Towards Trustworthy and Equitable Healthcare AI: Defining Fairness Domains and Metrics for Consensus

Hyun Jin Han[1,2], Shinyoung Park[3,4], Soomi Jo [3,4], Hae Sun Suh [1, 3, 4*].

1 College of Pharmacy, Kyung Hee University, Seoul, Republic of Korea
2 Vista Health Korea Ltd. Seoul, Republic of Korea
3 Department of Regulatory Science, Graduate School, Kyung Hee University, Seoul, Republic of Korea
4 Institute of Regulatory Innovation through Science (IRIS), Kyung Hee University, Seoul, Republic of Korea

*** Corresponding author**

PEBD
Pharmaceutical Economics
Big Data Analysis and Policy Lab

BACKGROUND

- Regulatory agencies such as Korea's MFDS, the U.S. FDA, and the EMA have issued guidance addressing transparency, accountability, and patient safety. However, limited alignment across agencies poses challenges to the global AI evaluation and regulatory adoption of AI.
- AI is increasingly applied in healthcare, enhancing diagnosis, treatment, and resource allocation. Concerns regarding fairness and equity remain, as AI may exacerbate existing disparities, and no consensus has been reached on the domains and metrics for assessing AI fairness.

OBJECTIVES

- This study synthesizes the literature on fairness in healthcare AI through an umbrella review and examines regulatory perspectives on trustworthy AI, with the aim of proposing a collaborative framework.

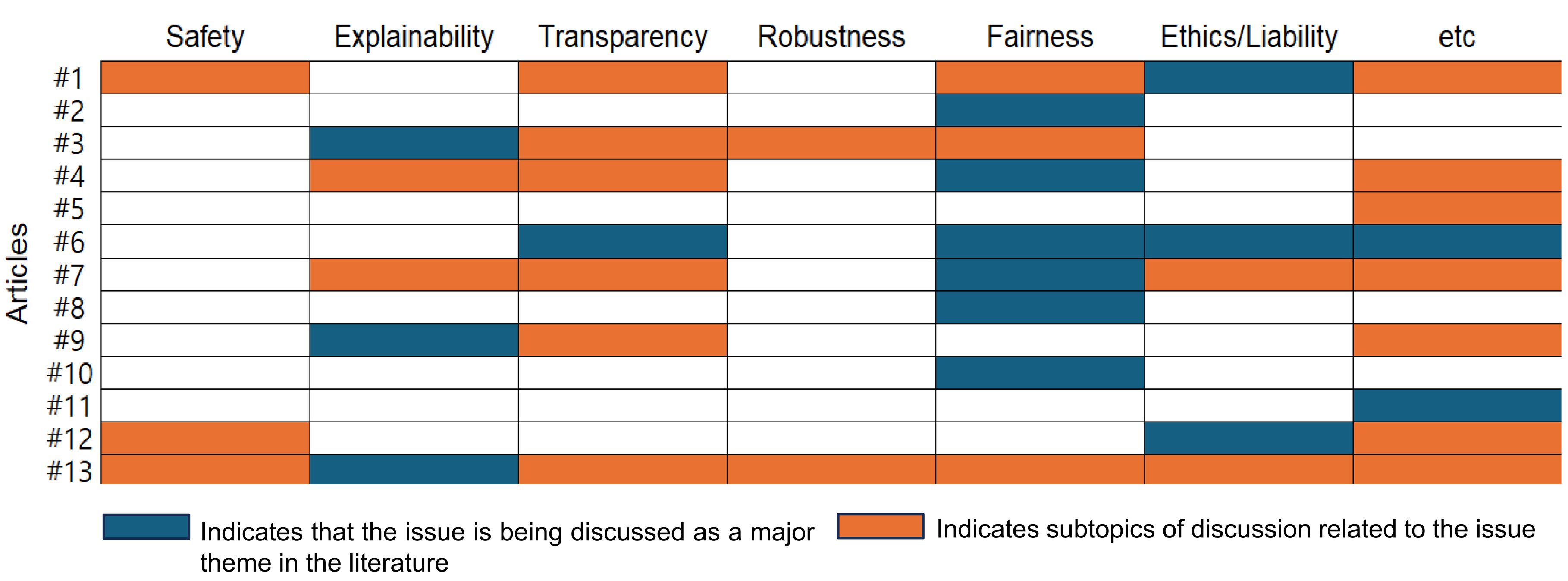
METHODS

- **Design**: Umbrella review of systematic reviews or scoping reviews, along with a review of regulatory publications.
- **Databases**: MEDLINE, CENTRAL, EMBASE, IEEE Xplore, ACM Digital Library (through July 2024).
- **Keywords**: AI, machine learning, deep learning, supervised/unsupervised learning, reinforcement learning, trustworthiness, fairness, and health.
- **Analysis**: Data charted by five trustworthiness properties—fairness, explainability, transparency, safety, and robustness. Systematic/scoping reviews on fairness were synthesized.
- **Regulatory review**: MFDS (Korea), FDA (U.S.), and EMA (Europe) were analyzed using keywords *'artificial intelligence', 'trustworthy',* and *'medical'* through descriptive and thematic analyses.

RESULTS

1. Main trustworthiness domains and fairness consideration

- From 933 records identified, 98 studies were initially selected according to the inclusion and exclusion criteria. After full-text review, 85 studies were excluded, leaving **13 studies as the final** candidates.
- Fairness was the most commonly identified major theme across the reviewed records.



2. Domains and metrics for fairness evaluation of AI for healthcare

Domain	Causes	Examples of metrics
Algorithmic fairness	Group fairness	Demographic parity, Equal opportunity, Equal odds
	Individual fairness	Generally defined distance metric, Domain-specific distance metric, Learned distance metric
	Label bias	Cohen's Kapa, Fleiss' Kapp, Krippendorf's Alpha
Data Fairness	Minority bias	Disparity impact ratio, Equal opportunity, Equal odds, Subgroup accuracy, Representation ratios
	Missing data bias	Missing rate by group, Missing data imbalance ratio, Little's MCAR test, Imputation error by group
	Informativeness bias	Mutual information, Group-wise feature importance shift
	Measurement bias	Cohen's Kappa, Fleiss' Kappa, Krippendorf's Alpha
	Temporal bias	Statistical parity difference, Equal opportunity difference, Disparate impact ratio, Calibration by group
Human Cognitive bias	Group attribution bias	Data drift over time by subgroup, Prediction drift per group, Exposure inequality index
	Implicit bias	Accuracy, Equal opportunity, Equal odds, Predictive equity, Calibration-in-the-large, False negative rate parity
	In-group bias	Mutual information, Feature attribution
	automation bias	Disagreement impact rate, Override rate by ground truth, User trust differential
	Feedback loop	Data drift over time by subgroup, Exposure inequality index, Prediction Drift per Group
Interaction-related biases	Rejection bias	Selection-Induced Label Bias, Missing label rate by group, counterfactual outcome disparity
	Privilege bias	Benefit distribution, Welfare parity, Opportunity imbalance index
	Informed mistrust	Adoption rate by group, intervention compliance gap, Trust calibration curve
	Agency bias	User override rate by group, Actionability disparity
	Intersectional fairness	Worst-case disparity, Pairwise evaluation

3. Regulatory perspectives on the trustworthiness and fairness of AI

Organization	Perspectives and recommendations
ISO & IEC	• **ISO/IEC TR 24028:2020**: Describes AI system trustworthiness as follows: ① Stakeholder concerns regarding AI and data use must be addressed in a transparent and accessible manner. ② Trustworthiness of AI systems depends on their technical robustness, controllability, and verifiability throughout the entire lifecycle. • **ISO/IEC 22989:2022**: Identifies components of AI trustworthiness—robustness, reliability, resilience, controllability, predictability, transparency, bias, and fairness. Discusses bias adjustment as a means to achieve AI fairness.
ITU & WHO	• **FG-AI4H**: Emphasizes that governments and WHO/ITU member states hold responsibility for the governance of healthcare services, while stakeholders developing and operating AI-based systems are responsible for ensuring their proper functioning. • **FG-AI4H DEL0.1**: Defines *Trustworthy AI* as AI that meets stakeholder expectations regarding characteristics such as bias, explainability, and provenance. Defines *fairness* as ensuring that all individuals are treated equitably, without discrimination, neglect, manipulation, domination, or abuse.
TTA	• **023 Guidelines for Developing Trustworthy AI – Healthcare Sector**: Identifies diversity, accountability, safety, and transparency as requirements for AI trustworthiness. • **Diversity**: Defined as the absence of discriminatory or biased practices in AI learning processes and outputs affecting individuals or groups, and the assurance of equal access to AI benefits, regardless of factors such as race, gender, or age. • **Related attributes**: Fairness, justice • **Related keywords**: Bias, discrimination, prejudice, diversity, equality
MFDS	• **Trustworthiness and Fairness**: Emphasizes the need to consider the possibility of inaccurate outputs caused by erroneous or biased data, as well as the reproducibility of outputs. • **Guidelines for Approval and Review of AI-based Medical Devices**: Define *bias* as any systematic difference in how specific objects, individuals, or groups are treated compared with a control group. This includes all actions such as perception, observation, representation, prediction, or decision.
FDA	• **Suggestion on trustworthy AI**: Recommends including diverse demographic groups in datasets, ensuring explainability, addressing bias, and securing transparency.
EMA	• **Component of trustworthy AI assessment**: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Accountability, Societal and environmental well-being, Diversity, Non-discrimination, and Fairness.

Abbreviation: ISO, International Organization for Standardization; IEC, Electrotechnical Commission; ITU, International Telecommunication Union; WHO, World Health Organization; TTA, Telecommunications Technology Association; MFDS, Ministry of Food and Drug Safety; FDA, Food and Drug Administration; EMA, European Medicines Agency.

CONCLUSIONS

- Our findings reveal that fairness in healthcare AI encompasses four key dimensions, with regulatory bodies consistently emphasizing fairness and explainability.
- These insights provide a foundation for developing transparent, ethical, and equitable AI applications in healthcare, and underscore the need for a collaborative regulatory framework to address the challenges in AI fairness.

ACKNOWLEDGMENT

This research was supported by a grant(RS-2024-00331719, 21153MFDS601) from the Ministry of Food and Drug Safety in 2025.

First author: hyunjin.han@gmail.com/hhan@vista.health