_label  # add the labels to new_data dataframe

df_label.value_counts()  # this should show 339 0s & 74 1s, i.e. 82% of users are not hesitant,
                         #almost in line with the paper which have 81%
```

```
0    339
1     74
```

By using 0.7 as the threshold from the average, we can split the values into a binary label to detect vaccine hesitation among the volunteers.

So now if we do (339/(339+74)), we get 0.8208, about 82% which is close to the value reported in the paper.

> COVID-19 vaccination rollout in Kenya. This cross-sectional study involved community health volunteers in four counties: Mombasa, Nairobi, Kajiado, and Trans-Nzoia, representing two urban and two rural counties, respectively. COVID-19 vaccination intention among community health volunteers was 81% (95% CI: 0.76–0.85). On individual binary logistic

This way we can say the label generated was close to the actual labels that was removed before publishing the dataset (for licensing or whatever reason it may be).