Explainable AI for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences

Jun Hou, MS[1], Lucy Lu Wang, PhD[2]
[1]Virginia Tech, Blacksburg, VA; [2]University of Washington, Seattle, WA

Abstract

Explainable AI (XAI) techniques are necessary to help clinicians make sense of AI predictions and integrate predictions into their decision-making workflow. In this work, we conduct a survey study to understand clinician preference among different XAI techniques when they are used to interpret model predictions over text-based EHR data. We implement four XAI techniques (LIME, Attention-based span highlights, exemplar patient retrieval, and free-text rationales generated by LLMs) on an outcome prediction model that uses ICU admission notes to predict a patient's likelihood of experiencing in-hospital mortality. Using these XAI implementations, we design and conduct a survey study of 32 practicing clinicians, collecting their feedback and preferences on the four techniques. We synthesize our findings into a set of recommendations describing when each of the XAI techniques may be more appropriate, their potential limitations, as well as recommendations for improvement.

INTRODUCTION

Clinical decision support systems (CDSS) powered by machine learning and AI have the potential to assist in medical decisions and improve patient outcomes. However, to meaningfully support clinicians, AI-powered CDSS must be trustworthy and interpretable, allowing clinicians to assess the utility and applicability of model predictions. Explainable AI (XAI) techniques have been proposed to improve model interpretability, especially for neural network and other blackbox models.[1] While XAI techniques have been applied to CDSS,[2] a comprehensive understanding of clinician preferences and perceptions regarding XAI applications in these systems remains largely unexplored.

Prior work on clinical XAI tends to focus on explanatory accuracy, in terms of which models are applicable,[3] how to integrate XAI methods for different healthcare tasks,[4] or which datasets are available to train on.[5] While these works consider XAI for improving model interpretability, they do not incorporate user studies with clinical practitioners to understand whether XAI methods meet their needs. Solely focusing on technical performance metrics can lead to a gap between how AI developers and healthcare professionals view and assess AI tools.

We fill this validation gap by surveying clinical workers about the utility of popular XAI methods applied to text-based CDSS. We conduct a study with 32 clinicians (predominantly nurses along with physicians, technicians, and administrators), asking them to interact with and compare several XAI methods, and eliciting feedback about the design and utility of these methods. We compare four methods in our study, including LIME (Local Interpretable Model-Agnostic Explanation),[6] Attention-based span highlights,[7] exemplar patient retrieval, and free-text rationales produced by a large language model (LLM), which represent all explanation forms categorized in the review by Chaddad et al.[4] We apply these methods to explain outputs for a model trained to predict in-hospital mortality in an intensive care setting, a task studied extensively in prior work.[8,9,10] We select this task because it mirrors decision-making processes common in disease diagnosis, supporting the potential to generalize our findings to other similar scenarios. Based on participants' questionnaire results, we answer the following research questions: (1) how these XAI methods can be improved, (2) what tasks they potentially support, and (3) how they compare.

After conducting thematic analysis on the results, we synthesize a set of recommendations for designing and implementing XAI methods in an ICU setting. Our findings underscore the importance of creating both efficient, generalized tools and specialist-sensitive options tailored to varying levels of clinical expertise. We also observe a strong preference among clinicians for free-text rationales, highlighting their potential to enhance communication between healthcare providers and patients. However, our participants also emphasize the importance of evidence-based XAI approaches, such as similar patient retrieval, in building trust between clinician users and AI systems.

BACKGROUND & RELATED WORK

Explainable AI (XAI) in healthcare Feature-map methods such as LIME[6] and SHAP (SHapley Additive exPlanations)[11] have been explored repeatedly in clinical settings,[12,13] as have textual explanation forms.[14,15] Shen et al.[16] applied an example-based XAI method, which retrieves similar clinical cases through Case-Based Reasoning. In clinical settings, the focus is on post-hoc local explanations that balance accuracy and clarity and provide detailed insights for individual patient cases.[17] Therefore, we focus on post-hoc instance-level explanations in this work, comparing four XAI methods that span the explanation forms discussed in Chaddad et al.[4]

In-hospital Mortality Prediction In-hospital mortality prediction aims to estimate the risk of a patient dying during their hospital stay, and is crucial for prioritizing treatment strategies and resource allocation. Prior studies have investigated this task in the ICU setting using clinical text and time series Electronic Health Record (EHR) data.[10,18] Performance on the task was significantly enhanced by leveraging LLMs in Van Aken et al.[9] Naik et al.[8] then integrated patient-specific retrieved literature as input into predictive models to enhance performance. These works highlight the continuing focus on this critical task using diverse methodologies, driving our choice of this task.

Understanding Physician Perspectives and Preferences A literature review conducted by Antoniadi et al.[19] revealed the importance of XAI in building trustworthy AI/machine learning-based CDSS, as well as the lack of user studies in their development. A systematic review conducted by Jung et al.[20] also found that prior work on XAI in healthcare lacked a consensus evaluation framework for assessing the success of the XAI method. We hope to approach these recommendations from a user-centered perspective, based on what clinicians identify as useful aspects of XAI applied in Natural Language Processing (NLP)-powered CDSS.

MATERIALS & METHODS

Data & Task The data for our study is derived from MIMIC-III,[21] a collection of de-identified health records from 46,520 patients who stayed in the critical care units of Beth Israel Deaconess Medical Center between 2001-2012. We adopt the early-detection mortality prediction task from Van Aken et al.,[9] which uses patient admission notes to predict whether a patient will experience in-hospital mortality. Each patient's admission note is semi-structured free text, consisting of the sections Chief complaint, Present illness, Medical history, Admission Medications, Allergies, Physical exam, Family history, and Social history. We use train/test splits from their work,[9] which consist of 30,420 patients in the survived class and 3,534 patients in the mortality class in the train split; and 8,797 patients in the survived class and 1,025 patients in the mortality class in the test split.

Predictive Model We train predictive models on mortality prediction and select examples for explanation generation and inclusion in our survey. We use UmlsBERT[22] as our base model because it was found to be the most effective for in-hospital mortality prediction in prior work.[8] UmlsBERT is a semantically-enriched model pretrained on MIMIC-III and the UMLS Metathesaurus; a single linear layer is then added for adaptation to downstream classification tasks. In our case, we finetune the model on mortality outcomes from MIMIC-III following the non-literature-augmented model variant introduced by Naik et al,[8] achieving 87.86 micro-F1 and 66.43 macro-F1 on a held-out test set.

Implementation of XAI Methods We apply post-hoc XAI methods to our model predictions to create explanations. We experiment with feature map, textual, and example-based explanation forms,[4] specifically LIME,[6] Attention-based explanations,[7] exemplar explanations through similar patient retrieval, and free-text rationales from an LLM:

- **LIME**[6] is a model-agnostic XAI method that provides feature-based explanations. LIME perturbs the input data by altering or removing features and observes corresponding changes in the model's prediction. Following LIME, we train an explainer module that treats each word in the input document as a feature, and identifies relevant features in a given patient's admission note that contribute to the prediction result for each of the two outcome classes: mortality (positive) and survived (negative). We present these explanations by highlighting words identified as relevant features for each class, with the intensity of the highlight reflecting the magnitude of the feature's importance. We introduce a percentile variable to limit the number of highlighted words to only the top percent of features based on their importance scores.
- **Attention-based** explanations use weights from attention-based models[7] to explain model decisions, making it a model-specific method. We apply the approach outlined by Falaki et al.[23] to extract attention weights from our UmlsBERT model's [CLS] tokens and recombine subword tokens into words for visualization. We highlight the top 30% of words by attention weights, with the intensity of the highlight defined by the weight value.

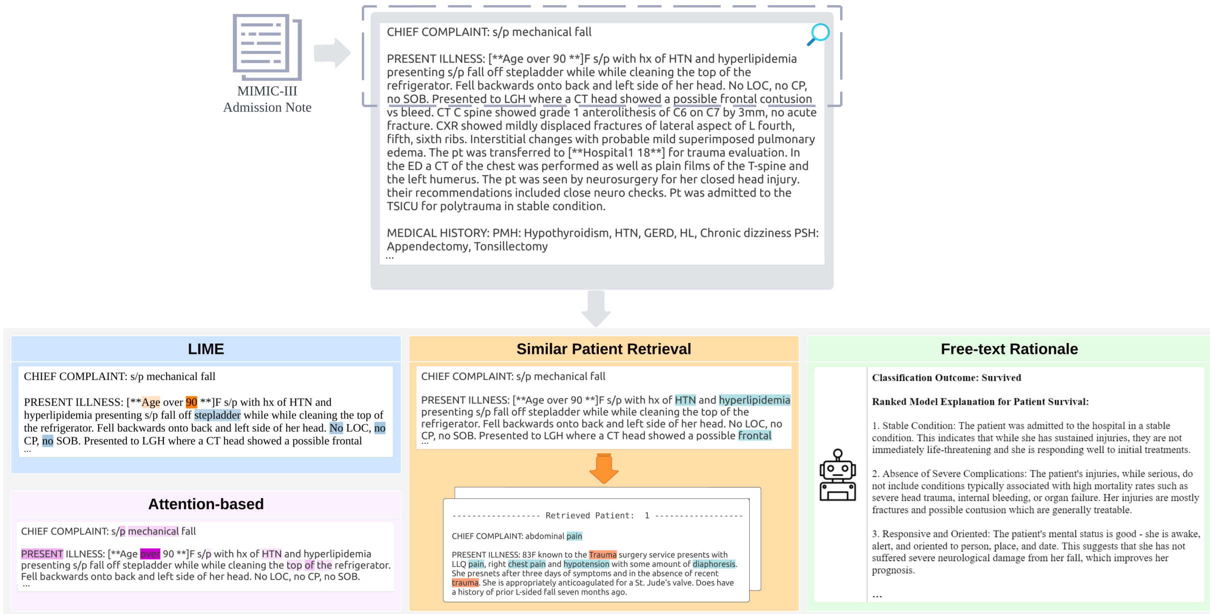


Figure 1: Example outputs of the four XAI methods applied to an MIMIC-III admission not, with color code and intensity indicating feature importance as detailed in the respective sections.

- **Similar Patient Retrieval** is an exemplar-based, model-agnostic XAI method aimed at producing explanations by retrieving the closest examples from a training dataset. We fine-tune UmlsBERT on mortality prediction and semantic textual similarity through contrastive learning following the SentenceBERT framework.[24] We embed all patients with this finetuned model, then apply k-nearest neighbor retrieval using cosine similarity to identify similar patients. At inference time, we retrieve the top-3 most similar patients from the training split with the same outcome as what is predicted to show as exemplars. To facilitate visual comparison of similarities and differences between retrieved patients and the test patient, we apply Named Entity Recognition (NER) using scispaCy[25] and highlight matching entities between test and retrieved notes in orange and non-matching entities in blue.
- **Free-text Rationales** are a model-agnostic XAI method that attempt to generate human-comprehensible explanations in natural language. Recently, this has often been achieved by prompting LLMs such as GPT-4,[26] as we do in this work. Our prompts can be found in our Github. We sample six admission notes and their ground-truth labels from the train split based on the method proposed in Liu et al.[27] for use as in-context examples. We then present the test note and ask for the top 3 reasons for and against the predicted outcome label.

Several of these methods are *model-agnostic*, i.e., explanations are generated by a separate model than the one making the prediction. These methods leverage surrogate models and perturbation techniques to approximate prediction model behavior; because the mechanism of the explanatory model differs from the prediction model for these methods, researchers have questioned the fidelity of their explanations.[1] For this reason, we also include the *model-specific* method (Attention-based explanations) in our investigation.

Figure 1 shows examples of the four XAI methods applied to an example patient admission note. In each case, we give the admission note, model prediction, and in some cases the model itself, as input into the XAI method. We then conduct additional postprocessing of the results to facilitate visualization and comparison. Details and code implementations for all methods can be found on Github: https://github.com/JuneHou/XAI_MOR_Survey.git

Survey Design To explore clinician perceptions of XAI methods, we design a survey eliciting feedback for each of the implemented XAI methods and comparing across them. We structure our survey into key sections as shown in Figure 2. Following survey completion, we follow up with a subset of participants to better understand their preferences and the

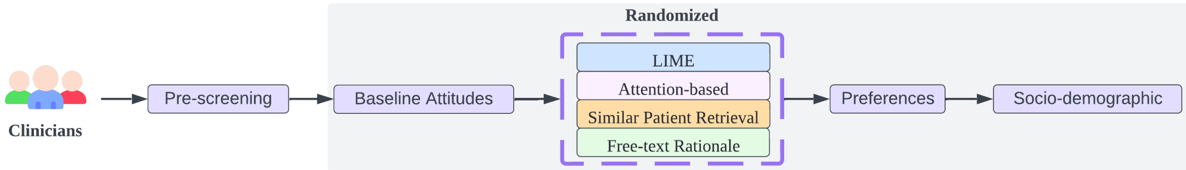


Figure 2: Survey design and workflow

rationales behind their answers. We describe the design of each survey section below:

- **Baseline Attitudes**: Personal experiences with AI systems have been found to impact user perceptions of AI.[28,29] Therefore, we ask participants to choose a set of five values from Jakesch et al.[30] that they consider most important for clinical AI systems and to indicate their attitudes towards AI on a 5-point Likert scale.
- **XAI Perceptions & Preferences**: We show participants four samples, one of each of the XAI methods with order randomized. Each XAI technique used in our study is accompanied by a detailed explanation of how it works at the beginning of the respective section. To offset patient- and outcome-specific biases, we sample two patients for each of the mortality and survived outcomes to include. We review XAI outputs to ensure that examples shown to users in the survey do not contain obviously incorrect or irrelevant information.

  After each example, the participant answers Likert-scale questions on the understandability, reasonableness, and usefulness of that XAI method; these facets are inspired by the evaluation of explainability and interpretability described by Saeed et al.[31] A free-text field is provided to allow the participant to express pros and cons in their own words. For each XAI method, we design additional questions specific to the characteristics of that method, e.g., whether the percent of words highlighted or number of similar patients retrieved are too many or too few.

  Following the evaluation of all four XAI samples, the participant is asked to rank the understandability and reasonableness of all four methods, and indicate their preference among them. We further ask the participant to assess the time efficiency of each method, and whether each method achieved the goals of enhancing confidence, broadening perspective, and increasing trust, three criteria defined in prior work as goals of XAI.[19]
- **Socio-demographic Information** We collect self-reported gender, age, and race/ethnicity from participants, along with their highest level of education and number of years of clinical experience. Participants also reported their job title, which we categorize into different job positions for reporting.

Study Recruitment Participants are eligible if they are clinical practitioners with more than two years of clinical experience OR medical school students with at least two years of training. We required that all participants be over 18 years of age, have at least a bachelors degree, and are located in the US. Participants were recruited using the Upwork platform, and compensated at rates of \$15-\$25 per hour for completing pre-screening, the main survey, and any followups. Our study was found exempt by the IRB at the University of Washington (STUDY00019118).

Results

We recruited 32 clinical practitioners to participate in our study. A majority of participants identified as white (75%), followed by hispanic/latino/a/x (12.5%), South Asian (6.3%), African-American/Black (3.1%), and Southeast Asian (3.1%). 78.1% of participants identified as female. Participant distribution across age groups is more balanced.

The highest level of education obtained were community college (3.1%), undergraduate degrees (53.1%), Masters degrees (25.0%), medical degrees (15.6%), and doctorate degrees (3.1%). The experience levels of participants varied, though a large majority had over 5 years of experience in clinical medicine (12.5% with 2-5 years, 37.5% with 5-10 years, 40.6% with 10-20 years, and 9.4% with >20 years). Most survey participants are registered nurses or nurse practitioners (78.1%), with others who identified as doctors/physicians (9.4%), researchers (6.3%), or other (6.3%).

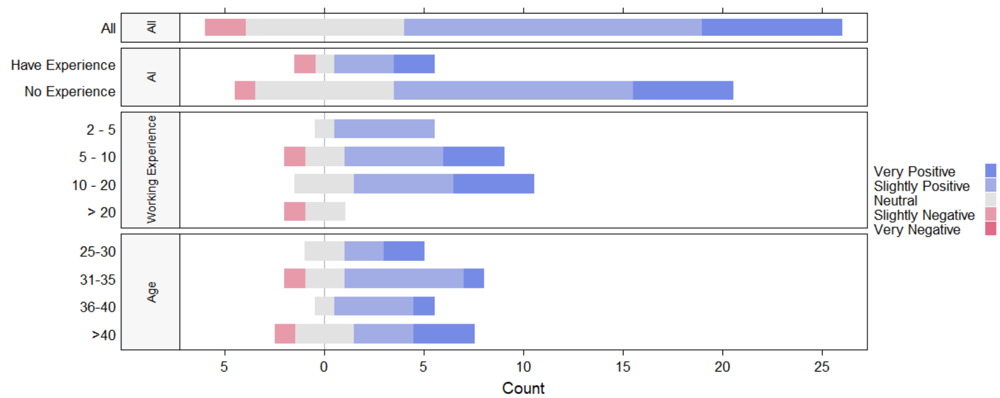
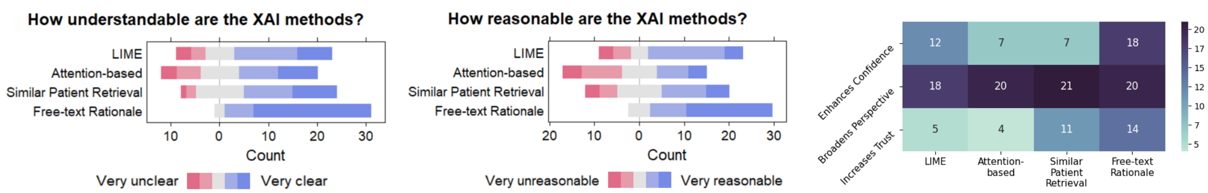


Figure 3: Participant attitudes toward AI by demographic group.



(a) Understandable and reasonable quantified by Likert scale question

(b) Heatmap for Functionality

Figure 4: Overview of practitioners' evaluations on the effectiveness and utility of XAI methods, covering understandability, reasonability, and key functional goals.

Summary Findings

What Clinicians Value in AI Clinicians were most likely to consider *safety* and *performance* important. Clinicians without AI experience selected *privacy* more often, while those with AI experience cared more about *beneficence*. *Accountability*, *human autonomy* and *transparency* were also rated as relatively important among all participants.

Attitudes Towards AI Figure 3 reports participant self-reported attitudes towards AI, collectively and split into different demographic groups. Overall distributions in attitudes toward AI are similar across groups. Very experienced clinicians (>20 years experience) in our cohort do not exhibit any positive sentiments towards AI; however, we note the small sample size (n=3).

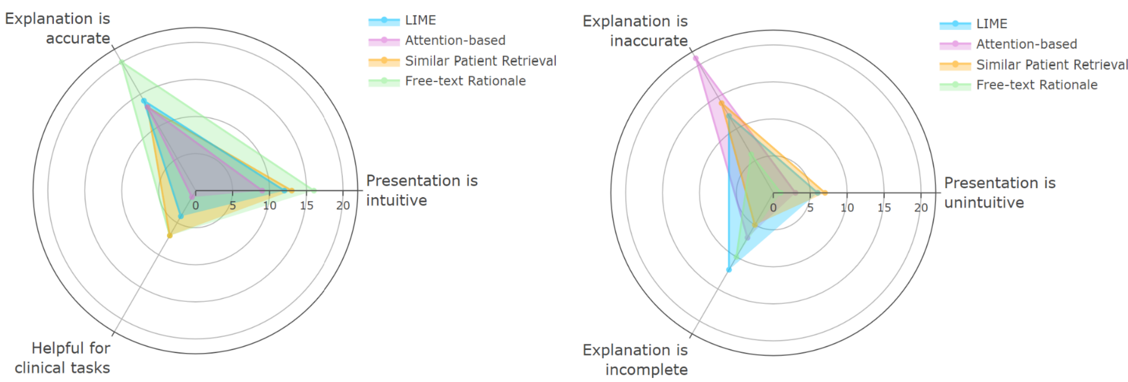
Attitudes Toward XAI Methods Regarding how understandable and reasonable each XAI method is from the clinician's perspective (Figure 4(a)), we observe similar results for the four XAI methods. Free-text rationales were found to be the most understandable and reasonable, with no negative responses. Attention-based explanations were found to be the least reasonable and hardest to understand. LIME and similar patient retrieval received a similar amount of positive and negative feedback along both dimensions, though similar patient retrieval was rated as more understandable. Clinician preferences toward the quantity of information provided, such as the percent of words highlighted or the number of similar patients and rationales shown, generally lean towards preferring more information.

Practitioner Preference Overall, free-text rationales were the most preferred method (n=15), with LIME second (n=12). When considering all four methods, no participant ranked Attention-based explanations first. The understandability of the similar patient retrieval method showed an almost even distribution across ranks 1-4. However, this method was the least preferred for use in real clinical settings, with an average ranking of 4.

We ask clinicians to assess the time efficiency of their most preferred XAI method on a Likert scale with 1 as least efficient and 5 as most efficient For free-text rationales, which is ranked first 15 times, 13% of participants reported

Sentiment	Theme	Description
Positive	Presentation is intuitive	The output of the XAI techniques (visualizations or text) is presented in a way that is intuitive and easy to understand
	Explanation is accurate	Highlighted keywords and/or free-form explanations align with clinicians' expectations and thought processes
	Helpful for clinical tasks	Explanations are useful in clinical settings, and could assist with tasks like decision-making or patient communication
Negative	Presentation is unintuitive	The output of the XAI techniques (visualizations or text) require further clarification or explanation to be understood by clinicians
	Explanation is inaccurate	Explanations are irrelevant or do not align with clinicians' expectations
	Explanation is incomplete	Additional information is required to fully support decision-making processes

Table 1: Themes grouped by sentiment with descriptions.



(a) Positive sentiment (b) Negative sentiment

Figure 5: Radar plots of themes mentioned by participants for each XAI method, grouped by sentiment.

Minimal time saved, 20% *Moderate time saved*, 47% *Considerable time saved*, and 20% *Significant time saved*. For LIME, ranked first 12 times, the time savings in the same four categories were assessed as 0%, 25%, 67% and 8%, respectively. In other words, participants were more divided on the time saving ability of free-text rationales.

Figure 4(b) visualizes responses towards whether each XAI method succeeds in enhancing confidence, broadening perspective, and increasing trust. The objective of broadening the users' perspective was successfully met by all methods, with more than half of participants agreeing for each method. Enhancing confidence was most effectively addressed by free-text rationales, with 18 participants agreeing, followed by LIME, with 12 participants agreeing. Increasing trust was the most challenging goal, though many more participants indicated similar patient retrieval and free-text rationales as helping to meet this goal. Additional statistics can be found on our Github.

Thematic Analysis

We conduct qualitative coding to identify shared themes among participant attitudes towards XAI methods. The first author conduct open coding on survey responses to identify major themes, with feedback and iteration from other authors. This process led to the identification of 6 themes, which we present and describe in Table 1, organized by positive and negative sentiment. Radar plots in Figure 5 compare the number of participants who raised each theme for each XAI method. Below, we summarize the qualitative feedback for each XAI method included in our study. Participants are identified by pseudonyms P1-32.

LIME The visualization technique of LIME is favored by 12 practitioners, with reasons including *"shade intensity was very helpful in showing how important the phrase was"* (P7) and *"can be helpful in drawing quick conclusions"*

(P4). The use of orange and blue for negative and positive features was described as intuitive by 6 practitioners. Highlighted words were found to match the clinician's thought process (14 participants); P15 mentioned *"highlighting words such as 'intubated' and 'unresponsive' is good."*, and P17 confirmed, *"I found favorable that the system considered more influential (darker shade) the age of the patient."* P30 offered one potential use of LIME in clinical settings by expressing that *"The red highlighted words did correlate with a mortality risk and were helpful in identifying these risks in the text."* Despite generally positive feedback, some highlighted words were considered not correlated with the outcome (P8) or were located in regions of the admission notes considered irrelevant for prediction (P21 and P28). Furthermore, several participants (P8 and P15) expressed a desire for more unique span highlights rather than duplicates of already highlighted words.

Attention-based Highlights The attention-based method received the most feedback for the need to improve the quality of highlights, with 21 participants mentioning the low relevancy of highlights to the outcome. Several participants were especially bothered by highlighted words like 'of' or 'with,' described by P12 as "random" and P22 as "distracting from the outcome." P2 discussed potential benefits, such as the model *"highlighting important words such as 'cardiac' and 'received CPR' because my brain thinks those are important too,"* suggesting areas where the model performs well in identifying relevant information. Regarding visualization, the single-color highlights were favored by 9 participants for their simplicity. P13, as the only participant commenting on utility for clinical tasks, mentioned that the highlights *"contribute to the respective viewer taking a Quick Look back to make sure no information is missed,"* highlighting the potential of visualization techniques to aid in clinical decision-making.

Similar Patient Retrieval Similar patient retrieval was found to be clinically helpful by 7 clinicians: for administrative tasks (P3), planning of treatment (P4, P19, P25), supporting diagnosis (P9, P12, P19), and *"guiding and auditing manifestation and intervention"* (P11). P2 commented that the highlighted NER terms aligned with their thought process, while P17 discussed accurate retrieval as a valuable feature of this technique, especially in terms of *"similar age (elderly),... abnormal electrocardiogram signs"* and other shared treatments and symptoms. However, additional clarification in terminology is suggested by P28, who noted *"The use of short hand/abbreviations should be minimized. As this could lead to confusion."* P27 discusses how this could be used to improve retrieval relevance: *"Consistent use of the same language... (eg s/p cardiac arrest vs s/p PEA arrest) would help to pull similar patients."*

Regarding visualizing NER overlap, 9 participants liked this feature, which allows for *"Quickly referencing similarities in past medical history and treatment"* (P9). P1 says *"color related to higher match is helpful."* P14 suggests the potential for color to convey more detailed categories, e.g., *"Meds: orange, Procedures: yellow, vitals: purple,"* to assist with comparison. However, there is room for improvement in the quality of explanations (14 participants), such as replacing non-contributory medication highlights with more influential words (P16), and wanting higher similarity between queries and retrievals, especially in areas like chief complaints.

Free-text Rationales Free-text rationales received the most positive feedback, with 16 participants indicating that its outputs were intuitive and 15 indicating that the explanations were accurate. Rationales were found to enhance predictability through conciseness of presentation (P16, P17, P19) and accurate reasoning (P23, P24). Participants indicate rationales can aid in clinical tasks such as prognosis (P11, P17, P30), prioritization (P13), decision-making (P15, P19), and treatment planning (P27), offering clarity for understanding (P11, P21) and facilitating communication with non-experts like patients (P12, P30). For improvement, participants suggested considering more contextual details like *"high doses of pressors in the mor[t]ality rate"* (P1) and the strength of the rationale, e.g., *"the medication allergy reasoning was weak"* (P8). Incorporating quantifiable scores (P7) and *"evidence-based protocols in the rationales, e.g. AHA"* (P11) could provide further support for outcome predictions, with P21 suggesting that *"Adding confidence levels or percentages would significantly improve how trustworthy this algorithm is."* For presentation enhancements, P31 suggests *"Bold faced and highlighted words for important info"*. On the other hand, use of jargon and excessive reasoning were found to reduce explanation clarity (P4).

DISCUSSION

Our analysis of survey results raises questions about the goals and benefits of XAI methods, and how to increase their relevance and utility to clinical practitioners. We synthesize these findings into additional recommendations below.

The importance of providing evidence Similar patient retrieval, while critiqued for its accuracy and effectiveness, demonstrated a greater ability to build trust than feature-based methods such as LIME and Attention-based highlights. Using historical cases as evidence to support decisions closely mirrors how practitioners rely on past experiences in clinical decision-making. To be successful from the practitioners' perspective, our findings suggest the importance of not only generating accurate and informative rationales but also incorporating evidence-based support (as with the exemplars in similar patient retrieval) in model explanations. While free-text rationales were received positively by participants, the lack of grounded evidence needs to be considered. Combining free-text rationales with retrieved exemplars or externally retrieved evidence (as in Naik et al.[8]) could help address these issues in future work.

Potential of free-text rationales to bridge communication Explanations in natural language that reflect the cognitive processes of clinicians can serve as a communication bridge within the healthcare system. P27 mentioned that in time-sensitive scenarios, generated content can function similarly to a nurse's note for communicating with a doctor. Additionally, generated rationales can be an educational tool, e.g., P27 described *"Especially in the trauma setting, the workflow is very fast, and you got residents attending...even if it's not a teaching hospital, it is extremely helpful."* The method also has the potential to bridge the gap between healthcare providers and patients by explaining symptoms and treatments in a non-technical way, as mentioned by P17, *"this model...it would be a good thing to not maybe show the family members, but to explain, okay, we use this model and this is what the outcomes are saying."* However, based on participant suggestions to simplify wording in free-text responses, future work should consider integrating plain language summarization[32,33] to enhance the understandability and efficiency of LLM output.

XAI for both structured and unstructured data In the critical care setting, especially in ICUs, treatment plans and bedside monitoring rely heavily on both structured data (such as vital signs and lab results) and unstructured data (such as clinical notes). Several participants mentioned the need to incorporate multimodal data into an XAI-enhanced CDSS. While this is beyond the scope of our study, we emphasize that real-world CDSS would likely take advantage of input predictors beyond the clinical note text, and that the importance of these predictors would also require explanation. This could be an addition to ongoing research that aims to create predictive frameworks combining unstructured textual data with structured data for clinical prediction.[34]

XAI tailored to different clinical workflows Responses from clinical practitioners revealed varied perceptions and preferences for XAI methods at different stages of patient care. In urgent care settings like the ICU or surgery, clinicians prioritize efficiency and clarity of explanations, favoring XAI methods that present key information quickly for reference. However, in less acute phases like post-surgical care, detailed explanations are in demand for analysis and treatment planning, a role well-served by a method like free-text rationales. For example, P17 states, *"Highlight saves time and we need that. If we had more time in clinical settings, I feel like the free text rationale gives a more in-depth reasoning."* To balance details and efficiency,[20] XAI methods must be tailored to specific clinical workflows.

Limitations Although we have covered all categories of post-hoc XAI methods from Chaddad et al.,[4] there are many methods that we could not include in this survey due to time and resource constraints (our survey was already long). Applying XAI methods to text data was not always straightforward, as some methods (like LIME) work better on feature-based models with lower-dimensional input. For similar patient retrieval in particular, we face many challenges in fine-tuning the retrieval model due to the lack of labeled datasets for patient semantic similarity search.

We use only the free-text admission notes from MIMIC-III as the inputs to our prediction models, whereas prediction models and XAI methods can be applied to other data formats, such as tabular and time series data. Furthermore, the predictive task in this study is limited to in-hospital mortality prediction. Future work should explore multimodal outcome prediction models as well as other clinical predictive tasks.

Although we have made significant efforts to recruit, our cohort is still relatively small, involving 32 clinical practitioners who are predominantly nurses. This may limit the generalizability of our findings. However, we did achieve a fairly diverse cohort in terms of age and experience, and many themes were consistently raised by most participants. In the future, we aim to validate our findings in real-world deployments, which we believe will offer valuable perspectives. We also plan to explore whether clinicians can effectively use XAI to identify hallucinations in LLM-powered decision support and mitigate the risks introduced by such systems.

Conclusion

Our survey results reveal the demands and preferences of healthcare practitioners towards the implementation of XAI in CDSS. By integrating clinicians in the evaluation process, we observed a strong preference for XAI techniques that replicate clinical reasoning, such as exemplar-based patient retrieval and free-text rationales. These methods enhance the interpretability and trustworthiness of AI-supported decision making, which can further help in realize the full potential of AI in clinical decision making, ensuring that CDSS are not only effective but also align with healthcare providers' needs. Moving forward, we aim to refine these methods by incorporating structured and unstructured data, tailoring XAI approaches to specific clinical workflows. This will improve the utility and efficacy of CDSS across diverse clinical settings, further supporting healthcare professionals in their decision-making processes.

References

[1] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 447–459, Suzhou, China, December 2020.

[2] Jinyue Feng, Chantal Shaib, and Frank Rudzicz. Explainable clinical decision support from text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1478–1489, 2020.

[3] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. Natural language processing of clinical notes on chronic diseases: Systematic review. In *JMIR Medical Informatics*, page e12239. JMIR Publications Inc., 2019.

[4] A. Chaddad, J. Peng, J. Xu, and A. Bouridane. Survey of explainable AI techniques in healthcare. In *Sensors (Basel, Switzerland)*, page 634. MDPI, 2023.

[5] Flavio Di Martino and Franca Delmastro. Explainable ai for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, 56(6):5261–5315, 2023.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[7] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[8] Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States, July 2022. Association for Computational Linguistics.

[9] Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, April 2021.

[10] Mengqi Jin, Mohammad Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, Tiberiu Doman, Arun Ravi, Matthieu Liger, and Taha Kass-hout. Improving hospital mortality prediction with medical named entities and multimodal learning, 2019.

[11] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[12] Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11960–11969, Toronto, Canada, July 2023.

[13] Chandni Saxena, Muskan Garg, and Gunjan Saxena. Explainable causal analysis of mental health on social media data. In *Neural Information Processing*, pages 172–183, Cham, 2023. Springer International Publishing.

[14] Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, Iker García-Ferrero, Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, German Rigau, and Anar Yeginbergenova. Hitz@antidote: Argumentation-driven explainable artificial intelligence for digital medicine. In *Annual Conference of the Spanish Society for Natural Language Processing*, 2023.

[15] Zhuo Wang, Rong Hua Li, Bowen Dong, Jie Wang, Xiuxing Li, Ning Liu, Chenhui Mao, Wei Zhang, Liling Dong, Jing Gao, and Jianyong Wang. Can LLMs like GPT-4 outperform traditional ai tools in dementia diagno-

sis? maybe, but not today. *ArXiv*, abs/2306.01499, 2023.

[16] Ying Shen, Joël Colloc, Armelle Jacquet-Andrieu, Ziyi Guo, and Yong Liu. Constructing ontology-based cancer treatment decision support system with case-based reasoning. In *Smart Computing and Communication: Second International Conference, SmartCom 2017, Shenzhen, China, December 10-12, 2017, Proceedings 2*, pages 278–288. Springer, 2018.

[17] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. In *Journal of Biomedical Informatics*, page 103655. Elsevier BV, 2021.

[18] B. J. Marafino, M. Park, J. M. Davies, R. Thombley, H. S. Luft, D. C. Sing, D. S. Kazi, C. DeJong, W. J. Boscardin, M. L. Dean, and R. A. Dudley. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. In *JAMA Network Open*, page e185097, 2018.

[19] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[20] Jinsun Jung, Hyung-Mok Lee, Hyunggu Jung, and Hyeoneui Kim. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 9, 2023.

[21] A. Johnson, T. Pollard, and L. et al. Shen. Mimic-iii, a freely accessible critical care database. In *Scientific Data*, page 160035, 2016.

[22] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online, June 2021.

[23] Ala Alam Falaki and Robin Gras. Attention visualizer package: Revealing word importance for deeper insight into encoder-only transformer models, 2023.

[24] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[25] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[26] OpenAI. GPT-4 technical report, 2024.

[27] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022)*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.

[28] Philipp Brauner, Alexander Hick, Ralf Philipsen, and Martina Ziefle. What does the public think about artificial intelligence?—a criticality map to understand bias in the public perception of AI. *Frontiers in Computer Science*, 5:1113903, 2023.

[29] Feridun Kaya, Fatih Aydin, Astrid Schepman, Paul Rodway, Okan Yetişensoy, and Meva Demir Kaya. The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. *International Journal of Human–Computer Interaction*, 40(2):497–514, 2024.

[30] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 310–323, 2022.

[31] Waddah Saeed and Christian Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

[32] Yue Guo, Weijian Qiu, Yizhong Wang, and Trevor A. Cohen. Automated lay language summarization of biomedical scientific reviews. *AAAI*, 2021.

[33] Yue Guo, Tal August, Gondy Leroy, Trevor A. Cohen, and Lucy Lu Wang. APPLS: Evaluating evaluation metrics for plain language summarization. *arXiv*, 2023.

[34] Sayna Ebrahimi, Sercan O. Arik, Yihe Dong, and Tomas Pfister. LANISTR: Multimodal learning from structured and unstructured data, 2023.