User Study: Comparison between explainable AI models

Dear participant,

In this study we are comparing AI models for a Covid-19 prediction task. Based on an extensive history of Covid-19 patients and the respective outcomes (recovered/died), an AI model is trained to predict whether a patient is at a *high risk* of dying, and thus needs to be in the emergency station, or not.

To this end, *explainable* AI models are used. These AI can explain to a doctor why a certain decision is taken, so that the doctor can intervene in case a faulty decision is taken.

This study examines three different kinds of explainable AI models. The difference between each model is the type of explanation they give, which varies in its information content. Your task is the following:

1. You are shown two patients. For each patient, you will see the AIs prediction, as well as an explanation. Look at the example below from AI 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | | | Effect |
| Age: 55 | +28.1% | | |
| Race: black | +0.1% | | |
| Sex: male | -1.1% | | |
| Diabetes: Yes | | -3.6% | |
| Artery disease: No | | +7.3% | |
| … | | … | |
| Average survival %  Predicted survival % | | 54.1%  84.9% | |

Each attributes effect on survival rate is explained. The prediction is explained as the sum of effects: 28.1% + 0.1% - 1.1% - 3.6% + 7.3% + 54.1% = 84.9%

Positive effects (+28.1%) are good for survival, negative effects are (-3.6%) bad.

1. Examine both explanations, and assess whether they make sense to you.
2. You will see the explanation for two patients. Your task is to estimate, what the model would predict for a new patient. Use the two provided explanations and a calculator if needed.
3. There are 3 AI models to review in total.
4. Please complete a short survey in the end.

Thank you very much for participating, it really helps us a lot!

# AI 1

|  |  |
| --- | --- |
| *Patient 1*  Attribute | Effect |
| Age: 87 | -7.5% |
| Race: black | +4.0% |
| Sex: female | -1.3% |
| High blood pressure: No | +2.8% |
| High blood fat: No | -2.3% |
| Diabetes: No | -4.5% |
| Artery disease: Yes | -14.6% |
| Heart failure: No | +2.4% |
| Vascular disease: No | +2.1% |
| Hepatitis: No | +1.6% |
| Partial kidney failure: No | -4.6% |
| Full kidney failure: No | +3.1% |
| Asthma: No | -4.7% |
| Chronic lung disease: No | -3.5% |
| Dementia: Yes | -9.4% |
| Cancer: No | 0% |
| Average survival %  Predicted survival % | 54.1%  17.7% |

Each attribute is assigned a positive or negative impact, e.g., Age 87 => -7.5% chance of survival.

The predicted chance is the sum of effects + the average. For example, for patient 1: - 7.5% + 4.0% - 1.3% + … + 54.1% (average) = 17.7%

Attributes of patient 3 marked light blue are shared with patient 1, light green with patient 2.

Unmarked attributes are shared with both patient 1 and patient 2.

**Your task**: Estimate the model’s prediction for *patient 3* based on your intuitive understanding of the explanations. You may use a calculator/your phone as aid.

|  |  |
| --- | --- |
| *Patient 2*  Attribute | Effect |
| Age: 93 | -28.1% |
| Race: black | 0.0% |
| Sex: male | +4.8% |
| High blood pressure: Yes | -11.3% |
| High blood fat: No | -5.0% |
| Diabetes: No | -2.6% |
| Artery disease: No | +5.8% |
| Heart failure: No | 0.0% |
| Vascular disease: No | +4.5% |
| Hepatitis: No | 0.0% |
| Partial kidney failure: No | -4.3% |
| Full kidney failure: No | +1.4% |
| Asthma: No | +6.2% |
| Chronic lung disease: No | -2.4% |
| Dementia: No | 0.0% |
| Cancer: No | -4.3% |
| Average survival %  Predicted survival % | 54.1%  18.6% |

|  |  |
| --- | --- |
| **Task:** *Patient 3*  Attribute | Effect |
| Age: 87 |  |
| Race: black |  |
| Sex: male |  |
| High blood pressure: No |  |
| High blood fat: No |  |
| Diabetes: No |  |
| Artery disease: No |  |
| Heart failure: No |  |
| Vascular disease: No |  |
| Hepatitis: No |  |
| Partial kidney failure: No |  |
| Full kidney failure: No |  |
| Asthma: No |  |
| Chronic lung disease: No |  |
| Dementia: No |  |
| Cancer: No |  |
| Average survival % | 54.1% |
| Predicted survival % | **???** |

# AI 2

|  |  |  |
| --- | --- | --- |
| *Patient 4*  Attribute | Effect | Additional  Effect |
| **Age: 63** | **+1.3%** |  |
| **Sex: male** | **-4.7%** |  |
| **Diabetes: Yes** | **-4.5%** | **-8.7%** |
| High blood fat: Yes | -0.1% |  |
| Artery disease: No | +4.4% |  |
| Hepatitis: No | +0.7% | -3.4% |
| Race: black | +2.6% |  |
| High blood pressure: Yes | -0.4% | -2.8% |
| Vascular disease: No | +0.4% |  |
| Asthma: No | -0.5% | +0.2% |
| Heart failure: No | -0.8% |  |
| Partial kidney failure: No | -1.0% |  |
| Full kidney failure: No | -0.1% |  |
| Chronic lung disease: No | -0.3% |  |
| Dementia: No | -0.1% |  |
| Cancer: No | 0% |  |
| Average survival %  Predicted survival % | 54.1%  36.0% |  |

AI 2’s explanations are based on *combinations* of attributes. Attributes with a significant *combined effect* are grouped together.

“Patient 4 is a 63-year-old male *with* Diabetes; therefore, his chance of survival goes down an additional 8.7%.”

The prediction is the sum of individual effects + additional effects + average.

E.g., patient 4: 1.3% - 4.7% - 4.5% - 8.7% + … + 54.1% = 36%

Combinations marked **bold** are the ones that are shared with patient 6

|  |  |  |
| --- | --- | --- |
| *Patient 5*  Attribute | Effect | Additional  Effect |
| Age: 58 | +7.0% |  |
| Diabetes: Yes | -4.3% |  |
| High blood fat: No | -0.5% | -6.0% |
| **Race: black** | **+2.3%** |  |
| **High blood pressure: Yes** | **-0.6%** |  |
| **Artery disease: No** | **+4.3%** | **-4.0%** |
| **Cancer: No** | **+0.4%** |  |
| **Vascular disease: No** | **+0.1%** | **+0.7%** |
| Sex: female | +4.0% |  |
| Hepatitis: No | +0.2% | -0.6% |
| Heart failure: No | +0.2% |  |
| Partial kidney failure: No | -0.1% |  |
| Full kidney failure: No | -0.6% |  |
| Chronic lung disease: No | 0% |  |
| Dementia: No | +0.9% |  |
| Asthma: No | -0.5% |  |
| Average survival %  Predicted survival % | 54.1%  57.0% |  |

**Your task**: Estimate the prediction for patient 6.

|  |  |  |
| --- | --- | --- |
| **Task:** *Patient 6*  Attribute | Effect | Additional  Effect |
| Age: 63 |  |  |
| Sex: male |  |  |
| Diabetes: Yes |  |  |
| Race: black |  |  |
| High blood pressure: Yes |  |  |
| Artery disease: No |  |  |
| Cancer: No |  |  |
| Vascular disease: No |  |  |
| High blood fat: No |  |  |
| Hepatitis: No |  |  |
| Heart failure: No |  |  |
| Partial kidney failure: No |  |  |
| Full kidney failure: No |  |  |
| Chronic lung disease: No |  |  |
| Dementia: No |  |  |
| Cancer: No |  |  |
| Average survival %  Predicted survival % | 54.1%  **???** |  |

# AI 3

The third and last AI model provides the most informative explanations. An effect for is given for *each combination* of attributes. You are shown the combinations with the highest absolute effects. Combinations marked **bold** are the ones that are shared with patient 9. The prediction is the sum of effects as usual. Estimate the prediction for patient 9 (next page).

|  |  |
| --- | --- |
| *Patient 7*  Attribute | Effect |
| **Artery disease: Yes** | **-13.4%** |
| **Age:83.00** | **-11.9%** |
| Age:83.00 & Sex:female & Diabetes:Yes & Artery disease:Yes | -8.9% |
| **Age:83.00 & Race:white** | **+8.1%** |
| Age:83.00 & Race:white & Diabetes:Yes & Artery disease:Yes | -6.2% |
| Age:83.00 & sex:female & High blood pressure:Yes | -5.2% |
| Diabetes: Yes | -5.0% |
| High blood pressure:Yes & Diabetes:Yes | +4.8% |
| Age:83.00 & Race:white & High blood pressure:Yes | -4.3% |
| **Race:white & Artery disease:Yes** | **-4.2%** |
| **Age:83.00 & sex:female & Artery disease:Yes** | **+4.1%** |
| sex:female & Diabetes:Yes | +3.8% |
| Age:83.00 & High blood pressure:Yes & Artery disease:Yes | -3.5% |
| Age:83.00 & High blood pressure:Yes & Diabetes:Yes & Artery disease:Yes | +3.5% |
| Age:83.00 & High blood pressure:Yes & Diabetes:Yes | -3.4% |
| **sex:female** | **+3.2%** |
| … | … |
| Average survival %  Predicted survival % | 54.1%  15.6% |

|  |  |
| --- | --- |
| *Patient 8*  Attribute | Effect |
| **Artery disease: Yes** | **-12.8%** |
| Age:70 & High blood fat:Yes & Artery disease:Yes | -11.5% |
| **High blood pressure:No & High blood fat:Yes & Diabetes:No & Artery disease:Yes** | **+7.0%** |
| Age:70 & Race:black & sex:female | +6.8% |
| **High blood fat:Yes & Artery disease:Yes** | **+6.0%** |
| Age:70 & Race:black & sex:female & High blood fat:Yes | -5.6% |
| Age:70 & High blood pressure:No & Diabetes:No & Artery disease:Yes | +5.1% |
| Age:70 & High blood pressure:No & Diabetes:No | +5.0% |
| Age:70 & sex:female & High blood fat:Yes & Artery disease:Yes | +4.4% |
| Age:70 & High blood pressure:No & High blood fat:Yes & Diabetes:No | -4.3% |
| Age:70 & Race:black & sex:female & Artery disease:Yes | +4.2% |
| Age:70 & High blood fat:Yes & Diabetes:No | +4.0% |
| Age:70 & Race:black | -4.0% |
| **High blood pressure:No & High blood fat:Yes & Diabetes:No** | **-3.6%** |
| Age:70 & High blood pressure:No | -3.5% |
| Race:black & High blood fat:Yes & Artery disease:Yes | -3.4% |
| … | … |
| Average survival %  Predicted survival % | 54.1%  29.1% |

|  |  |
| --- | --- |
| **Task:** *Patient 9*  Attribute | Effect |
| Age: 83 |  |
| Race: white |  |
| Sex: female |  |
| High blood pressure: No |  |
| High blood fat: Yes |  |
| Diabetes: No |  |
| Artery disease: Yes |  |
| Heart failure: No |  |
| Vascular disease: No |  |
| Hepatitis: No |  |
| Partial kidney failure: No |  |
| Full kidney failure: No |  |
| Asthma: No |  |
| Chronic lung disease: No |  |
| Dementia: No |  |
| Cancer: No |  |
| Average survival % | 54.1% |
| Predicted survival % |  |

# Google Form

The main effort is done, thank you!

Please go over the provided explanations again (patients 1, 2, 4, 5, 7, 8 only). Think whether they make sense to you and why they do/ do not align with your intuitions. Then please fill out the form here <redacted>