ORIENT

tOwaRds Informed dEcisions iN colorecTal cancer screening

ML model for Shared Decision-Making Tool for CRC screening

Daiane M. Seibert and Karen Feyen

Kom op tegen Kanker

Prevention project realised with the support of KOTK (projectID: 12751)

DOMUS MEDICA ALLES VOOR HUISARTSEN · vito · Universiteit Antwerpen · THOMAS MORE · RESEARCH @THOMAS MORE

What will Orient try to do?

Provide the general practitioners (GPs) with the tools to discuss colorectal cancer screening and its benefits and harms with patients belonging to vulnerable groups, so that they are well-informed and can take a deliberate decision to participate or not in screening.

ORIENT

tOwaRds Informed dEcisions iN colorecTal cancer screening

Key messages about SDM tools

- SDM tools for cancer screening are **effective in increasing knowledge** on cancer and cancer screening of **vulnerable people.**
- SDM tools are **effective in reducing decision conflicts** among people considering cancer screening.
- SDM tools are **effective in increasing screening intentions among vulnerable people**, especially those who are considering **colorectal cancer screening**.
- Given the complexities of patients’ and clinicians’ preferences in SDM tool characteristics, **fostering collaboration** between patients and clinicians **during the creation of an SDM tool** for cancer screening is essential.

Key messages from the literature reviews performed in the project.

Key messages about SDM tools

- Out of numerous risk prediction models for CRC with good model performance, **only a few** of them can be potentially **integrated** into practical healthcare settings.

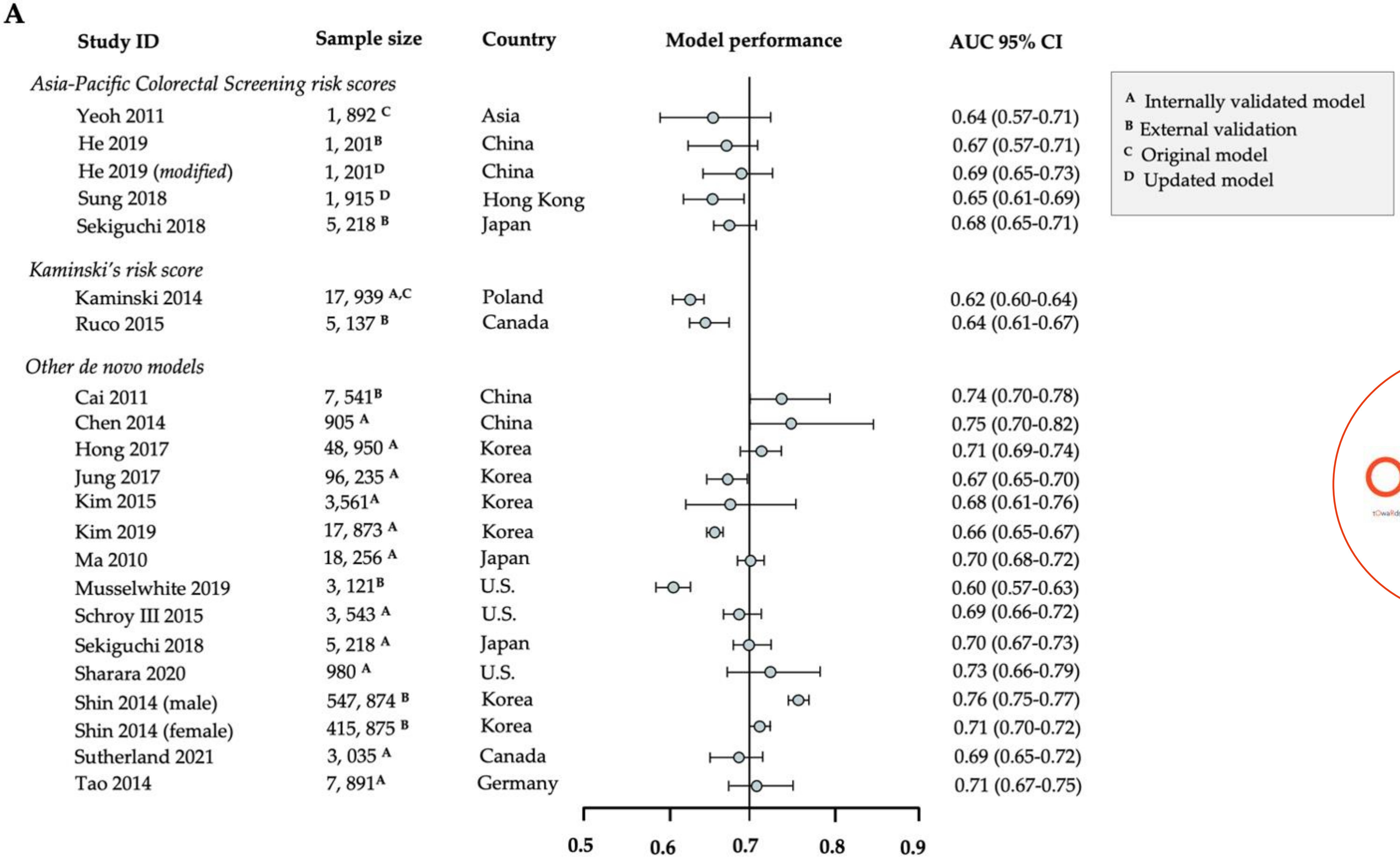
- It is crucial to establish a **standardized reporting** of risk prediction models in CRC that accounts for the model's interpretability, generalizability, and potential clinical utility.

Key messages from the literature reviews performed in the project.

The constraints

- The tool need to be simple in a way to be used in a 15-min appointment.
- Features based in laboratory tests do not fit our objectives. (Conventional models)
- Information needed to feed into the model will be collected by the doctor.
- Things to keep in mind:
  - The language barrier over the vulnerable population.
  - Lack of knowledge about CRC screening.

tOwaRds Informed dEcisions iN colorecTal cancer screening



Study ID	Sample size	Country	AUC 95% CI
Asia-Pacific Colorectal Screening risk scores			
Yeoh 2011	1, 892 ^{C}	Asia	0.64 (0.57-0.71)
He 2019	1, 201^{B}	China	0.67 (0.57-0.71)
He 2019 (*modified*)	1, 201^{D}	China	0.69 (0.65-0.73)
Sung 2018	1, 915 ^{D}	Hong Kong	0.65 (0.61-0.69)
Sekiguchi 2018	5, 218 ^{B}	Japan	0.68 (0.65-0.71)
Kaminski's risk score			
Kaminski 2014	17, 939 ^{A,C}	Poland	0.62 (0.60-0.64)
Ruco 2015	5, 137 ^{B}	Canada	0.64 (0.61-0.67)
Other de novo models			
Cai 2011	7, 541^{B}	China	0.74 (0.70-0.78)
Chen 2014	905 ^{A}	China	0.75 (0.70-0.82)
Hong 2017	48, 950 ^{A}	Korea	0.71 (0.69-0.74)
Jung 2017	96, 235 ^{A}	Korea	0.67 (0.65-0.70)
Kim 2015	3,561^{A}	Korea	0.68 (0.61-0.76)
Kim 2019	17, 873 ^{A}	Korea	0.66 (0.65-0.67)
Ma 2010	18, 256 ^{A}	Japan	0.70 (0.68-0.72)
Musselwhite 2019	3, 121^{B}	U.S.	0.60 (0.57-0.63)
Schroy III 2015	3, 543 ^{A}	U.S.	0.69 (0.66-0.72)
Sekiguchi 2018	5, 218 ^{A}	Japan	0.70 (0.67-0.73)
Sharara 2020	980 ^{A}	U.S.	0.73 (0.66-0.79)
Shin 2014 (male)	547, 874 ^{B}	Korea	0.76 (0.75-0.77)
Shin 2014 (female)	415, 875 ^{B}	Korea	0.71 (0.70-0.72)
Sutherland 2021	3, 035 ^{A}	Canada	0.69 (0.65-0.72)
Tao 2014	7, 891^{A}	Germany	0.71 (0.67-0.75)

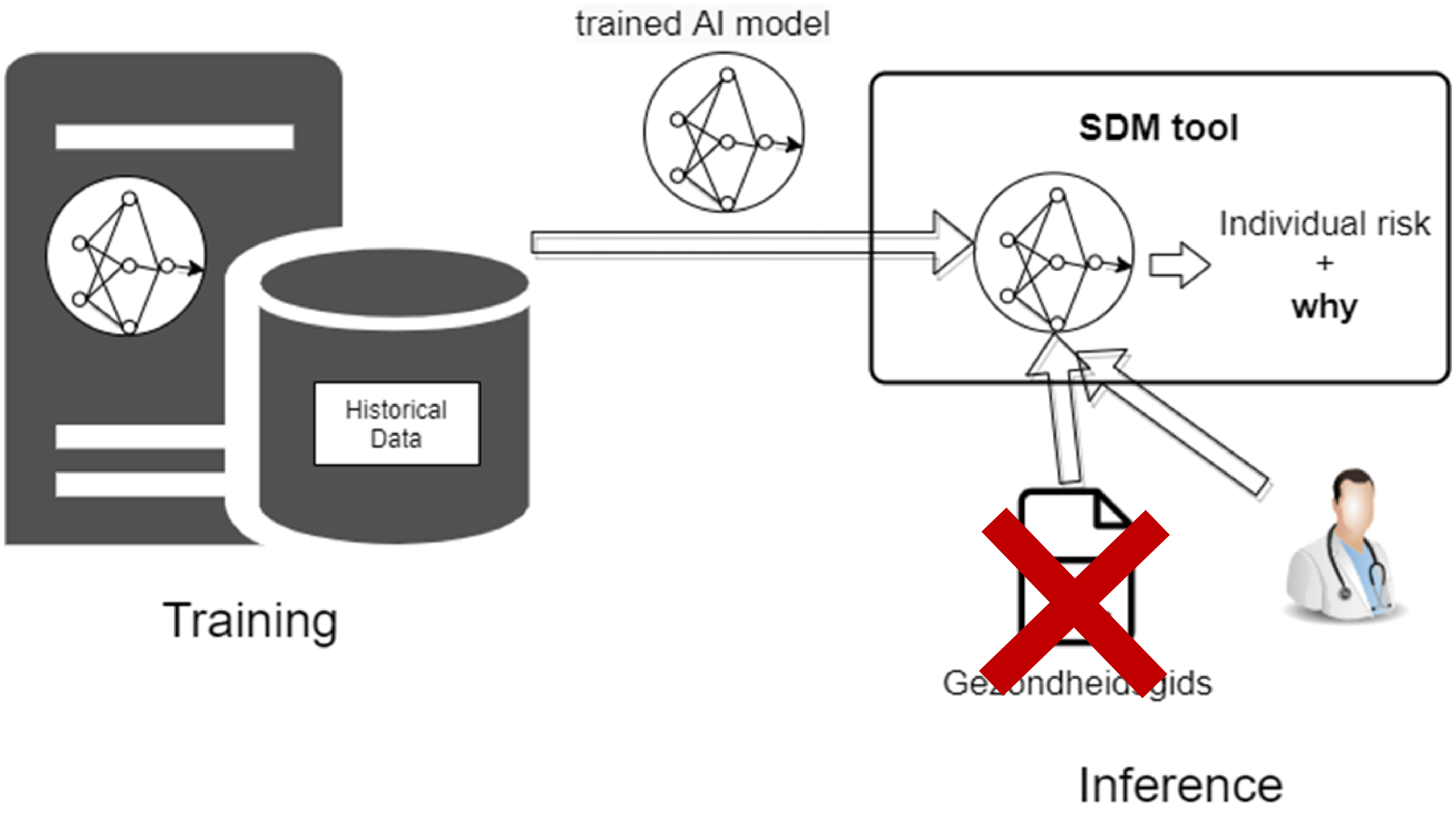
^{A} Internally validated model
^{B} External validation
^{C} Original model
^{D} Updated model

ORIENT
tOwaRds Informed dEcisions iN colorecTal cancer screening

Model performance of conventional risk models for colorectal cancer with 95% confidence interval (CI)

Info collected during the literature review performed in the project.

AI for ORIENT

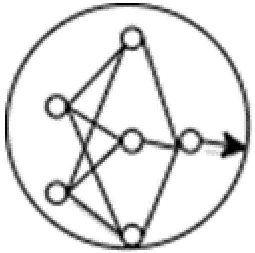


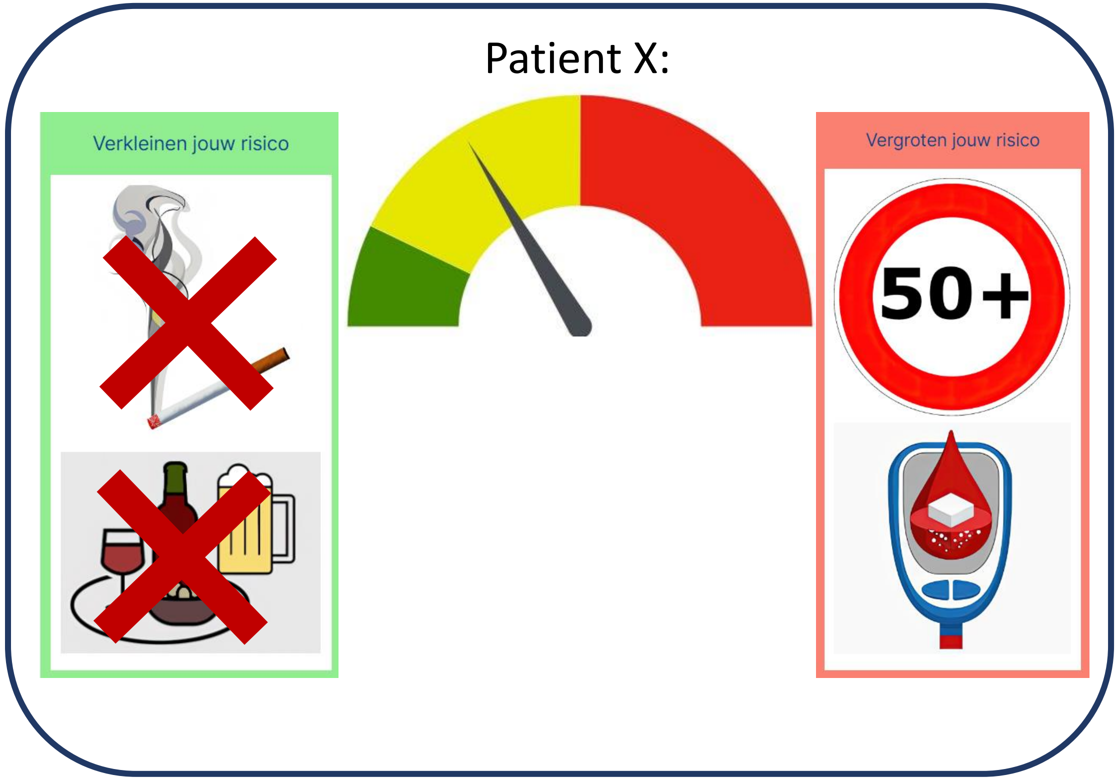
AI for ORIENT – Inference example

Patient X:

- 65 years old
- Former smoker
- Diabetes
- No Hypertension

⋮



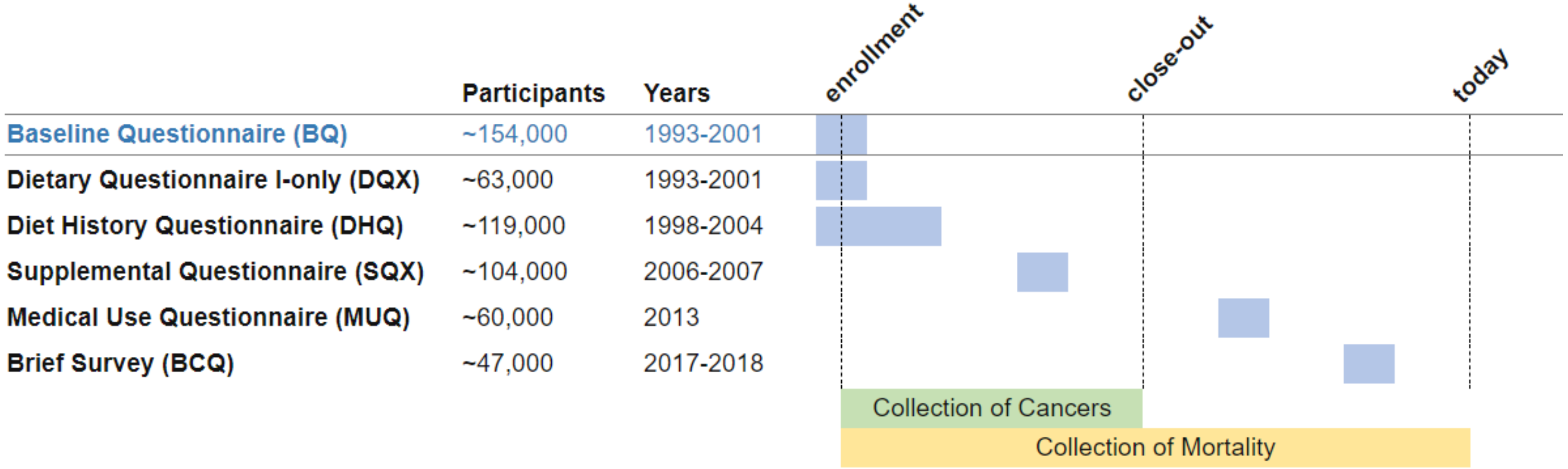


⇒ Using *Explainable* AI

Data: PLCO - The Prostate, Lung, Colorectal and Ovarian Cancer

PLCO study aimed to evaluate the impact of screening exams on reducing mortality from prostate, lung, colorectal, and ovarian cancers. Participants assigned to the control arm received usual care, whereas participants assigned to the intervention arm were invited to receive screening exams.

Enrolment period: 1993 - 2001
Data size: ~154,000
Screening period: 1993 - 2009
Data Collection: +500 risk factors, collected through 6 questionnaires



	Participants	Years
Baseline Questionnaire (BQ)	~154,000	1993-2001
Dietary Questionnaire I-only (DQX)	~63,000	1993-2001
Diet History Questionnaire (DHQ)	~119,000	1998-2004
Supplemental Questionnaire (SQX)	~104,000	2006-2007
Medical Use Questionnaire (MUQ)	~60,000	2013
Brief Survey (BCQ)	~47,000	2017-2018

enrollment
close-out
today
Collection of Cancers
Collection of Mortality

AI model: feature engineering

US dataset

- ± 150 000 patients
- 500 + risk factors

First risk factor selection.

E.g. factors removed:

- Screening & diagnosis features
- US specific
- Too specific

US dataset

- ± 150 000 patients
- 248 risk factors
- 1 label

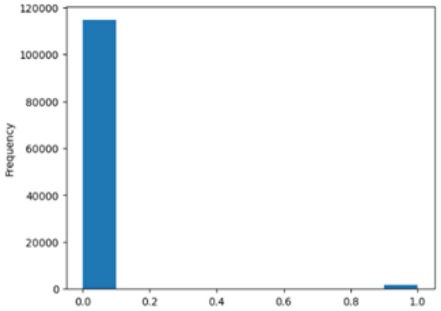
Second risk factor & patient selection.

E.g. removed:

- Patients with too many missing values
- Factors statistically too insignificant

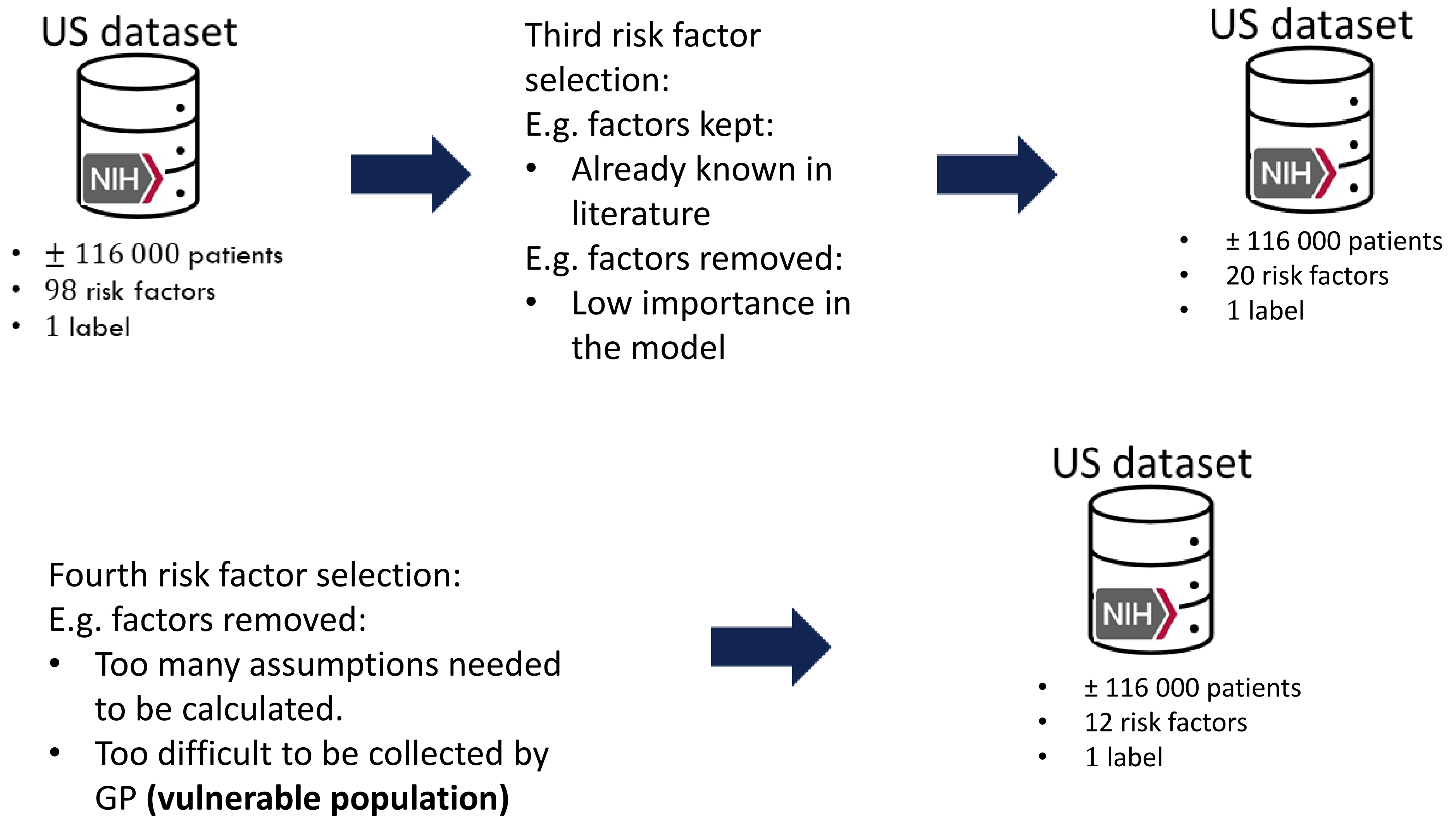
US dataset

- ± 116 000 patients
- 98 risk factors
- 1 label



- Balance ±1:70

AI model: feature engineering



US dataset

- ± 116 000 patients
- 98 risk factors
- 1 label

Third risk factor selection:
E.g. factors kept:

- Already known in literature

E.g. factors removed:

- Low importance in the model

US dataset

- ± 116 000 patients
- 20 risk factors
- 1 label

Fourth risk factor selection:
E.g. factors removed:

- Too many assumptions needed to be calculated.
- Too difficult to be collected by GP **(vulnerable population)**

US dataset

- ± 116 000 patients
- 12 risk factors
- 1 label

ORIENT
tOwaRds Informed dEcisions iN colorecTal cancer screening

Selected risk factors (Features)

1. **Sex:** *Male, Female*
2. **Age:** *Numeric*
3. ***Height****: Numeric*
4. ***Weight:*** *Numeric*
5. ***BMI:*** *Numeric*
6. **Hypertension:** Did the participant ever have high blood pressure? *Yes/No*
   Question modified: Do you take medication for hypertension ? *Yes/No*
7. **Heart problems:** Did the participant ever have coronary heart disease or a heart attack? *Yes/No*
   Question modified: Do you take medication for heart problems? *Yes/No*
8. **Diabetes:** Did the participant ever had diabetes? *Yes/No*
9. **Family History of Colorectal Cancer:** Colorectal cancer family history in first-degree relatives. Includes parents full-siblings and children. *Yes/No/Possible, cancer type not clear*
10. **Smoke history:** Participant's current cigarette smoking status. *Former smoker/ Current smoker / Never smoked*
11. **Smoke quantity:** In the time that you smoked, how many cigarettes per day (approximately)? *0 / 1-10 / 11-20 / 21-30 / More than 30*
12. **Alcohol drink history:** Alcoholic Beverages (drinks/day) – DHQ. *Does not drink / Drinks, but less than 7 drinks per week / Drinks, more than 7 drinks per week*
    Alcohol from Beer Wine and Liquor - Age 40-54 (drinks/day) – DHQ. *Did not drink / Drank, but less than 7 drinks per week / Drank, more than 7 drinks per week*

ORIENT

tOwaRds Informed dEcisions iN colorecTal cancer screening

Selected risk factors (Features)

1. **Sex:** *Male, Female*
2. **Age:** *Numeric*
3. ***Height***: *Numeric*
4. ***Weight:*** *Numeric*
5. ***BMI:*** *Numeric*
6. **Hypertension:** Did the participant ever have high blood pressure? *Yes/No*
   Question modified: Do you take medication for hypertension ? *Yes/No*
7. **Heart problems:** Did the participant ever have coronary heart disease or a heart attack? *Yes/No*
   Question modified: Do you take medication for heart problems? *Yes/No*
8. **Diabetes:** Did the participant ever had diabetes? *Yes/No*
9. **Family History of Colorectal Cancer:** Colorectal cancer family history in first-degree relatives. Includes parents full-siblings and children. *Yes/No/Possible, cancer type not clear*
10. **Smoke history:** Participant's current cigarette smoking status. *Former smoker/ Current smoker / Never smoked*
11. **Smoke quantity:** In the time that you smoked, how many cigarettes per day (approximately)? *0 / 1-10 / 11-20 / 21-30 / More than 30*
12. **Alcohol drink history:** Alcoholic Beverages (drinks/day) – DHQ. *Does not drink / Drinks, but less than 7 drinks per week / Drinks, more than 7 drinks per week*
    Alcohol from Beer Wine and Liquor - Age 40-54 (drinks/day) – DHQ. *Did not drink / Drank, but less than 7 drinks per week / Drank, more than 7 drinks per week*

#9
This factor is an exception factor in our tool.

Following the local guidelines, who has family history of colorectal cancer, should be advised to go through colonoscopy directly.

Selected risk factors (Features)

1. **Sex:** *Male, Female*
2. **Age:** *Numeric*
3. ***Height****: Numeric*
4. ***Weight:*** *Numeric*
5. ***BMI:*** *Numeric*
6. **Hypertension:** Did the participant ever have high blood pressure? *Yes/No*
   Question modified: Do you take medication for hypertension ? *Yes/No*
7. **Heart problems:** Did the participant ever have coronary heart disease or a heart attack? *Yes/No*
   Question modified: Do you take medication for heart problems? *Yes/No*
8. **Diabetes:** Did the participant ever had diabetes? *Yes/No*
9. **Family History of Colorectal Cancer:** Colorectal cancer family history in first-degree relatives. Includes parents full-siblings and children. *Yes/No/Possible, cancer type not clear*
10. **Smoke history:** Participant's current cigarette smoking status. *Former smoker/ Current smoker / Never smoked*
11. **Smoke quantity:** In the time that you smoked, how many cigarettes per day (approximately)? *0 / 1-10 / 11-20 / 21-30 / More than 30*
12. **Alcohol drink history:** Alcoholic Beverages (drinks/day) – DHQ. *Does not drink / Drinks, but less than 7 drinks per week / Drinks, more than 7 drinks per week*
    Alcohol from Beer Wine and Liquor – at your forties (drinks/day) – DHQ. *Did not drink / Drank, but less than 7 drinks per week / Drank, more than 7 drinks per week*

#6 and #7
Heart and Hypertension factor in our model decreased the risk. We discovered that previous studies indicated the use of standard medication for these conditions also reduced the risk. As a result, we decided to modify our research question.

ORIENT
tOwaRds Informed dEcisions iN colorecTal cancer screening

Data change

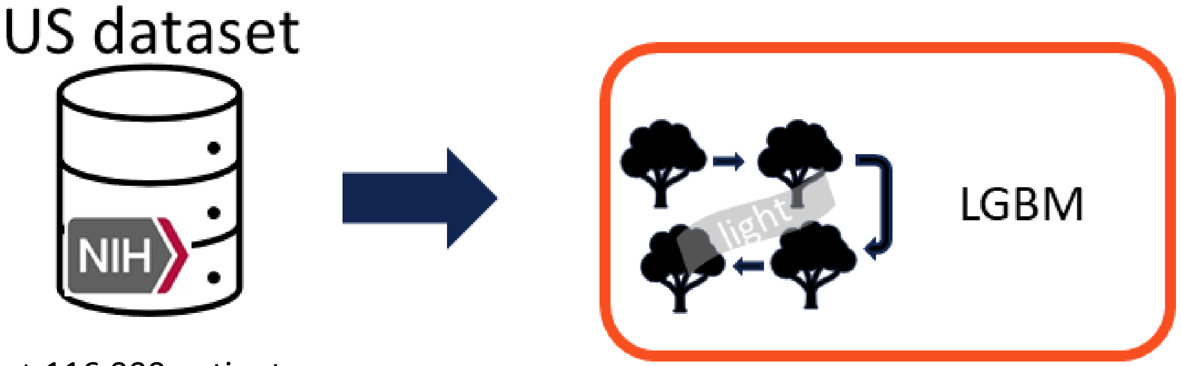
1. **Alcohol drink history:**
   - Alcoholic Beverages (drinks/day) – DHQ. *Does not drink / Drinks, but less than 7 drinks per week / Drinks, more than 7 drinks per week*
   - Alcohol from Beer Wine and Liquor – At your forties (drinks/day) – DHQ. *Did not drink / Drank, but less than 7 drinks per week / Drank, more than 7 drinks per week*

A regular beer is considered as one drink.

The threshold of 7 biers was defined based in the local dietary guidelines and, also the proportion of cancer incidence in our data.

2. ***Age:** The age was collected when all patients were enrolled.*
   - *We have the age that the cancer was discovered, for positive cases.*
   - *And what about negative cases?*
     - *Age that colonoscopy was performed (just for intervention arm)*
   - *Ages based in the colonoscopy date was added for negative cases.*
     Cases where a person died from colorectal cancer but was negative, were excluded.

Model



US dataset

LGBM

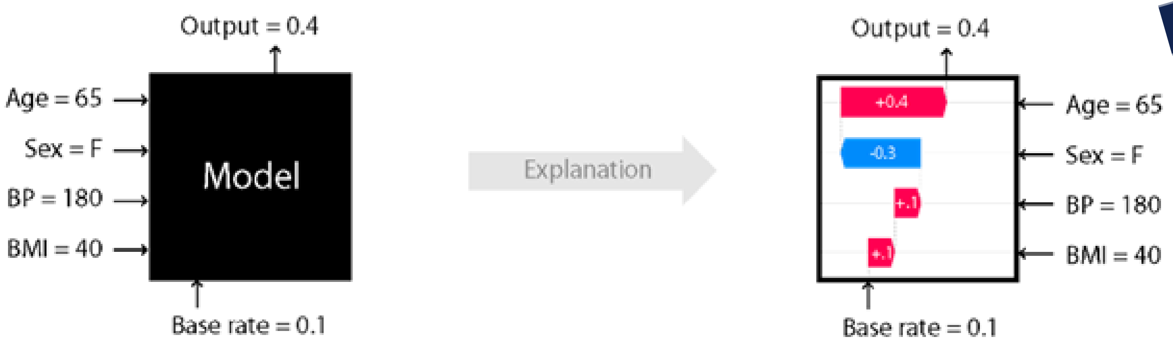
- ± 116 000 patients
- 12 risk factors
- 1 label

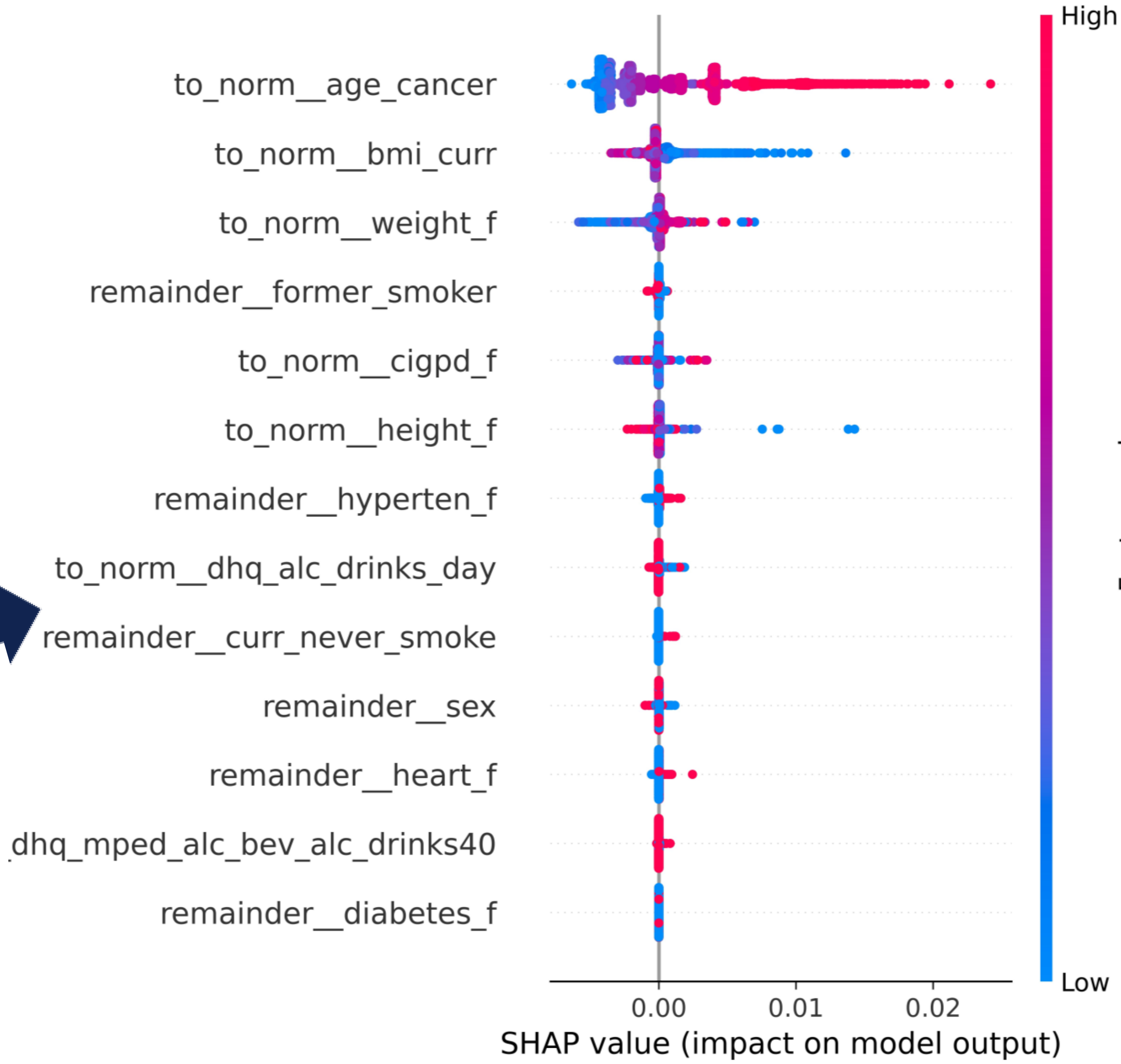
ORIENT

tOwaRds Informed dEcisions iN colorecTal cancer screening

Explainable AI

SHAP





Other terminologies used in the literature

Measures of prediction model performance

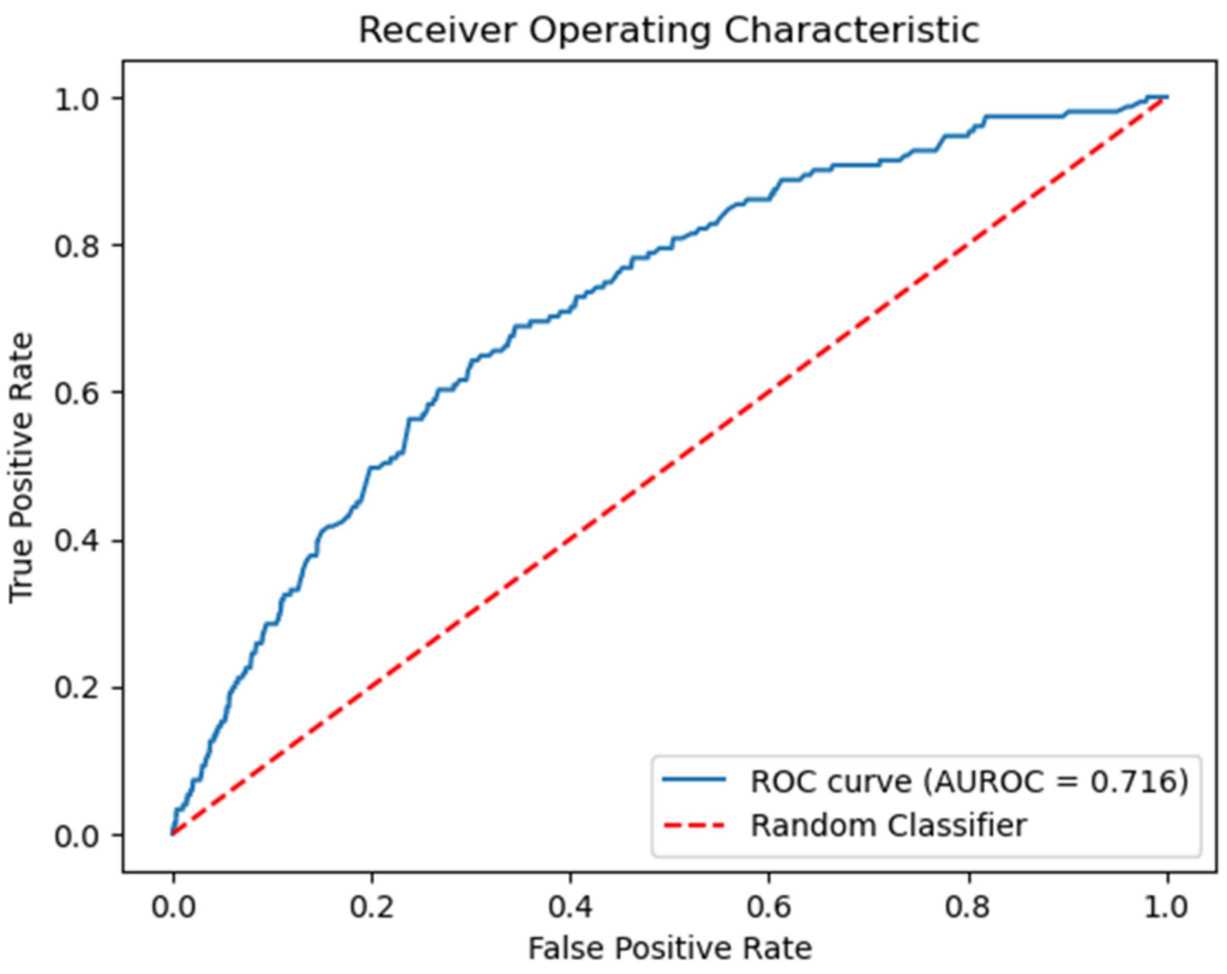
Terms	Definition
AUC	Area under the curve, in this case the receiver operating characteristic curve. A measure of discrimination. For prediction models based on logistic regression; this corresponds to the probability that a randomly selected diseased patient had a higher risk prediction than a randomly selected patient who does not have the disease
Calibration	Correspondence between predicted and observed risks is usually assessed in calibration plots or by calibration intercepts and slopes.
Sensitivity	The proportion of true positives in truly diseased patients
Specificity	The proportion of true negatives in truly non-diseased patients.
Positive predictive value	The proportion of true positives in patients classified as positive.
Negative predictive value	The proportion of true negatives in patients classified as negative.

Info from the literature reviews performed in the project.

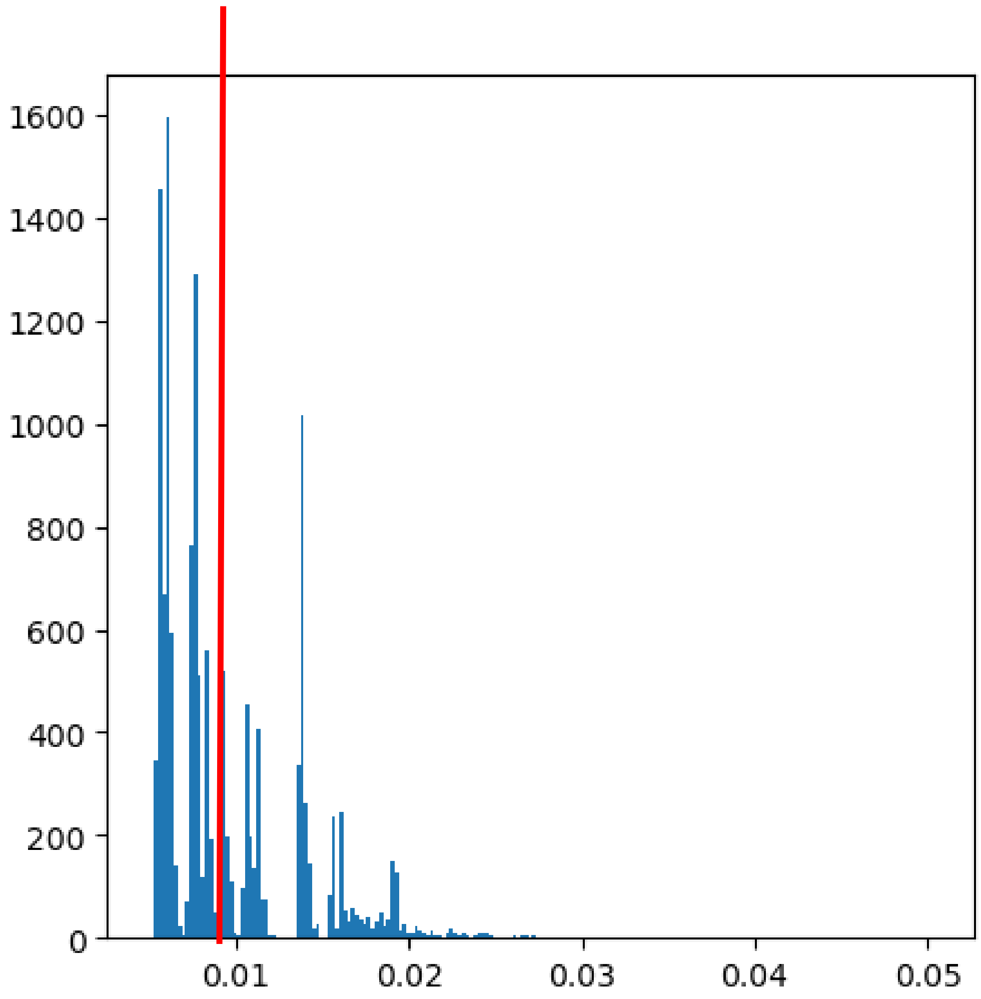
Model's performance

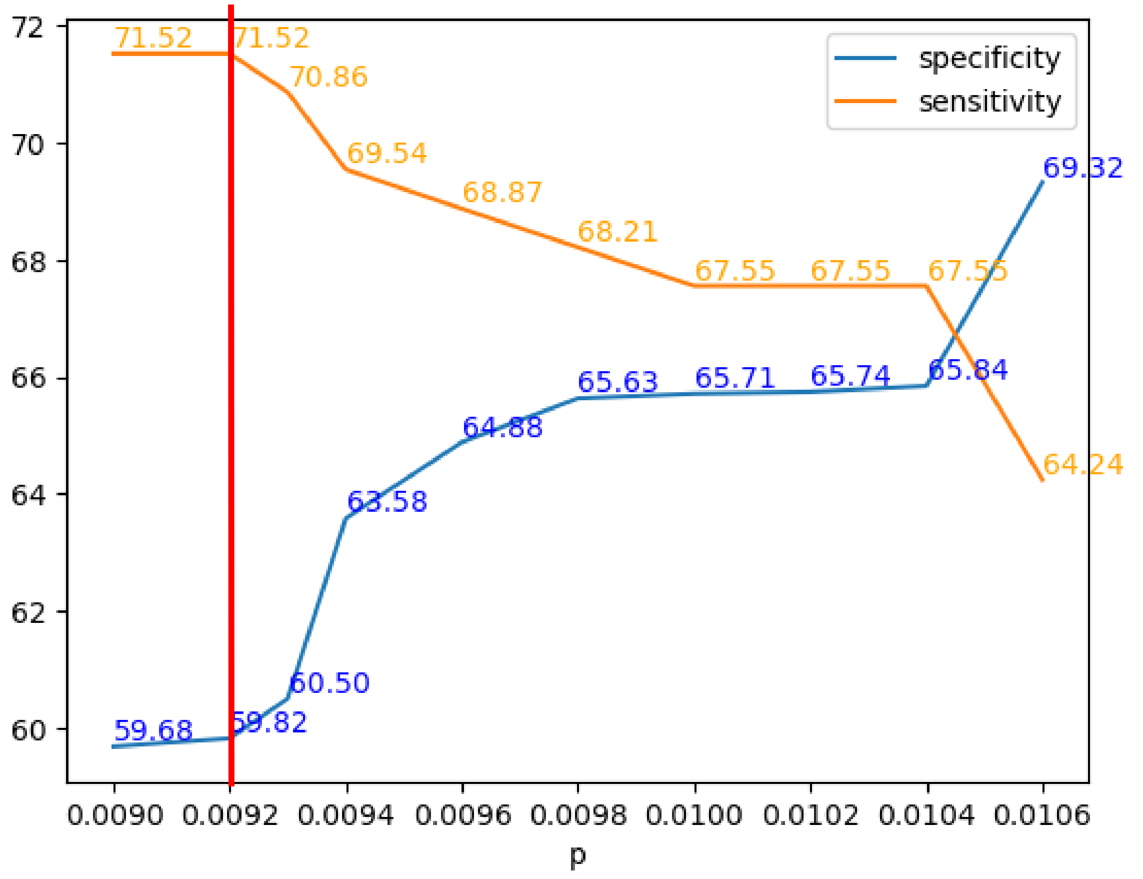
AUROC of our model is 0.716.
It is a metric used to evaluate the performance of binary classification models.

The Receiver Operating Characteristic (ROC) curve is a **graphical representation of the true positive rate (sensitivity) plotted against the false positive rate (1 - specificity)** as the discrimination threshold of the model varies.
The area under this ROC curve, the AUROC, is a single scalar value that summarizes the overall performance of the model across various threshold settings.

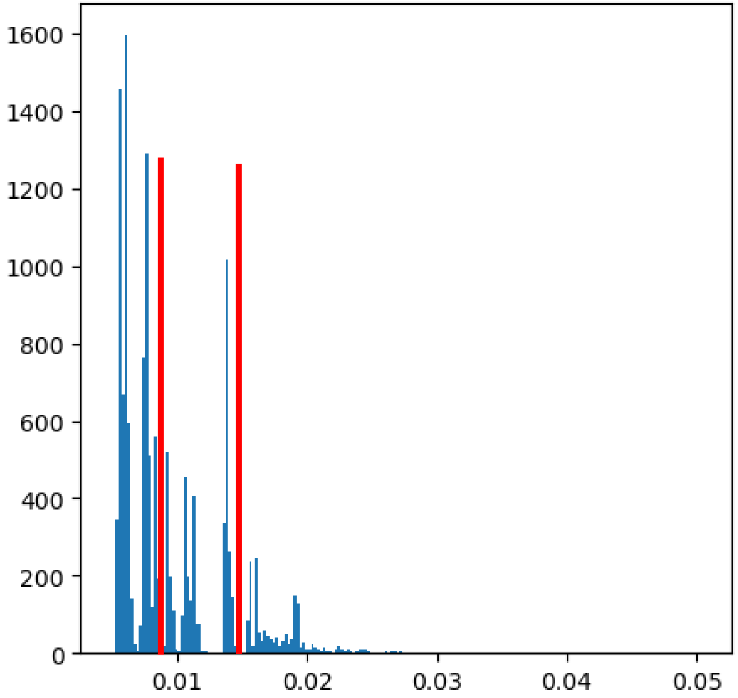


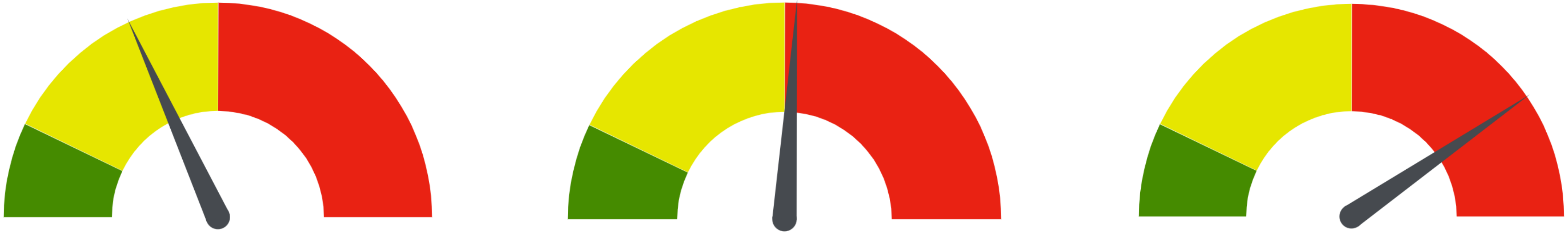
Model's performance





How to present the risk





tOwaRds Informed dEcisions iN colorecTal cancer screening

Thank you!

Questions?

You can contact us by e-mail:
daiane.seibert@thomasmore.be
karen.feyen@thomasmore.be

ORIENT
tOwaRds Informed dEcisions iN colorecTal cancer screening

THOMAS
MORE