06 July 2023 – PFIA, Health and AI day

Towards Trustworthy-AI-by-Design Methodology for Intelligent Radiology System _

Clotilde Brayé[1,2,3], Jérémy Clech[1], Arnaud Gotlieb[3], Nadjib Lazaar[2], Patrick Malléa[1]

nehs DIGITAL[1] LIRMM[2] simula[3]

Upcoming European AI regulation

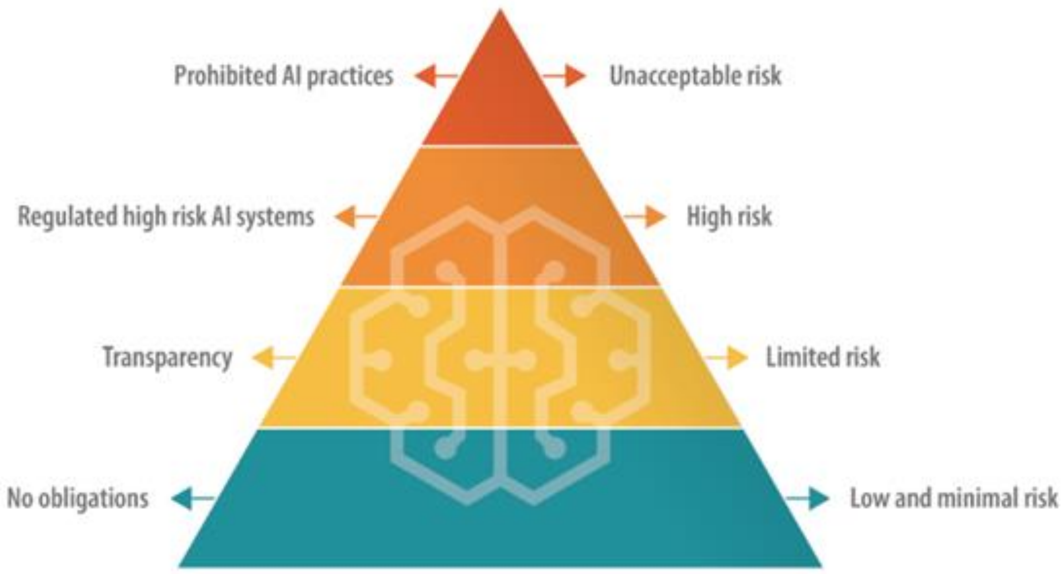
Ethical values and risk-based regulation

- 7 key requirements for lawful, ethical and robust AI

Trustworthy AI Requirements (TAIR)	
TAIR$_1$	Human agency and oversight
TAIR$_2$	Technical robustness and safety
TAIR$_3$	Privacy and data gouvernance
TAIR$_4$	Transparency
TAIR$_5$	Diversity, non-discrimination and fairness
TAIR$_6$	Societal and environnemental wellbeing
TAIR$_7$	Accountability

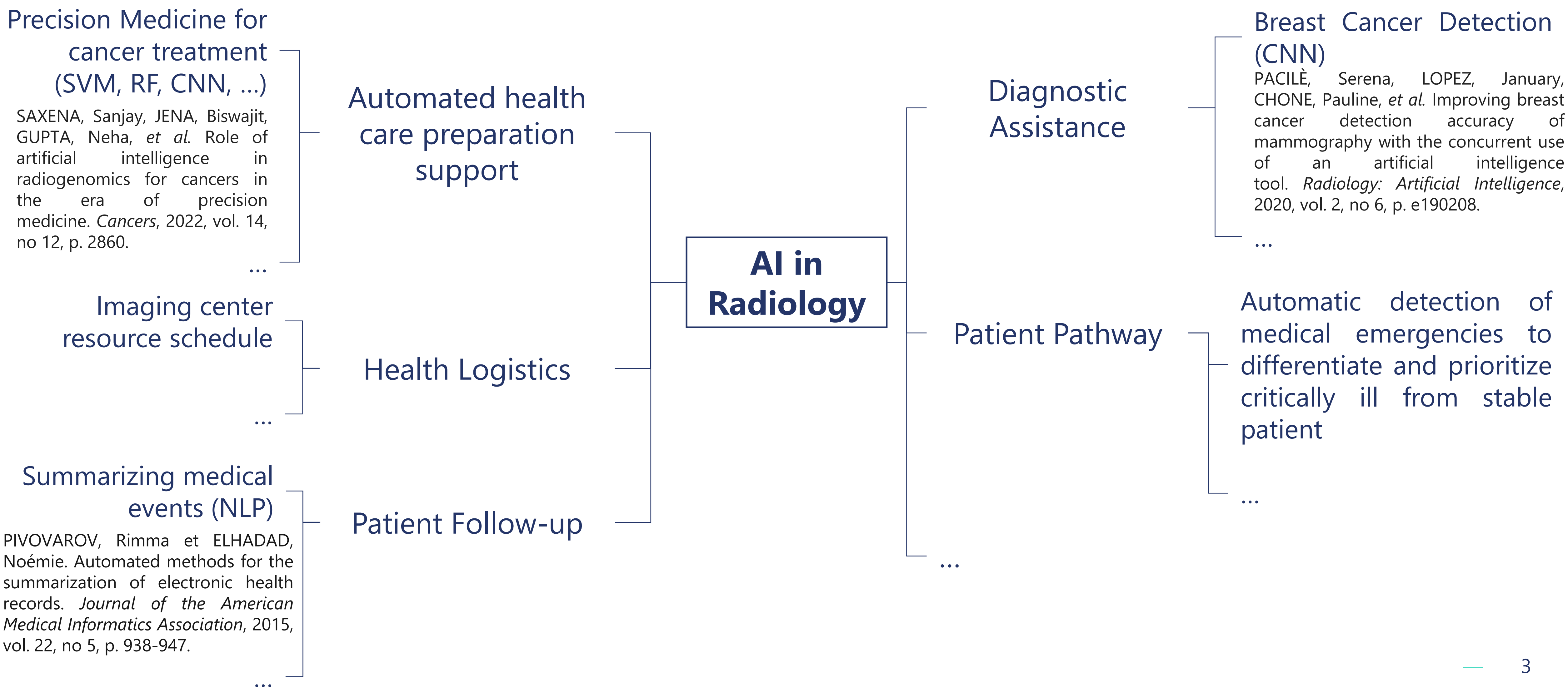
High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. *Publications Office of the European Union*, 2019

- Obligations depend on the risk level



MADIEGA, Tambiama André. Artificial intelligence act. European Parliament: *European Parliamentary Research Service*, 2021.

AI has expanded into all areas of radiology



AI in Radiology

Automated health care preparation support

Precision Medicine for cancer treatment (SVM, RF, CNN, …)

SAXENA, Sanjay, JENA, Biswajit, GUPTA, Neha, *et al.* Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine. *Cancers*, 2022, vol. 14, no 12, p. 2860.

…

Health Logistics

Imaging center resource schedule

…

Patient Follow-up

Summarizing medical events (NLP)

PIVOVAROV, Rimma et ELHADAD, Noémie. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 2015, vol. 22, no 5, p. 938-947.

…

Diagnostic Assistance

Breast Cancer Detection (CNN)

PACILÈ, Serena, LOPEZ, January, CHONE, Pauline, *et al.* Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiology: Artificial Intelligence*, 2020, vol. 2, no 6, p. e190208.

…

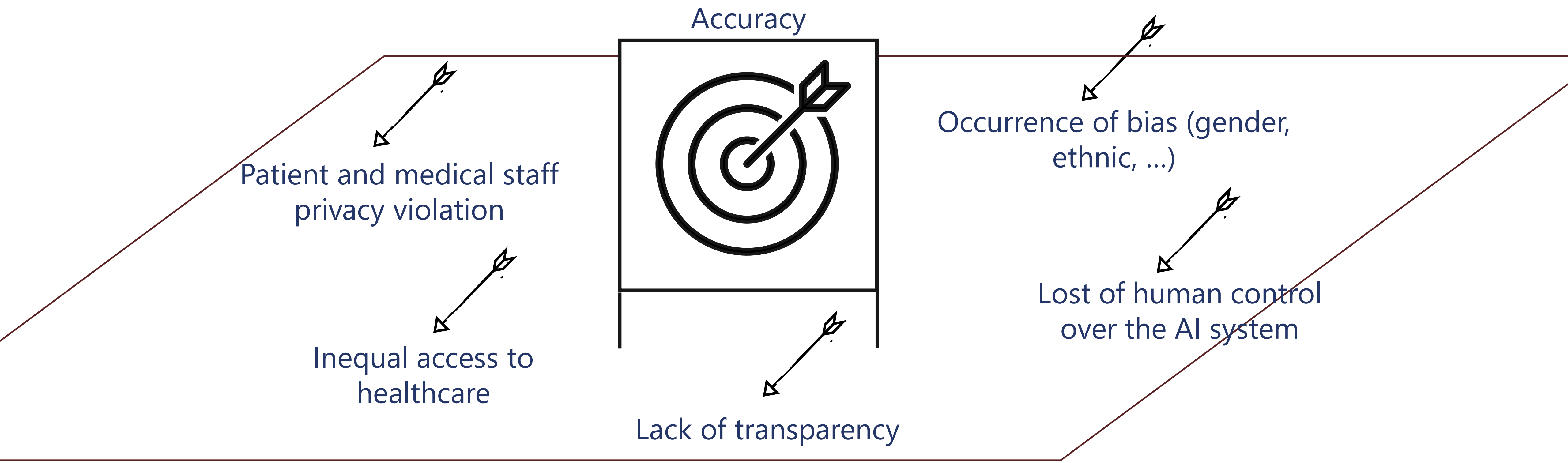
Patient Pathway

Automatic detection of medical emergencies to differentiate and prioritize critically ill from stable patient

…

…

PERFORMANCE-DRIVEN APPROACH



Accuracy

Patient and medical staff privacy violation

Inequal access to healthcare

Lack of transparency

Occurrence of bias (gender, ethnic, ...)

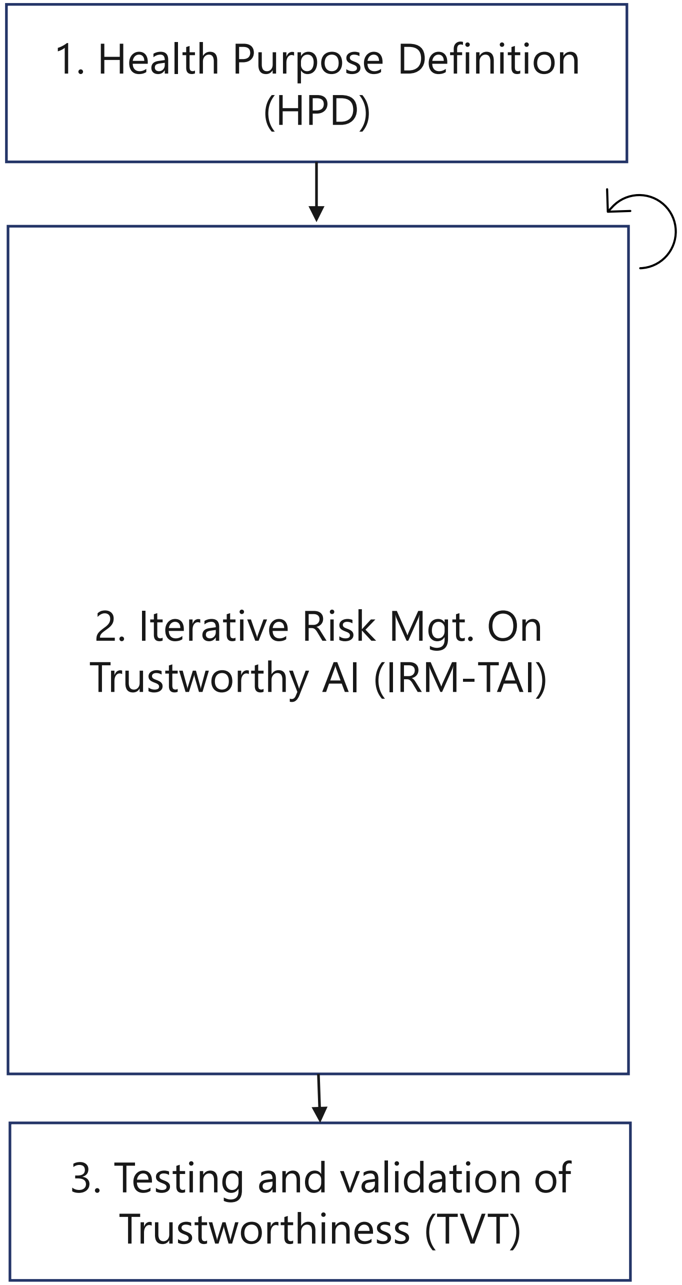
Lost of human control over the AI system

➢ Trustworthy-AI-by-Design

WU, Eric, WU, Kevin, DANESHJOU, Roxana, *et al.* How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 2021, vol. 27, no 4, p. 582-584.
RAJPURKAR, Pranav et LUNGREN, Matthew P. The Current and Future State of AI Interpretation of Medical Images. *New England Journal of Medicine*, 2023, vol. 388, no 21, p. 1981-1990.
SEYYED-KALANTARI, Laleh, ZHANG, Haoran, MCDERMOTT, Matthew BA, *et al.* Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 2021, vol. 27, no 12, p. 2176-2182.

TAID Methodology

- **TAID** for **T**rustworthy-**AI**-by-**D**esign Methodology
- TAID goal: minimise risks according to the 7 trustworthy AI requirements
- **Three-steps methodology** to **assess** AI system **risks** based on risk management*
- Give a **qualitative** evaluation of every choice regarding the AI system

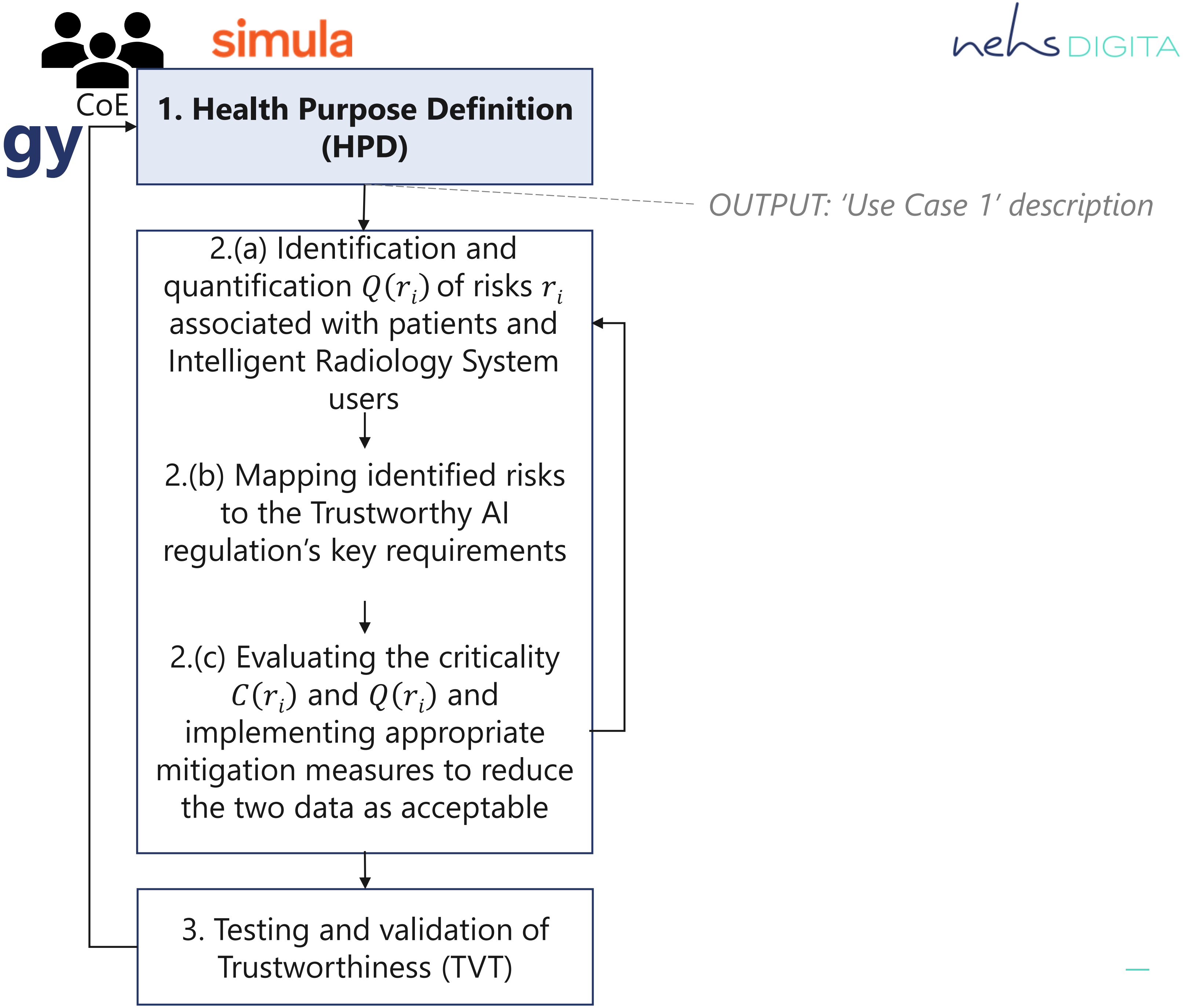

1. Health Purpose Definition (HPD)

2. Iterative Risk Mgt. On Trustworthy AI (IRM-TAI)

3. Testing and validation of Trustworthiness (TVT)

* ISO 14971 “Medical Devices - Application of risk management to medical devices”

TAID Methodology



CoE

1. Health Purpose Definition (HPD)

OUTPUT: 'Use Case 1' description

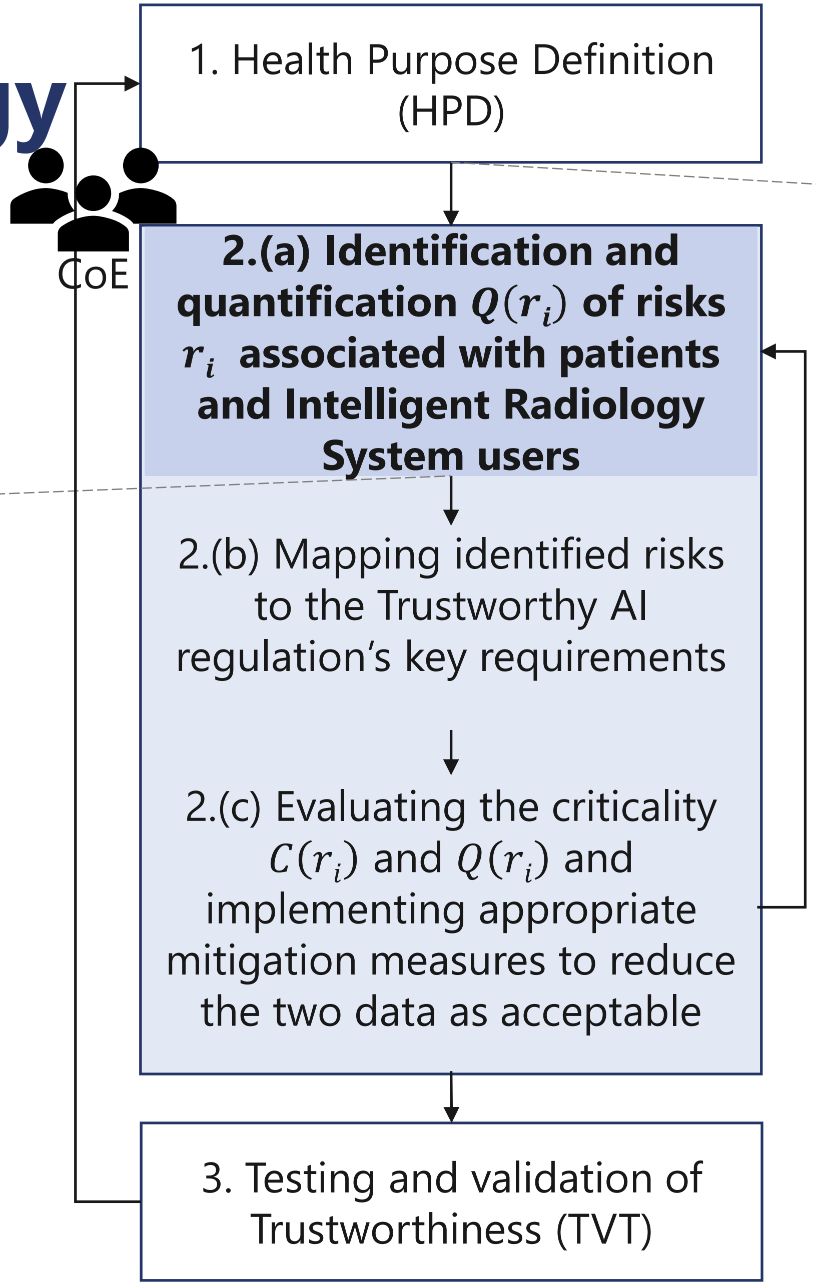
2.(a) Identification and quantification $Q(r_i)$ of risks r_i associated with patients and Intelligent Radiology System users

2.(b) Mapping identified risks to the Trustworthy AI regulation's key requirements

2.(c) Evaluating the criticality $C(r_i)$ and $Q(r_i)$ and implementing appropriate mitigation measures to reduce the two data as acceptable

3. Testing and validation of Trustworthiness (TVT)

TAID Methodology



1. Health Purpose Definition (HPD)

OUTPUT: 'Use Case 1' description

CoE

2.(a) Identification and quantification $Q(r_i)$ of risks r_i associated with patients and Intelligent Radiology System users

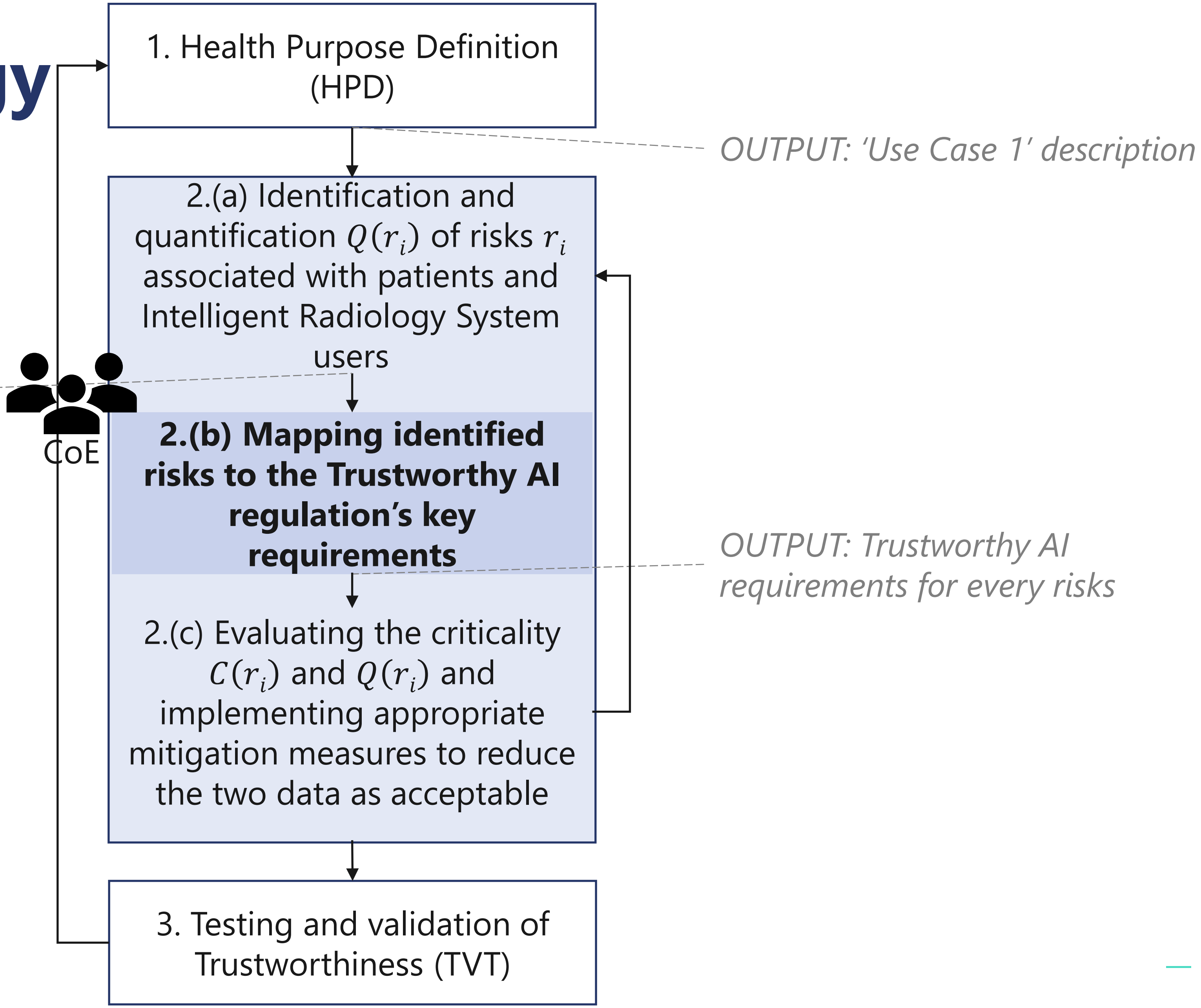
OUTPUT: List of risks and their quantification

2.(b) Mapping identified risks to the Trustworthy AI regulation's key requirements

2.(c) Evaluating the criticality $C(r_i)$ and $Q(r_i)$ and implementing appropriate mitigation measures to reduce the two data as acceptable

3. Testing and validation of Trustworthiness (TVT)

TAID Methodology



1. Health Purpose Definition (HPD)

OUTPUT: 'Use Case 1' description

2.(a) Identification and quantification $Q(r_i)$ of risks r_i associated with patients and Intelligent Radiology System users

OUTPUT: List of risks and their quantification

CoE

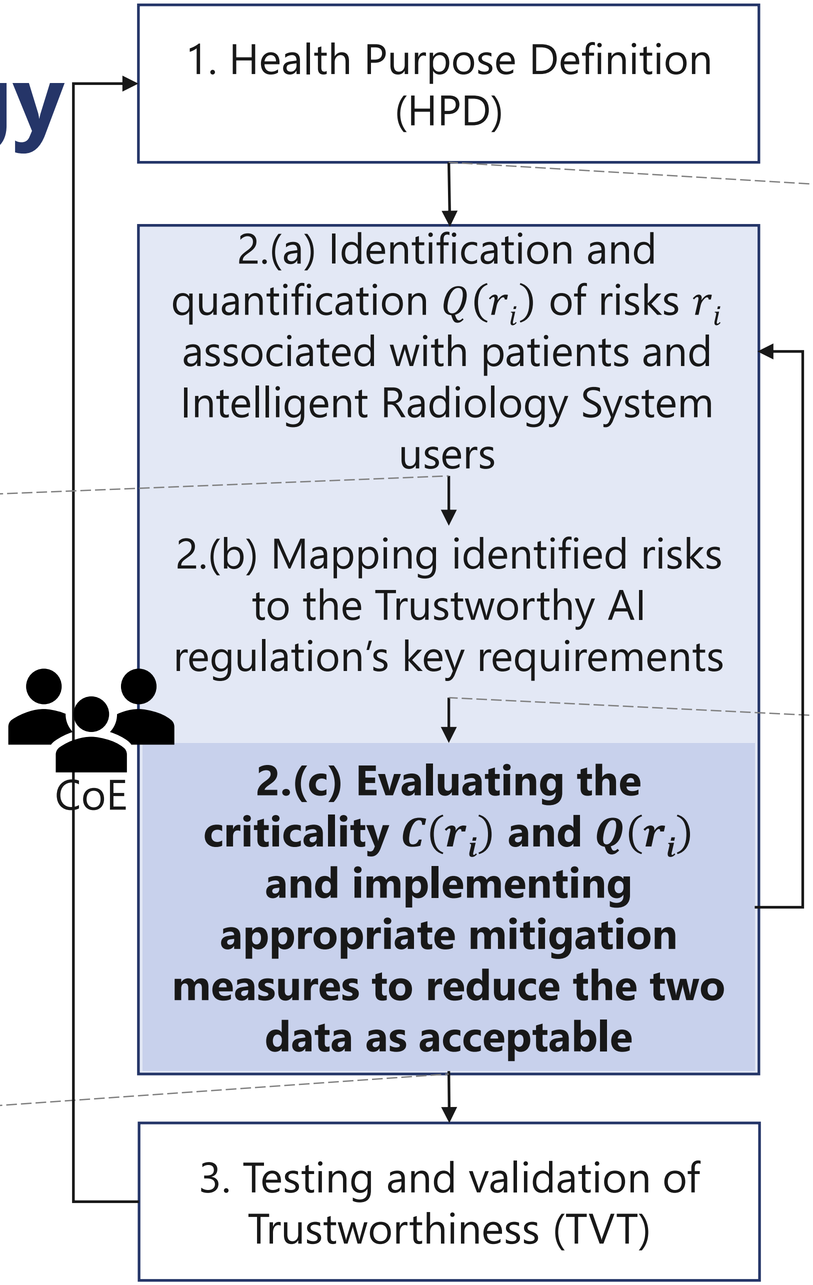
2.(b) Mapping identified risks to the Trustworthy AI regulation's key requirements

OUTPUT: Trustworthy AI requirements for every risks

2.(c) Evaluating the criticality $C(r_i)$ and $Q(r_i)$ and implementing appropriate mitigation measures to reduce the two data as acceptable

3. Testing and validation of Trustworthiness (TVT)

TAID Methodology



1. Health Purpose Definition (HPD)

2.(a) Identification and quantification $Q(r_i)$ of risks r_i associated with patients and Intelligent Radiology System users

2.(b) Mapping identified risks to the Trustworthy AI regulation's key requirements

2.(c) Evaluating the criticality $C(r_i)$ and $Q(r_i)$ and implementing appropriate mitigation measures to reduce the two data as acceptable

3. Testing and validation of Trustworthiness (TVT)

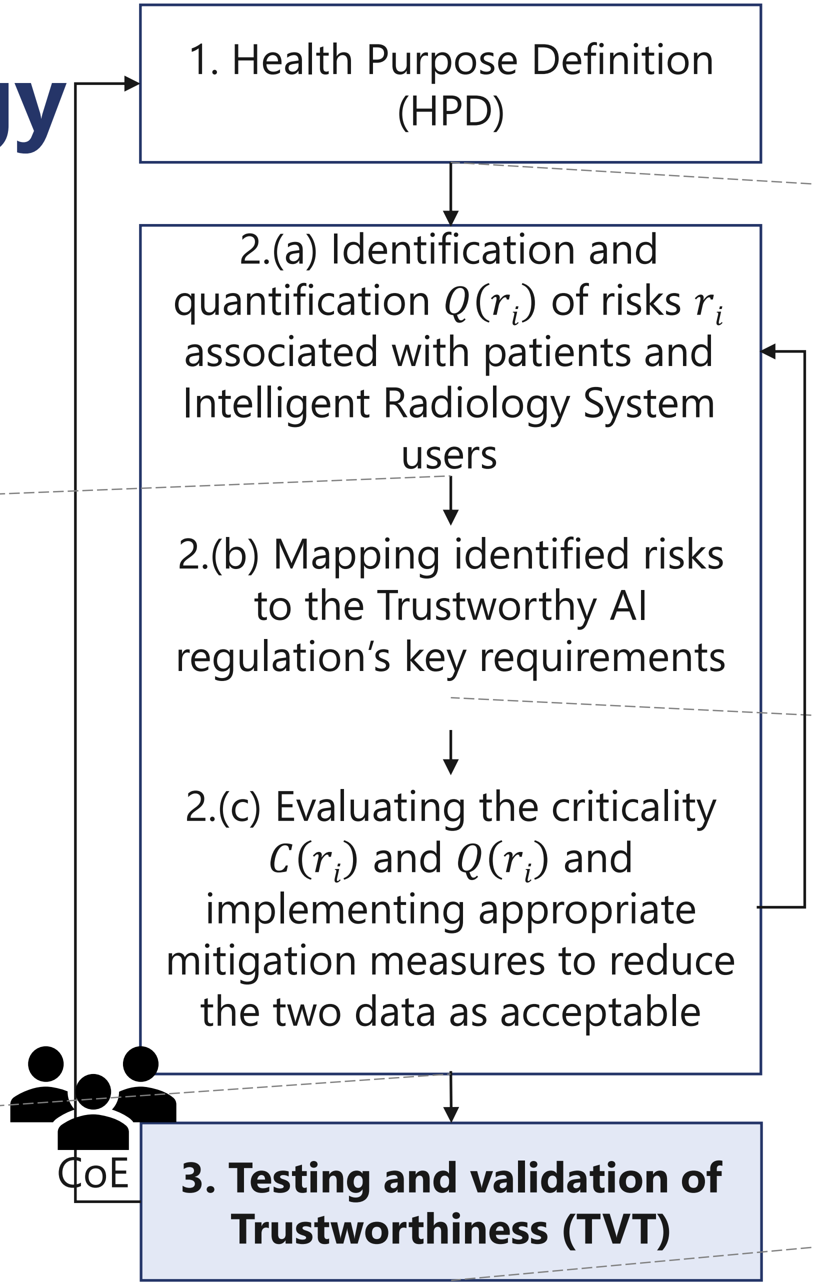
OUTPUT: 'Use Case 1' description

OUTPUT: List of risks and their quantification

OUTPUT: Trustworthy AI requirements for every risks

OUTPUT: List of risks, criticality and mitigations actions

TAID Methodology



1. Health Purpose Definition (HPD)

OUTPUT: 'Use Case 1' description

2.(a) Identification and quantification $Q(r_i)$ of risks r_i associated with patients and Intelligent Radiology System users

OUTPUT: List of risks and their quantification

2.(b) Mapping identified risks to the Trustworthy AI regulation's key requirements

OUTPUT: Trustworthy AI requirements for every risks

2.(c) Evaluating the criticality $C(r_i)$ and $Q(r_i)$ and implementing appropriate mitigation measures to reduce the two data as acceptable

OUTPUT: List of risks, criticality and mitigations actions

CoE

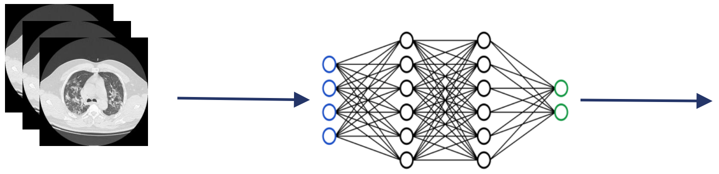
3. Testing and validation of Trustworthiness (TVT)

OUTPUT: Tests Report

Evaluation on two different use cases

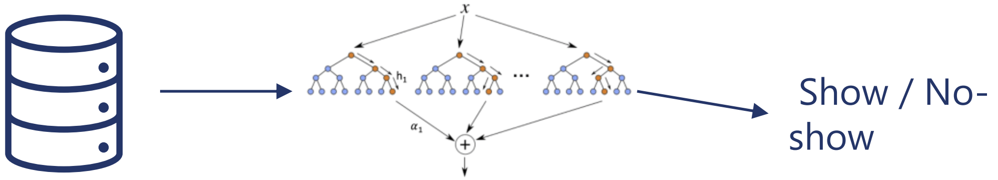
High-risk Systems (AI Act)

- ‘FIDAC’ : automatically detect COVID-19 on CT-scans using CNN



Low and minimal risk (AI Act)

- ‘NOSHOW’ : use DT/RF to estimate the likelihood of a patient attending a radiology appointment

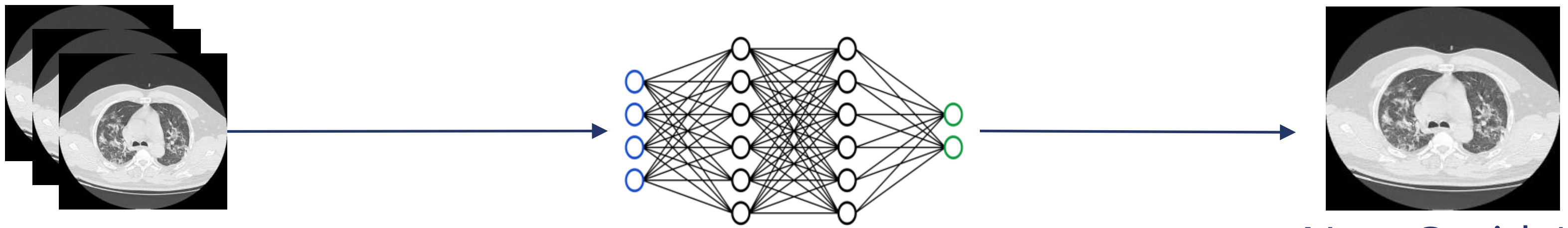


UC	Risks	Risk Description	TAIR_i Initial	TAIR_i Residual	Initial Criticality	Residual Criticality
FIDAC	F_R1	Personal data breaches	TAIR2, TAIR3	TAIR3	13.71	0.14
FIDAC	F_R2	Lack of explicability of the prediction	TAIR1, TAIR4, TAIR7	TAIR4, TAIR7	11.57	5.14
FIDAC	F_R3	Model attacks	TAIR2	TAIR2	2.57	2.57
FIDAC	F_R4	Wrong patient care	TAIR1, TAIR2, TAIR6	TAIR1, TAIR2, TAIR6	20.57	7.71
FIDAC	F_R5	Differences of performance depending on age or gender	TAIR2, TAIR3, TAIR5	TAIR2, TAIR3, TAIR5	7.71	3.86

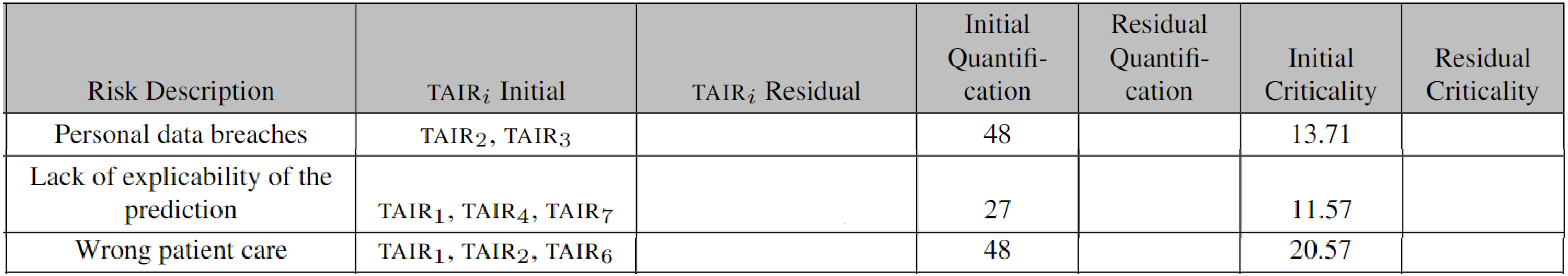
UC	Risks	Risk Description	TAIR_i Initial	TAIR_i Residual	Initial Criticality	Residual Criticality
NOSHOW	N_R1	Personal data breaches	TAIR2, TAIR3	TAIR3	7.71	0.14
NOSHOW	N_R2	Lack of explicability of the prediction	TAIR1, TAIR2, TAIR4, TAIR5, TAIR7	TAIR4, TAIR7	19.28	2.57
NOSHOW	N_R3	Model attacks	TAIR2	TAIR2	2.57	2.57
NOSHOW	N_R4	Patient categorisation	TAIR1,TAIR4, TAIR5	TAIR1, TAIR4, TAIR5	20.57	7.71
NOSHOW	N_R5	Excessive patient reminders	TAIR1, TAIR2, TAIR3, TAIR4, TAIR5	TAIR2, TAIR4, TAIR5	8.57	3.42
NOSHOW	N_R6	Disorganization of the center	TAIR1, TAIR2	TAIR2	5.14	1.28
NOSHOW	N_R7	Deterioration of the facility’s image	TAIR2, TAIR7	TAIR2, TAIR7	2.57	1.14
NOSHOW	N_R8	Inability of the facility to complete the planned medical exam	TAIR1	TAIR1	6.86	0.14
NOSHOW	N_R9	Equal access to healthcare	TAIR1	TAIR1	6.86	0.14

Evaluation on two different use cases

‘FIDAC’ : automatically detect COVID-19 on CT-scans using CNN



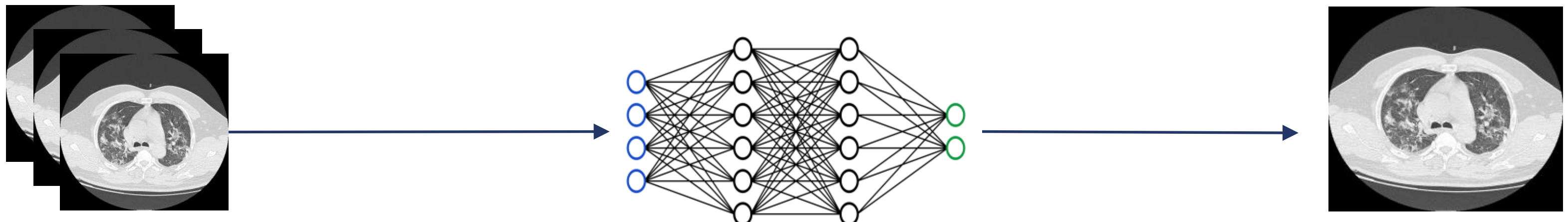
Non-Covid / Covid



Risk Description	TAIR$_i$ Initial	TAIR$_i$ Residual	Initial Quantification	Residual Quantification	Initial Criticality	Residual Criticality
Personal data breaches	TAIR$_2$, TAIR$_3$		48		13.71	
Lack of explicability of the prediction	TAIR$_1$, TAIR$_4$, TAIR$_7$		27		11.57	
Wrong patient care	TAIR$_1$, TAIR$_2$, TAIR$_6$		48		20.57	

Evaluation on two different use cases

‘FIDAC’ : automatically detect COVID-19 on CT-scans using CNN



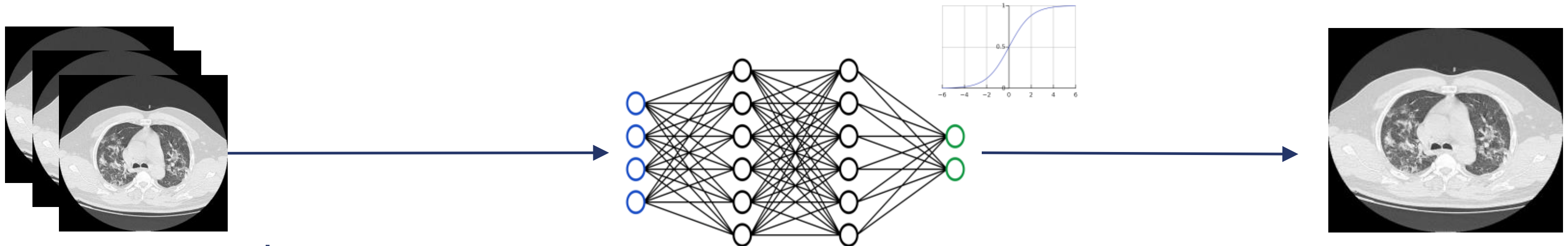
Anonymous Database

Non-Covid / Covid

Risk Description	TAIR$_i$ Initial	TAIR$_i$ Residual	Initial Quantification	Residual Quantification	Initial Criticality	Residual Criticality
Personal data breaches	TAIR2, TAIR3	TAIR3	48	1	13.71	0.14
Lack of explicability of the prediction	TAIR1, TAIR4, TAIR7		27		11.57	
Wrong patient care	TAIR1, TAIR2, TAIR6		48		20.57	

Evaluation on two different use cases

'FIDAC' : automatically detect COVID-19 on CT-scans using CNN



Anonymous Database

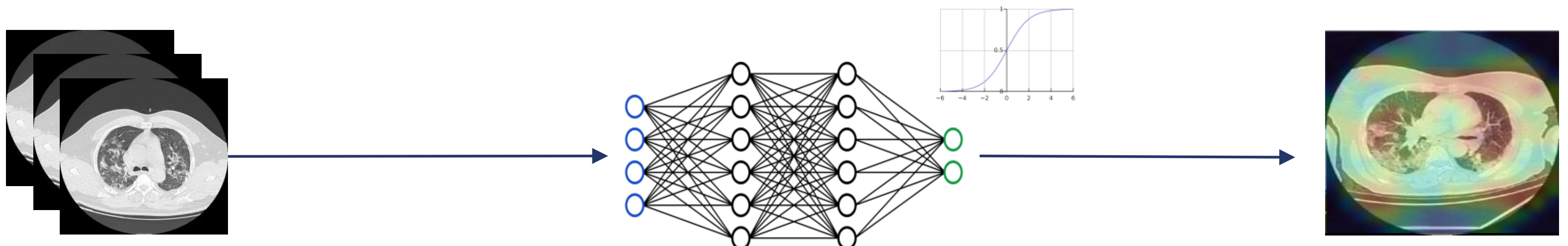
ResNet-50

Non-Covid / Covid

Risk Description	TAIR$_i$ Initial	TAIR$_i$ Residual	Initial Quantification	Residual Quantification	Initial Criticality	Residual Criticality
Personal data breaches	TAIR$_2$, TAIR$_3$	TAIR$_3$	48	1	13.71	0.14
Lack of explicability of the prediction	TAIR$_1$, TAIR$_4$, TAIR$_7$	TAIR$_4$, TAIR$_7$	27	18	11.57	5.14
Wrong patient care	TAIR$_1$, TAIR$_2$, TAIR$_6$		48		20.57	

Evaluation on two different use cases

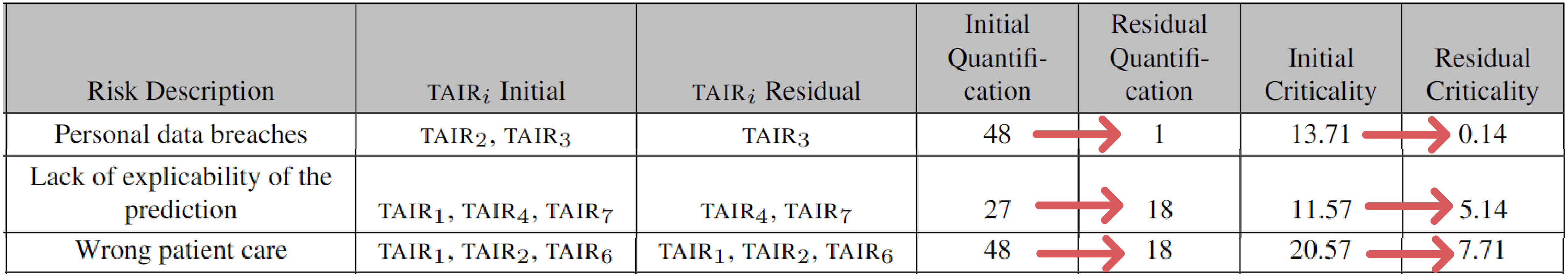
'FIDAC' : automatically detect COVID-19 on CT-scans using CNN



Anonymous Database

ResNet-50

Non-Covid / Covid and heatmap from Grad-CAM algorithm



Risk Description	TAIR$_i$ Initial	TAIR$_i$ Residual	Initial Quantification	Residual Quantification	Initial Criticality	Residual Criticality
Personal data breaches	TAIR2, TAIR3	TAIR3	48	1	13.71	0.14
Lack of explicability of the prediction	TAIR1, TAIR4, TAIR7	TAIR4, TAIR7	27	18	11.57	5.14
Wrong patient care	TAIR1, TAIR2, TAIR6	TAIR1, TAIR2, TAIR6	48	18	20.57	7.71

Initial limitations identified

θ?
η?

Arbitrarily selected thresholds

Missing risks ?

Mitigation measures may not be sufficient to reduce the risk

Conclusion…

- TAID offers a comprehensive framework for managing AI-related risks addressing all the seven trustworthy AI requirements during life-cycle of the AI system

- Risks identification and mitigation actions are similar for both use cases

… and future work

- Refine “ Test and validation of the trusworthiness” part

- Tradeoff between risk reduction and model performance

- Assess the deployment of TAID methodology

LIRMM[2] simula[3] nehs DIGITAL[1]

“Towards Trustworthy-AI-by-Design Methodology for Intelligent Radiology System” _

Clotilde Brayé[1,2,3], Jérémy Clech[1], Arnaud Gotlieb[3], Nadjib Lazaar[2], Patrick Malléa[1]

Thank you for your attention!