# Auditing the Surgical Outcome Risk Tool (SORT) for Fairness

## Background

### Fairness

The definition of fairness differs throughout literature, though most authors agree that it is something linked to the concepts of bias or discrimination. That is, an individual or group are treated unfairly due to a characteristic they have (Cambridge Dictionary). Note that this is not the same as being treated differently due to a characteristic, as some situations, such as giving more medications to someone of a heavier weight, require different treatment. Identifying whether an Artificial Intelligence (AI) model is fair to specific groups is a large area of research. Within this area several metrics have been used to identify if a model is being fair including equalised odds (Moustakidis et al., 2020), causal reasoning (Loftus et al., 2018) and search-based fairness testing (Perera et al., 2022). The first of these notions, and the one that we will use, falls into the most common category of fairness tests: group fairness notions (Makhlouf et al., 2021). Group fairness requires different sub-groups to be treated equally by considering the average for the group. This measure doesn’t consider individuals within the groups merely the group as a whole. This is implemented through statistical tests, which are easy to use though often implemented incorrectly.

We have created a fairness tool that audits supervised models using electronic healthcare models (EHRs) for fairness. This tool uses the fairness notion statistical parity, which was picked following the conditions laid out in Makhlouf et al. (2021), as we have the possibility of measurement bias and a lack of explanatory variables. Statistical parity requires the probability of a positive result to be independent of the subgroup. We consider the difference in statistical parity between subgroups, so if the notion is met then the difference is 0. The statistical method is distinct from concepts of fairness which might include consideration for expected proportional differences depending on group membership. For example, 79% of new cases of laryngeal cancer in the USA in 2022 were in men but, although the statistical parity would identify inequality with respect to sex, there is no expectation that diagnoses of laryngeal cancer should be equal across sexes (American Cancer Society, 2022). In this situation, a model that was 100% correct would not meet statistical parity. Therefore, when there is a known disproportional diagnosis between subgroups, we consider the expected difference in statistical parity and allow for a comparison between the found difference and expected difference.

### SORT

The Surgical Outcome Risk Tool (SORT) is a preoperative risk prediction tool for risk of death within 30 days of inpatient surgery (Protopapa et al., 2014; Wong et al., 2020). SORT is a logistic regression model that requires 7 variables, of which one is the clinician’s subjective risk assessment of mortality, and returns a percentage mortality risk. These percentage mortality risks can be grouped, as the clinicians subjective risk assessments are, or split into likely to survive (<5%) and other. The model is well-validated in high-income countries. National bodies (e.g., the Royal College of Surgeons in the UK) mandate the use of SORT to support the consent to surgery process, as well as for decisions on resource allocation including postoperative critical care admission. It is therefore important that SORT provides equitable predictions across population groups.

Initial tests were carried out using results of the SORT model. Our tool will compare the SORT results to actual patient mortality and the SORT results to clinicians’ assessment.

### Submission

In addition to this document, my submission also includes an interactive Jupyter notebook for testing statistical parity with supervised learning models in medical AI. The remainder of this document explains the conditions on using the notebook, an explanation on how it works, and the results.

## Auditing Tool

### Conditions

To use the tool effectively there are requirements on the system, data and model being tested. More details on this are listed below.

System requirements:

* Access to jupyter notebook,
* Access to pandas library,
* Access to os.path.

Model requirements:

* It must be a supervised model, i.e. have a ground truth.

Data requirements:

* Contain the ground truth,
* Contain the model predictions,
* Contain the protected characteristic,
* CSV format.

### Description

The flowchart in Figure 1 shows the path followed by the tool. The first three nodes represent input actions taken by the user. The following rectangular nodes represent actions taken by the tool and oval nodes represent outputs from the tool. There is a separation on the third node dependent on whether the user knows of an expected bias in their data, for example they are working with an illness that is more prevalent in one group than another. The difference between each path following this node is the additional work regarding this expected bias.

Users must have an understanding of their data and an introductory level of python, but knowledge of fairness testing or AI is not required. The tool will allow people to check AI models without needing to know how the model works.

### Diagram Description automatically generated

Figure 1: Flowchart for auditing tool

### Policy/Law

This tool is easy to use and requires only a small amount of data, that can be anonymised data in most situations. Since the tool is able to be downloaded onto a local machine high risk data does not need to be uploaded to third party servers. This means that it can be easily used in accordance will GDPR and data sharing requirements. As this tool considers difference it also meets the requirements for the Equality Act 2010, a requirement for models to meet in the UK.

## SORT

We used this tool to perform two tests on SORT. The first was comparing the model’s prediction for 30-day mortality with whether the patient died. As the predictions were given over a continuous range, we took less than 0.05 to be survived, and deceased otherwise, as has been done in the original literature on SORT. When investigating the true mortality data with regards to sex, we found there was low chance of bias as there was a difference in statistical parity of less than 0.1. Therefore SORT can fairly predict likelihood of death without a sex bias.

The second test was to take the clinicians’ subjective risk assessment as a ground truth and compare with the model’s predictions, again with regards to sex. It should be noted that in the most recent update of SORT the clinicians’ decisions are a factor in the model. This allows for a comparison of the six percentage groups that split 0% chance of death to 100%, rather than a binary comparison. A difference in statistical parity was found for each prediction group, leading to a range rather than a single result. All of these were less than 0.1, with the largest being 0.053. Therefore, when considering sex, SORT is fairly mimicking the clinicians subjective risk assessment.

As there is no expected disparity between sexes for surgical outcomes, we did not use the “expected bias” part of the tool in this example.

## Conclusion

### Summary

This tool has shown that SORT, a nationally mandated model, is fair to use, in regards to sex and the fairness notion chosen. The auditing tool meets the requirements for the equality act 2010 and GDPR. This tool is easy to download, use and understand for someone with a basic understanding of statistics. Whilst this tool has been created to analyse SORT, it will also work for any supervised model that uses electronic healthcare records.

### Future Work

There are number of ways we hope to improve this tool in the future. The current version can only take discrete data, so in situations such as SORT results must be grouped, thereby losing information. Statistical parity is only suitable in certain situations and a variety of testing methods may make this tool suitable in more situations.

## References

American Cancer Society 2022. Throat Cancer Statistics | Cases of Throat Cancer Per Year. [Accessed 30 September 2022]. Available from: https://www.cancer.org/cancer/laryngeal-and-hypopharyngeal-cancer/about/key-statistics.html.

Cambridge Dictionary n.d. Cambridge Dictionary. [Accessed 5 October 2021]. Available from: https://dictionary.cambridge.org/dictionary/english/discrimination.

Loftus, J.R., Russell, C., Kusner, M.J. and Silva, R. 2018. Causal Reasoning for Algorithmic Fairness.

Makhlouf, K., Zhioua, S. and Palamidessi, C. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*. **58**(5), p.102642.

Moustakidis, S., Papandrianos, N.I., Christodolou, E., Papageorgiou, E. and Tsaopoulos, D. 2020. Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness. *Neural Computing and Applications*.

Perera, A., Aleti, A., Tantithamthavorn, C., Jiarpakdee, J., Turhan, B., Kuhn, L. and Walker, K. 2022. Search-based fairness testing for regression-based machine learning systems. *Empirical Software Engineering*. **27**(3), p.79.

Protopapa, K.L., Simpson, J.C., Smith, N.C.E. and Moonesinghe, S.R. 2014. Development and validation of the Surgical Outcome Risk Tool (SORT). *BJS (British Journal of Surgery)*. **101**(13), pp.1774–1783.

Wong, D.J.N., Harris, S., Sahni, A., Bedford, J.R., Cortes, L., Shawyer, R., Wilson, A.M., Lindsay, H.A., Campbell, D., Popham, S., Barneto, L.M., Myles, P.S., Collaborators, S.-2: E. and Moonesinghe, S.R. 2020. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: An international prospective cohort study. *PLOS Medicine*. **17**(10), p.e1003253.