Analysis of fairness metrics for anonymization in electronic health records

Mariela Rajngewerc[*1,3], **Laura Ación**[2,3] **and Laura Alonso Alemany**[1]

[1] Sección de Computación, FAMAF, UNC, Argentina
[2] Instituto de Cálculo, UBA, Argentina
[3] CONICET, Argentina
[*] Email: marielaraj@gmail.com / Web: https://marielaraj.github.io/

In this work, **we show the strengths and limitations of different fairness metrics**, illustrating them as applied to the bias analysis of anonymization algorithms of electronic health records (EHR). We show how different fairness metrics highlight certain aspects of the behavior of these algorithms while obscuring others.

We need ethical insights

- Classical metrics to evaluate machine learning models are usually **aggregates**.
- Aggregate metrics provide **no insights** into the differential behavior of the model across subgroups (bias).
- The analysis of bias must consider the **characteristics of the problem** (unbalanced groups, differences in the amount and the density of sensitive information in each group).
- The impact of bias must be assessed to detect, mitigate, or even prevent possible **harm when working with human data**.

Errors in anonymization

Input	Expected outcome
María Gomez (DNI 12.345.678) acude a la consulta con dolor cervical. El 10/01 asistió a turno y fue evaluada por la Dra. Susana Lopez quien le recetó antiinflamatorio.	X X acude a la consulta con dolor cervical. El X asistió a turno y fue evaluada por la X quien le recetó antiinflamatorio.

Label 1 represents the class of interest, i.e., that a word is considered sensitive information.
Label 0 are words with no sensitive information.

Ŷ (predicted values) \ Y (actual values)	Y=1	Y=0
Ŷ=1	TP María Gomez presenta dolor cervical.	FP La paciente presenta dolor cervical.
Ŷ=0	FN María Gomez presenta dolor cervical.	TN La paciente presenta dolor cervical.

Characteristics of the problem:

- Most of the words in the EHR are not personal protected information.
- False Negatives (FN) expose sensitive words.
- The amount of sensitive information of certain groups can be much higher than that of others.

What do metrics focus on?

In the following, we discuss four well-known fairness metrics that, by definition, include an analysis of the distribution of errors over each group.

Treatment equality

$$\frac{FN_a}{FP_a} = \frac{FN_b}{FP_b}$$

✓ Both types of errors contribute to the final value of the metric.

✗ The actual number of accurate predictions of the class of interest (TP) is not considered in this metric.

⚠ The same value would be obtained for a group with 100 FP, 100 FN, and 100 TP or for a group with 100 FP, 100 FN and 0 TP, so, both groups have the same $\frac{FN}{FP}$ but in the latter case, all the sensitive information of one of the groups has been exposed.

Equal Opportunity

$$\frac{TP_a}{TP_a+FN_a} = \frac{TP_b}{TP_b+FN_b}$$

✓ All true sensitive information samples (Y=1) from both groups are considered.

✗ If the sensitive information samples from the groups are unbalanced, the same value for equal opportunity for both groups could be achieved, even when more sensitive information from one of the groups is revealed.

⚠ Suppose group A has a total of 2000 true sensitive information samples (Y=1) and group B has a total of 2 true sensitive information samples. In cases where $TN_a = 1000, TP_a = 1000, TN_b = 1, TP_b = 1$ both groups have equal opportunity, however, more sensitive information is being revealed for group A.

Equalized odds

$$\frac{FP_a}{FP_a+TN_a} = \frac{FP_b}{FP_b+TN_b} \quad \wedge \quad \frac{TP_a}{TP_a+FN_a} = \frac{TP_b}{TP_b+FN_b}$$

✓ The true sensitive information and true non-sensitive information are used in the comparison.

✗ Problems with unbalanced samples and the comparison related to the predictions from the non-sensitive information samples may not be captured.

⚠ When there are many more samples of true non-sensitive information (Y=0) than of sensitive information (Y=1) in both groups and the model has good accuracy, differences in the number of misidentified non-sensitive information (FP) result in similar equalized odds values because they are minimized by the big number of TN. This implies that more relevant information is deleted for one of the groups.

Conditional use accuracy equality

$$\frac{TP_a}{TP_a+FP_a} = \frac{TP_b}{TP_b+FP_b} \quad \wedge \quad \frac{TN_a}{TN_a+FN_a} = \frac{TN_b}{TN_b+FN_b}$$

✓ Both types of errors (FN and FP) are considered.

✗ The differences across groups in errors classifying true sensitive information samples as non-sensitive (FN) may not be clearly distinguished.

⚠ Models with good accuracy are expected to have high TN values, so both groups would have $\frac{TN}{TN+FN} \sim 1$, despite differences in exposing sensitive information (FN).