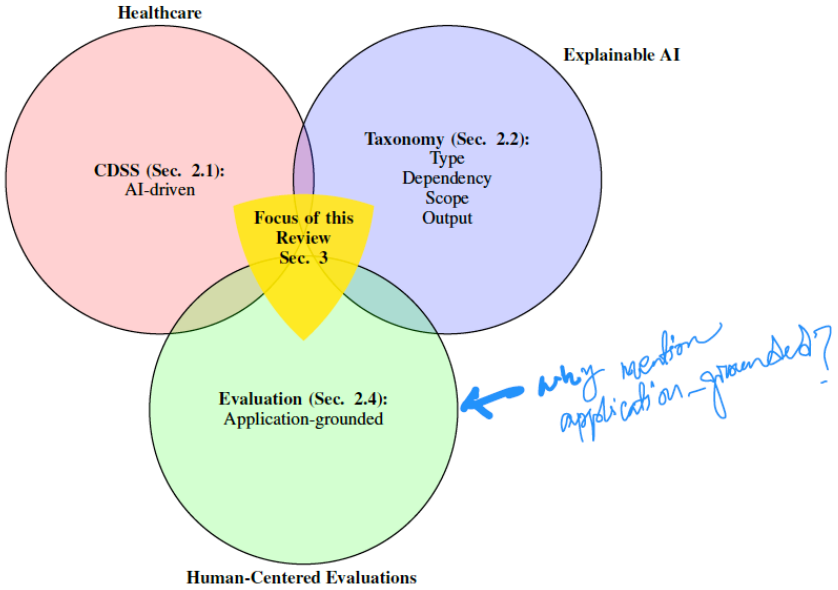
A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems

Alessandro Gambetti, Qiwei Han, Hong Shen, and Cláudia Soares

Presented by: Abdullah Mamun

Date: 29 April 2026

Email: a.mamun@asu.edu



Paper

abdullah-mamun.com

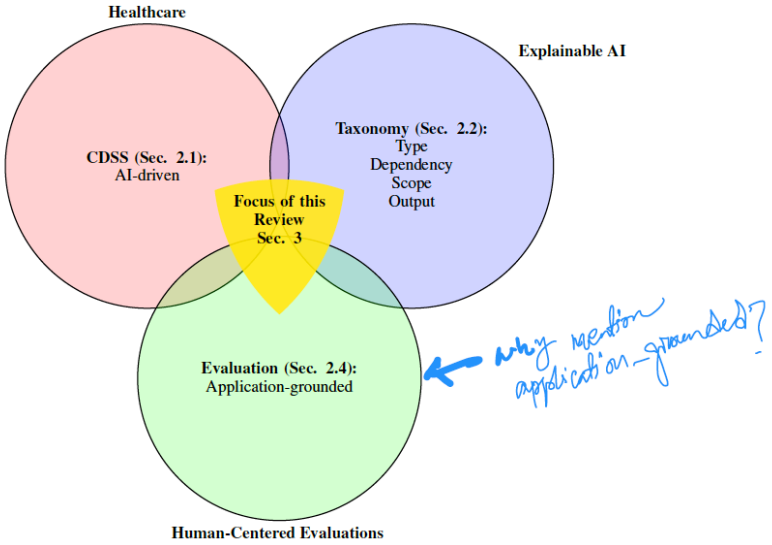
𝕏: @AB9Mamun

The Challenge: Trusting the Black Box in Medicine

Explainable AI (XAI) is essential for the adoption of Clinical Decision Support Systems (CDSS). While AI models can be highly accurate, clinicians are hesitant to trust them without understanding their reasoning, especially when patient lives are at stake. However, the real-world effectiveness of XAI methods is inconsistently evaluated.

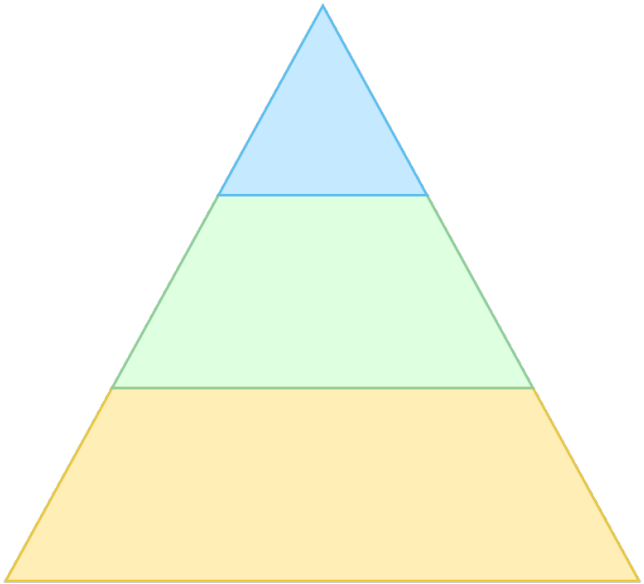
The Core Problem

- AI in CDSS offers huge potential, but 'black-box' models hinder trust.
- Diagnostic errors have life-threatening consequences.
- Regulatory frameworks like GDPR mandate explainability.



Background: Levels of Human-Centered Evaluation (HCE) – Based on Doshi-Velez and Kim’s position paper

Evaluating an XAI system isn't just about technical accuracy. Human-centered evaluations measure its real-world utility. There's a trade-off between realism and cost.



Application-Grounded (High Fidelity, High Cost): Real tasks with real domain experts (e.g., doctors diagnosing real cases). The gold standard for medical CDSS.

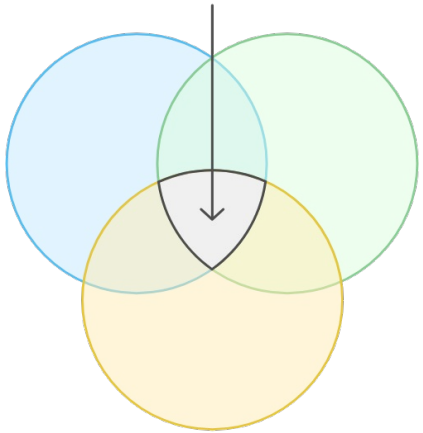
Human-Grounded (Medium Fidelity, Medium Cost): Simplified tasks with laypeople. Good for general usability but not for clinical specifics.

Proxy Evaluation (Low Fidelity, Low Cost): Automated tasks without human participants. Useful for early-stage prototypes but not reflective of real-world use.

Bridging the Gap: The Paper's Contribution

This paper presents a systematic survey of 31 human-centered evaluations (HCE) of XAI in CDSS to understand how explanations are assessed in practice. The goal is to identify gaps and propose a new framework for developing trustworthy and clinically effective XAI-based CDSS.

Focus of this Survey



Background: A Taxonomy of XAI Methods

This survey uses a taxonomy based on three key criteria:

Type: Intrinsic vs. Post Hoc

Is the model inherently transparent (like a decision tree), or does it require a separate method to explain its decision after the fact (like SHAP on a neural network)?

Dependency: Model-Specific vs. Model-Agnostic

Is the explanation method tied to a specific model architecture (e.g., GradCAM for CNNs), or can it be applied to any model type?

Scope: Local vs. Global

Does the explanation apply to a single prediction for one patient (local), or does it describe the model's overall behavior across all data (global)?

Survey Methodology & Scope

PRISMA Process

The authors conducted a systematic review using the PRISMA methodology. They screened over 3,000 published after 2017 (motivated by the DARPA XAI program).

(1) top-tier conference venue (CORE A or A) or SJR Q1 journal (PRISMA "Screening"); (2) publication date after 2017, motivated by the DARPA XAI Program announced in May 2017"*

- Databases: ACM, Web of Science, PubMed
- Final Selection: 31 relevant papers identified
- Focus: Studies with explicit XAI and HCE methodologies in a clinical context.

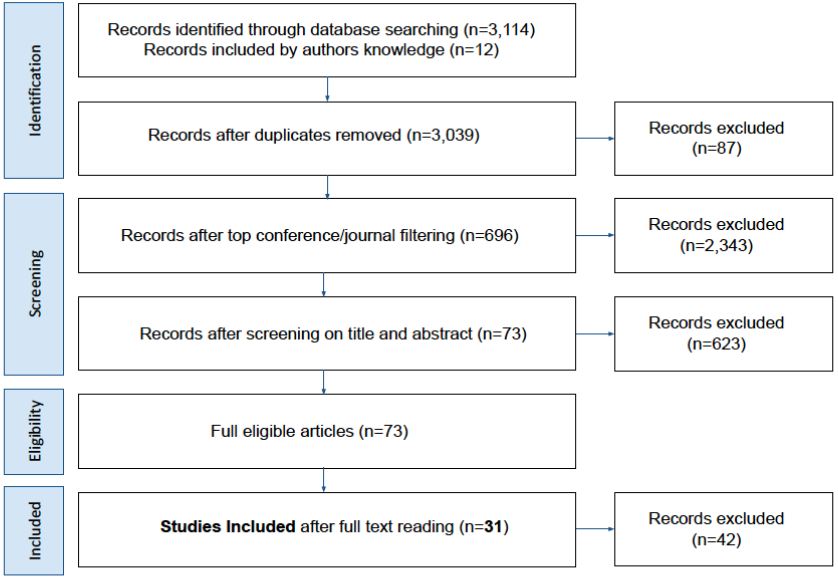


Figure 2: PRISMA Flow Diagram detailing the systematic process of identifying, screening, and selecting studies for inclusion.

Distribution of Research Over Time

The analysis of the 31 selected papers reveals a significant and growing interest in the human-centered evaluation of XAI in CDSS, with most publications appearing in just the last few years.

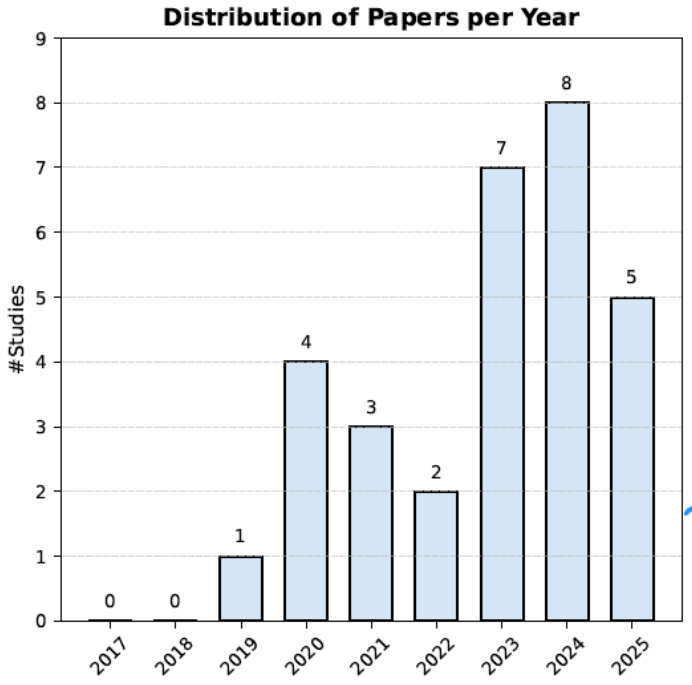


Figure 3: Histogram of retrieved papers per year.

All the papers and Methods

Table 1: Summary of Papers by ML Models, XAI Methods, XAI Taxonomy, HCE Methodologies, Medical Fields, and Clinician Perceptions.

Paper	Field	Model(s)	XAI(s)	XAI Taxonomy Sec. 2.2	HCE(s)	Perc.
[Barda *et al.*, 2020]	Critical Care	RF	SHAP	PH, AGN, LOC	FG	↑
[Brennan *et al.*, 2019]	Critical Care	GAM, RF, +	-	INT, LOC	TA	↑
[Ellenrieder *et al.*, 2023]	Radiology	CNN	I. Gradients	PH, SPEC, LOC	TA	↑
[Hwang *et al.*, 2022]	Sleep Medicine	CNN	Saliency	PH, INT, LOC	TA, I	→
[Kumar. *et al.*, 2020]	Critical Care	RF	LIME	PH, AGN, LOC	S	↑
[Kovalchuk *et al.*, 2022]	Endocrinology	RF, XGB, +	SHAP	PH, AGN, LOC	S, I, +	→
[Matthiesen *et al.*, 2021]	Cardiology	RF	LIME	PH, AGN, LOC	I	→
[Neves *et al.*, 2021]	Cardiology	KNN, CNN, +	LIME, +	PH, AGN, LOC	S	→
[Pumplun *et al.*, 2023]	Radiology	CNN	I. Gradients	PH, SPEC, LOC	S	↑
[Rajashekar *et al.*, 2024]	Gastroenterology	RF	PDP, ICE, ALE	PH, AGN, LOC/GLOB	I, S, +	↑
[Sabol *et al.*, 2020]	Oncology	CNN	X-CFCMC	PH, AGN, LOC	S	↑
[Singh *et al.*, 2021]	Ophthalmology	CNN	DeepTaylor	PH, AGN, LOC	S	↑
[Sivaraman *et al.*, 2023]	Critical Care	XGB	SHAP	PH, AGN, LOC	TA	→
[Zhang *et al.*, 2024]	Critical Care	LSTM	Attention	INT, LOC	I	→
[Bienefeld *et al.*, 2023]	Critical Care	Tree	SHAP	PH, AGN, LOC	S, FG+	→
[Bhattacharya *et al.*, 2023]	Endocrinology	Logit	SHAP	PH, AGN, LOC	I, S+	→
[Abraham *et al.*, 2023]	Critical Care	-	SHAP	PH, AGN, LOC	I	↑
[Cabitza *et al.*, 2025]	Radiology	CNN	CAM	PH, SPEC, LOC	S	→
[Chanda *et al.*, 2024]	Dermatology	CNN	GradCAM	PH, SPEC, Local	S	↑
[Chari *et al.*, 2023]	Endocrinology	BERT	SHAP	PH, AGN, LOC	TA	↑
[Famiglini *et al.*, 2024]	Radiology	CNN	CAM	PH, SPEC, LOC	S	↑
[Gombolay *et al.*, 2024]	Neurology	DT	-	INT, LOC	S	→
[Gu *et al.*, 2020]	Oncology	XGBoost	SHAP	PH, AGN, LOC	S	↑
[He *et al.*, 2024]	Critical Care	DT, RF, +	SHAP	PH, AGN, LOC/GLOB	TA	↑
[Hur *et al.*, 2025]	Critical Care	Boosting	SHAP	PH, AGN, LOC	S	↑
[E. Ihongbe *et al.*, 2024]	Radiology	CNN	LIME, +	PH, AGN/SPEC, LOC	S	↑
[Jing *et al.*, 2025]	Geriatrics	XGB	SHAP	PH, AGN, LOC	S	↑
[Jung *et al.*, 2025]	Critical Care	DT, RF, +	SHAP	PH, AGN, LOC	I	↑
[Kayadibi *et al.*, 2025]	Dentistry	CNN	LIME	PH, AGN, LOC	S	↑
[Rainey *et al.*, 2024]	Radiology	CNN	GradCAM	PH, SPEC, LOC	S	→
[Singla *et al.*, 2023]	Radiology	CNN	Counterf.	PH, SPEC, LOC	S	↑

Note—"↑" indicates a Positive clinician perception; "→" indicates a Neutral perception; "↓" indicates a Negative perception.
In Model, XAI and HCE, "+" indicates more methods used. Abbreviations: FG: Focus Group; TA: Think-Aloud; I: Interview; S: Survey;
RF: Random Forest; CNN: Convolutional Neural Network; XGB: XGBoost; DT Decision Tree; Logit: Logistic Regression; GAM: Generalized Additive Model; KNN: K-Nearest Neighbor; LSTM: Long Short-Term Memory; BERT: Bidirectional Encoder Representations from Transformers;
PH: Post Hoc; AGN: Agnostic, SPEC: Specific; INT: Intrinsic; LOC: Local; GLOB: Global.

Aggregate Findings: A Snapshot of the Field

Across the 31 studies, clear patterns emerge in the choice of XAI methods, evaluation techniques, and clinical application areas.

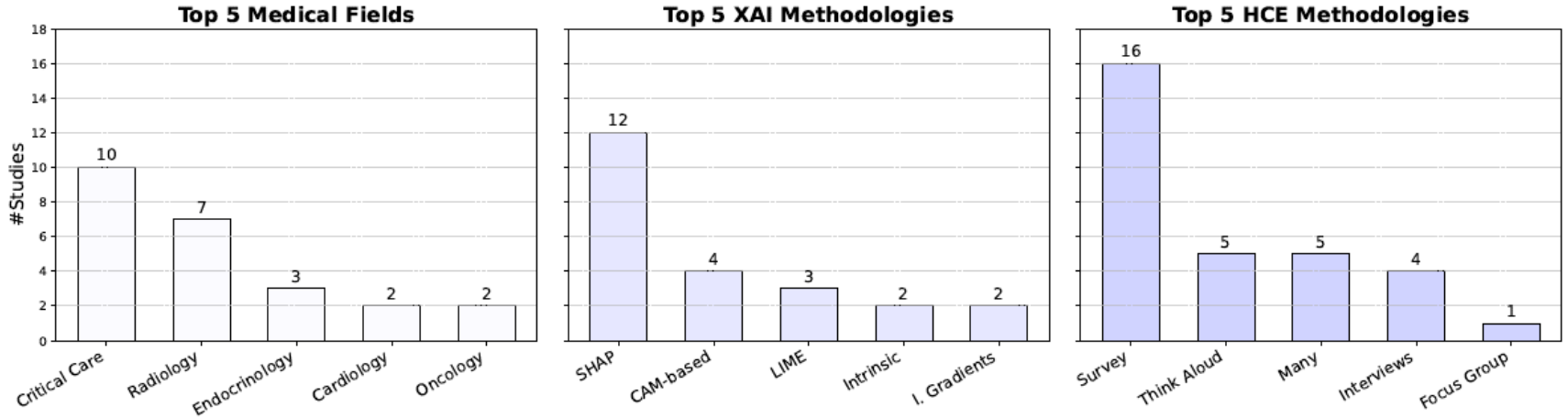


Figure 4: Distribution of top XAI and HCE methodologies and top medical fields where CDSS were used.

Discussion: Key Themes from the Survey

The evaluation results were not uniformly positive. The authors synthesized the findings into four main themes that capture the nuances of XAI adoption in clinical practice.

Table 2: Summary of discovered themes related to the works surveyed, pointing out clinical outcomes and clinical takeaways.

Themes and Sub-themes	Papers	Clinical Outcomes
1. Improvements in Risk Assessments	[Brennan *et al.*, 2019] [He *et al.*, 2024] [Jung *et al.*, 2025] [Hur *et al.*, 2025] [Kumar. *et al.*, 2020] [Gu *et al.*, 2020] [Jing *et al.*, 2025] [Rajashekar *et al.*, 2024] [Famiglini *et al.*, 2024] [Rainey *et al.*, 2024] [Sabol *et al.*, 2020] [Kayadibi *et al.*, 2025] [Singh *et al.*, 2021] [Chanda *et al.*, 2024] [Pumplun *et al.*, 2023] [Singla *et al.*, 2023]	Documented Adoption Enhanced Diagnostics
1.1 Enhanced Support for Novice Clinicians	[Neves *et al.*, 2021]	Better Novice Support
1.2 Reduction of False Learning	[Ellenrieder *et al.*, 2023]	Optimal Learning
1.3 Enhancement of Clinical Planning	[Abraham *et al.*, 2023]	Enhanced Planning
2. Preference for Established Clinical Practice	[Kovalchuk *et al.*, 2022] [Matthiesen *et al.*, 2021] [Zhang *et al.*, 2024] [Sivaraman *et al.*, 2023] [Hwang *et al.*, 2022] [Bhattacharya *et al.*, 2023]	Seek for Clinical Explanations Plausibility
2.1 XAI not Always Actionable	[E. Ihongbe *et al.*, 2024] [Chari *et al.*, 2023]	Prefer Actionable Explanations
3. Disparities Across Clinical Roles / Seniority Levels	[Gombolay *et al.*, 2024] [Barda *et al.*, 2020] [Cabitza *et al.*, 2025]	Implement Adaptive Explanations
4. Frictions Between Stakeholders for Development	[Bienefeld *et al.*, 2023]	Integrate Iterative Feedback Loops

Key Findings: Intrinsic vs. Post Hoc Models

Intrinsic Models (Sec 3.1)

Only two surveyed studies used inherently interpretable models (e.g., decision trees, linear models).

- MyRiskSurgery: Showed positive perception and improved physician risk assessment.
- Neurology Study: Found that one-size-fits-all explanations don't work; performance degraded for experienced neurologists, suggesting a need for personalization.

Post Hoc Models (Sec 3.2)

The vast majority of studies used post hoc methods, applying an explanation layer on top of a black-box model. These were predominantly model-agnostic and focused on local (patient-level) explanations.

- Dominant Method: SHAP was the most widely adopted technique.
- Common Goal: Identifying critical clinical features for disease prediction and outcome assessment.

Deep Dive: Post Hoc Methods in Practice

A mixed but mostly positive reception among clinicians.

Critical Care

SHAP explanations improved clinician alignment with the CDSS and increased trust, but also came with a learning curve and could increase cognitive effort.

Sepsis Treatment

A study identified four clinician behaviors: Ignore, Negotiate, Consider, and Rely. This shows explanations are often treated as supplementary evidence, not a definitive instruction.

Endocrinology & Cardiology

Clinicians often preferred simpler, actionable, data-centric visualizations (bar charts) over more complex SHAP plots. Contextual information was more valued than raw feature importance.

Overall Sentiment

Explanations were valued for clarity and augmenting confidence. However, limitations persisted, including the need for user adaptation and a preference for less cognitively demanding solutions.

Conclusion & Key Takeaways

While XAI generally enhances trust, its effectiveness is highly variable. A 'one-size-fits-all' approach to explainability fails in the complex world of clinical medicine. The path forward requires a new development paradigm.

- CDSS as an Augmentative Tool: The goal is to enhance, not replace, clinician decision-making.
- The Socio-Technical Gap is Real: Simply adding an explanation is not enough; the entire system must be designed with human context in mind.
- Iterative, Stakeholder-Centric Design is Essential: Involving all stakeholders (especially clinicians and patients) from the beginning is crucial for building trust and ensuring clinical viability.
- Rigorous Evaluation is Non-Negotiable: Given the high stakes, application-grounded evaluations with real domain experts must be the standard.

Thank You!

Email: ***a.mamun@asu.edu***

Paper

abdullah-mamun.com

𝕏: @AB9Mamun