Appendix E: Use case HC2 detailed findings

E.1 Fair-by-Design: Fair Output Interpretation Methodology (FOIM) Workshop Report

E.1.1 Attendees

- Technical Expert from Akkodis
- Social Sciences Expert from ThinkTankPeriod
- Social Sciences Expert from ThinkTankPeriod
- Social Sciences Expert from ThinkTankPeriod
- Worker Representative from Eurocadres
- Socio-technical Expert from UMU (Umeå University)
- Sociotechnical Expert from Phillips
- Technical Expert from UNIBO (University of Bologna)
- Technical Expert from UNIBO (University of Bologna)
- Moderator from ALLAI
- Moderator from ALLAI

E.1.2 Use Case Summary

The AEQUITAS use case of Bias-aware Prediction of ICU Healthcare Outcomes was evaluated in this workshop[53]. The AI system aims to fairly predict patient care outcomes/outputs, in particular in-hospital mortality, and the length of stay in the Intensive Care Unit (ICU). Concretely, the system provides a score to health officials, upon which they decide the length of stay. This output is generated by continuously monitoring patient data.

The AI model has been developed using the MIMIC-III dataset, which is a comprehensive collection of e-medical records from a hospital in Boston, USA collected between 2001 and 2012. The dataset comprises over 53,000 ICU admissions, with 26 tables and 324 attributes. These include information such as demographics, mortality data, vital signs, and lab results. The main concern is the inherent bias in the MIMIC-III dataset, due to an overrepresentation of older patients, which was evaluated in this workshop.

As no such system is currently in use, for the purpose of the workshop a number of assumptions were made regarding the systems functioning, such as the types of inputs the system would take into consideration when predicting in hospital mortality and length of patient stay. The output scenarios were also fictional cases about decisions that such a system may make regarding the stay of patients with various characteristics and symptoms. The aim of these scenarios was to challenge the workshop participants to

[53] Note that due to the confidential nature of the PRE use case's algorithms and decision making, we used a related use case to test the FOIM methodology.

Funded by the European Union

identify potentially discriminatory outcomes, the same way that a doctor might need to do.

E.1.3 Methodology

The assessment followed the Fair Output Interpretation Methodology (FOIM), which is grounded in EU AI Act's Article 14 "Human Oversight". The aim of FOIM is to have an interdisciplinary and a multi-stakeholder group assess the fairness of the AI system's outputs within the context it will be used in. FOIM assists AI users and/or deployers in critically assessing the outputs of AI systems before actual deployment to ensure fair implementation. Given that AI systems are developed in diverse environments with various tasks in mind, it is paramount to address the technical, ethical, legal, and social constraints to fairness in AI model outputs.

For a model or an algorithm's outputs to be deemed fair, AI users must complete and fill in information for the 'Self-identifying interpretation biases" assessment. Such assessment includes the below mentioned overview of the system, a questionnaire to self-assess biases and an impact assessment before deployment. Details include the following:

Self-identifying Interpretation Biases-Overview of the system

- Purpose identification
- Data Source and Representation
- Outcome evaluation

Questionnaire to Self-Assess Biases

- Consumer bias
- Mode Confusion Bias
- Cognitive Bias
- Dunning-Kruger Effect
- Loss of Situational Awareness
- Rashomon Effect
- Selective Adherence
- Presentation Bias

Impact Assessment Before Deployment

- AI Risk Level
- Impact of the System: Impact on Health, on Safety, on Dignity, on Freedoms, on Equality, on Citizen Rights and on Justice.

E.1.4 Process

Prior to the workshop, participants were sent the FOIM workshop template. The additional general information provided consisted of a brief explanation of the use case, including key details of the AI model's intended functionality, the specific aims of the

Funded by the European Union

model and a detailed description of the dataset. Furthermore, FOIM template information included necessary sections and questions to complete the assessment.

The workshop was held on WEBEX and a brief introduction was given to the participants. During the session the moderators explained the goals and objectives of the workshop and provided a summary of the FOIM methodology. The workshop took a plenary format with the moderators assisting and facilitating an open discussion, whilst guiding the participants through the methodology. After the introduction, the moderators presented the AI use case, output scenarios and proceeded to conduct the Self-identifying Interpretation Biases assessment of the model's output requirements. Each question was asked and then discussed by participants to assess whether the AI model's outputs successfully met the requirements for them to be deemed fair and free of bias. During the workshop participants actively provided feedback on the quality of the methodology to ensure its validity. This specific workshop completed the Self-identifying Interpretation Biases assessment.

E.1.5 Results

The methodology outlined in the FOIM template was used to evaluate whether the AI model's outputs are fair and free of bias. The results stem from the plenary session discussion and cover the entirety of the completed Self-identifying Interpretation Biases assessment. Due to time constraints, this session was unable to cover the Questionnaire to Self-Assess Biases and the Impact Assessment Before Deployment. Instead, the consortium participants decided to work independently on the two remaining sections, with results to be shared at a later date.

E.1.5.1 Self-identifying Interpretation Biases-Overview of the system

Purpose identification

The AI use case was found to be a single and predictive purpose AI system, with the intended purposes of informing decisions, calculating a score, and ranking within the context of mortality and ICU stay for critical patients. The relevant fairness Implications include training people on how to use the system, drawing attention to the shared documentation, shifting the legal responsibility under the EU AI Act to producers when relevant, and undertaking third-party conformity assessments (DEKRA).

Data Source and Representation

The training data used was not representative of the target population (i.e., EU patients) or the intended use case, as it was sourced from a hospital in Boston, United States. In addition to limited ethnic representation, contextual differences were identified linked to differing healthcare operations. For example, the U.S. insurance system has been linked to certain fraudulent practices, such as over-prescription of treatments to secure coverage or financial benefits for hospitals. This undermines the appropriateness of the data for other geographical contexts. Furthermore, the dataset failed to account for temporal shifts. These issues raise fairness concerns due to inherent data biases stemming from the non-representative nature of the dataset. It was emphasized that

Funded by
the European Union

such shortcuts should be critically considered when evaluating outcomes to mitigate potential discrimination against any group.

Outcome evaluation

To assess the success of the system's outcomes, participants proposed evaluating its impact on population diversity. This included conducting an impact risk analysis for identified groups, asking whether the system's outcomes affected certain populations differently. Additionally, participants recommended using a matrix that integrates social and legal perspectives alongside medical considerations. One suggested success metric was the level of trust doctors place in the system, as it may indicate the system's practical applicability to real-world cases. However, it was noted that physician subjectivity could influence this metric and should be carefully considered. Finally, participants highlighted that transparency and user agency should also be key indicators of success. This could involve assessing whether users are provided with sufficient and understandable information enabling them to make meaningful decisions based on both the system's outputs and their own expertise.

When assessing whether historical or other forms of bias had been identified and addressed in the system's design, it was found that historical biases had not been considered. As a result, there is a risk that past injustices could be perpetuated through the AI system. The implication was a need for comprehensive documentation to ensure developers and providers actively mitigate bias, particularly against marginalized groups, such as Black communities. While users are able to provide direct feedback to the AI provider, there are currently no built-in mechanisms to control system outcomes, including the absence of a 'stop' function. Consequently, it was concluded that the system should not be deployed until the necessary fairness considerations are addressed for fair outcomes.

E.1.6 Areas of Improvement

The methodology was an appropriate medium to address and assess AI model output fairness. Questions remained mostly relevant and flowed cohesively with one another, which provided a clearer view of the EU AI Act's Article 14 criterion. However, areas of improvement were highlighted, which includes clarity, and reformulation of certain methodology sections and questions.

The section on Mode Confusion Bias includes inherently difficult questions to answer. Doctors should be trained to operate a system but also might not be aware of other medical staff errors in data entry which permeate into the AI decision-making. Expanding the section to provide better scope of the mode might benefit the assessment.

The section Dunning-Kruger Effect requires a sensitivity check, to ensure AI users/ doctors can answer the question genuinely. Provided the nature of such section and question, doctors might not be pleased, if their expertise is questioned. This can result in reduced cooperation from the doctor's side. Furthermore, clarification on Question 2

Funded by
the European Union

in terms of which experts were consulted could yield more insight into output interpretation.

Funded by
the European Union