Handling missing values in clinical machine learning: Insights from an expert study

Lena Stempfle STEMPFLE@CHALMERS.SE
Chalmers University of Technology & Gothenburg University, Sweden

Arthur James ARTHUR.JAMES@APHP.FR
Sorbonne University, GRC 29, AP-HP, DMU DREAM,
Department of Anaesthesiology and Critical Care, Pitié-Salpêtrière Hospital, Paris, France

Julie Josse JULIE.JOSSE@INRIA.FR
PreMeDICaL Inria-Inserm, France

Tobias Gauss TGAUSS@PROTONMAIL.COM
Déchocage - Bloc des urgences, Pole Anesthésie-Réanimation, CHU Grenoble Alpes, France

Fredrik D. Johansson FREDRIK.JOHANSSON@CHALMERS.SE
Chalmers University of Technology & Gothenburg University, Sweden

Abstract

Inherently interpretable machine learning (IML) models offer valuable support for clinical decision-making but face challenges when features contain missing values. Traditional approaches, such as imputation or discarding incomplete records, are often impractical in scenarios where data is missing at test time. We surveyed 55 clinicians from 29 French trauma centers, collecting 20 complete responses to study their interaction with three IML models in a real-world clinical setting for predicting hemorrhagic shock with missing values. Our findings reveal that while clinicians recognize the value of interpretability and are familiar with common IML approaches, traditional imputation techniques often conflict with their intuition. Instead of imputing unobserved values, they rely on observed features combined with medical intuition and experience. As a result, methods that natively handle missing values are preferred. These findings underscore the need to integrate clinical reasoning into future IML models to enhance human-computer interaction.

1. Introduction

Missing values in healthcare data pose substantial challenges in predictive modeling, requiring effective methods to address them (Austin et al., 2021; Wells et al., 2013). These missing values occur in incomplete medical records and often arise intentionally, due to specific testing protocols, or unintentionally, due to time constraints or data collection errors (Salgado et al., 2016). The resulting gaps can significantly impact the reliability and applicability of predictive models, particularly at "test time" when they are used for decision-making. Interpretable machine learning (IML) models have become essential in high-stakes decision-making areas such as healthcare (Ustun and Rudin, 2019a), sustainability (Jin et al., 2022), and criminal justice (Wang et al., 2023; Zhang et al., 2022), where transparency and accountability are critical. These models offer insights into decision processes, ensuring that predictions can be understood and trusted by domain experts. IMLs, such as decision trees, linear models, and rule-based models, are extensively studied for their intuitive structure and ease of interpretation (Molnar, 2020; Rudin, 2019). They are widely used in tasks like classification, regression, and risk scoring (Fürnkranz et al., 2012; Margot and Luta, 2021). As with most ML models, IML models to date rely on complete inputs, making them particularly vulnerable to missing data. In retrospective studies, a typical solution is complete-case analysis, where cases with missing values are excluded (Little and Rubin, 2019). While effective for analysis, this approach is unsuitable at test time, as it can leave many patients without predictions. An alternative is the *impute-then-regress* strategy which imputes missing values be-

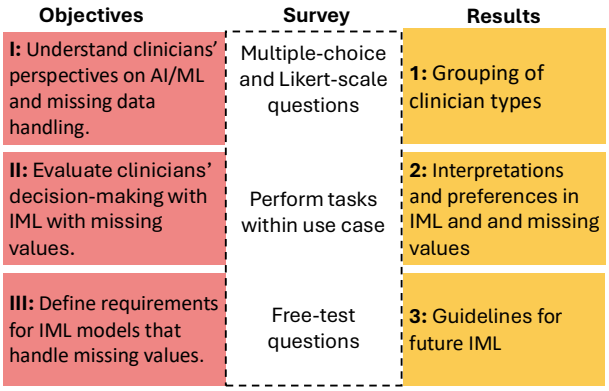


Figure 1: Overview of study objectives, key survey question types, and findings on best practices for clinical interpretable machine learning with missing values at test time.

fore prediction, ensuring consistency between training and deployment tasks (Rubin, 1976). However, impute-then-regress is often biased when the missingness of predictive features depends on unobserved variables. Moreover, complex imputation methods can reduce interpretability, complicating deployment and decision-making. Despite extensive research on handling missing values (Van Buuren and Groothuis-Oudshoorn, 2011; Le Morvan et al., 2020; Chen et al., 2023; Austin et al., 2021; Luo, 2022; Stempfle and Johansson, 2024), little is known about how medical domain experts interact with missing values when using prediction models.

In this work, we survey clinicians to gain qualitative insights into their handling of missing values using IMLs, providing a real-world perspective on decision-making. Our evaluation is structured around three key objectives (shown in Figure 1). First, we aim to understand clinicians' attitudes toward artificial intelligence/machine learning (AI/ML) in healthcare and their current practices for clinical decision-making with missing values (Objective I). Next, we present a use case aimed at identifying common interpretations and usages of three different IML approaches and preferences for strategies to handle missing values (Objective II). Finally, we define requirements for future IML models that handle missing values at test time, ensuring they align with clinician preferences and can be seamlessly integrated into clinical workflows (Objective III).

Our survey focuses on a real-world scenario in handling missing values for interpretable clinical prediction. This complements and contrasts with the rich literature on quantitative and theoretical comparisons between, e.g., different imputation methods. We emphasize domain expertise through direct quotes and publish our full questionnaire, enabling collaborative research on IML and missing values to identify best practices across healthcare systems.

Our main contributions are as follows: 1) We clustered the survey responses of 55 survey participants to analyze clinicians' attitudes toward AI/ML and their practices for handling missing values, providing an overview of existing workflows. 2) We assessed clinicians' use of three different IML approaches in scenarios with missing values, analyzing their reasoning, preferences, and challenges in leveraging these models for decision-making. 3) Based on clinician feedback, we formulated guidelines for designing IML models that handle missing values natively, ensuring better alignment with clinical workflows and facilitating real-world adoption.

Our findings highlight a gap between missing value handling in machine learning research, which often employs overly simplistic or too complex imputation methods, and clinical decision-making, where clinicians prioritize observed data and medical intuition at the time of decision-making. Future approaches should focus on reducing reliance on imputation by adopting more interpretable strategies, with models that natively handle missing values.

2. Survey methodology

Participant recruitment. We conducted an online survey using LimeSurvey [1] in July–August 2024 to assess clinicians' understanding of IML models and handling of missing values in decision-making. The survey involved 214 clinicians from 29 trauma centers across France, all part of the TraumaBase [2] network, which collects data on over 50,000 trauma cases annually. Responses were assumed to be largely independent due to the geographical spread of centers. To ensure clinical relevance, we focused on predicting hemorrhagic shock (Dutton et al., 2010), a critical ICU scenario where missing information is common. In trauma ICUs, data is often incomplete due to lab results being unavailable before critical decisions are needed. Trauma ICU clinicians, accustomed to clinical scores that align with IML frameworks, are

1. https://www.limesurvey.org
2. https://www.traumabase.eu/en_US

well-suited to evaluate IML models under test-time missingness. To boost response rates, we focused on a single use case and kept the survey within 20–25 minutes (details in Appendix A). Participants provided informed consent before participation.

Survey development. The survey was developed with the input of trauma physicians and machine learning researchers to ensure its relevance to both clinical and ML communities. To maintain linguistic accuracy and cultural relevance, it was provided in English and French by native speakers. The full survey in both languages is included in Appendix B.

Survey design. The survey consisted of 41 questions covering demographics, multiple-choice questions, ranking tasks, Likert-scale statements, and open-ended questions, aligned with the survey's three objectives (Figure 1). The *first part* of the survey focuses on understanding clinicians' perspectives on AI/ML models within healthcare and their current approaches to handling missing values during decision-making (Objective I). It includes multiple-choice and Likert-scale questions to capture detailed insights and measure attitudes consistently. Questions on current practices referenced trauma-related risk scores, such as the Red Flag (Hamada et al., 2018) or ABC score (Nunez et al., 2009), ensuring familiarity and serving as a warm-up by linking familiar tools to the challenges of missing values. The *second part* of the survey examines a specific use case (Objective II), where clinicians predicted hemorrhagic shock for a patient with missing values using various IML models. It assesses how clinicians engaged with model outputs for decision-making, focusing on interpretability and handling incomplete data. Multiple-choice and ranking tasks uncovered preferences and priorities when interpreting IML models (details in the next paragraph). Finally, the *third part* comprises open-ended questions, allowing clinicians to share insights, contributing to guidelines for future IML research (Objective III).

IML and missing values in a use case. The second part of the survey presented a patient case with a missing value and predictions by three IML models (Figure 2). Participants were asked how they would handle the missing value in their decision-making process. This allows us to investigate whether clinicians would rely on additional information, such as a subpopulation's average feature value, assign zero for the missing value, or apply their domain knowledge to estimate it. Additionally, we asked for their preferences in imputation strategies for different IML methods to better understand their approach to decision-making.

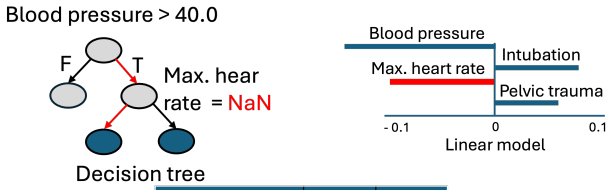
The survey asked questions about three well-established IML models: risk scores (Ustun and Rudin, 2019b), linear models (Hastie et al., 2009), and decision trees (Breiman, 2017). Since our primary interest was interpretability rather than predictive performance, we ensured all models had comparable accuracy (Molnar, 2020; Stiglic et al., 2020).

Handling missing values is challenging for these models, as they lack native support and often rely on imputation for predictions. Linear models require explicit handling of missing values since each feature contributes to the prediction based on its learned coefficient. If a feature is missing, its contribution is unknown unless imputed, meaning the model implicitly assumes the imputation strategy is correct. Rule-based risk scores require all features referenced in a rule to be available; otherwise, rule activation may be uncertain or impossible. Decision trees struggle with missing values due to their hierarchical structure, where each node depends on feature availability. If a feature is missing, the tree may follow a default branch (e.g., XGBoost Chen and Guestrin (2016)) or use imputed values, both potentially introducing bias. While XGBoost learns the default direction during training, this approach lacks interpretability at test time. Alternatively, decision trees can use **surrogate splits** (Feelders, 1999), identifying alternative features that mimic the primary split, enabling predictions without explicit imputation.

We use methods feasible for handling test-time missingness: imputation and native approaches:

- **Native methods**: These are model-inherent mechanisms for handling missing values, such as surrogate splits in decision trees, which rely on alternative features to make predictions (Feelders, 1999).
- **Implicit imputation**: Clinicians were presented with missing values without automated imputation and asked to estimate them using available features and their clinical judgment.
- **Explicit imputation**: These included simple statistical replacements, such as mean or zero imputation (Rubin, 1976).
- **Black-box methods**: More complex imputation techniques, such as Multiple Imputation by Chained Equations (MICE) (Van Buuren and

Patient ID	Blood pressure	Max. heart rate	Intubation	Pelvic trauma
36	42.0	NaN	Yes	Non



Feature	Points	
Blood pressure	1	1
Max. heart rate	2	+?
Intubation	1	+ 1
Pelvic trauma	1	+1
Sum Points	**Score**	**= ...**

	0	1	2	3	4	5
Risk score	<5%	12%	32%	66%	88%	< 99%

Risk score

Figure 2: Clinicians were given a patient sample (top) and three interpretable ML models (middle: decision tree, logistic regression; bottom: risk score) predicting hemorrhagic shock. We assessed their interaction with the models with missing values.

Groothuis-Oudshoorn, 2011), were used to assess how opaque imputation models influence clinician trust and decision-making.

Data analysis. We analyzed multiple-choice responses using descriptive statistics. For questions with responses on the Likert scale or categorical responses, we applied Factor Analysis for Mixed Data (FAMD) in R (Lê et al., 2008) to explore latent structures and identify response patterns. FAMD, the mixed data counterpart of Principal Component Analysis (PCA), accommodates both continuous and categorical variables, making it particularly well suited for questionnaire data (Josse and Husson, 2012). Since our study focuses on missing values, FAMD provides a natural framework where handling missingness is integrated into the analysis. To prevent bias from discarding incomplete responses, we employed the Expectation-Maximization (EM) algorithm to estimate principal components while maximizing the observed likelihood. This approach ensures that missing value estimation aligns with imputation, where minimizing the weighted least squares criterion or maximizing the observed likelihood produces equivalent results. To further analyze clinician groupings (Finding 3.1) based on AI/ML attitudes and incomplete data practices, we applied clustering to the FAMD results, allowing for a multivariate characterization of participants beyond isolated variables. Missing values in the survey answers, such as unanswered questions, were imputed using PCA for continuous variables and the most frequent category for categorical data. This process was implemented using the missMDA package (Josse and Husson, 2016), which relies on iterative Singular Value Decomposition (SVD) (Hastie et al., 2015), ensuring that imputed datasets preserve the same principal components as the original, maintaining the dataset's underlying structure. Free-text responses were analyzed for central tendency and variability. Figure 4 in the Appendix illustrates the full experimental setup.

Formal validation. We applied systematic methods to formally validate our qualitative survey and enhance the reliability, credibility, and accuracy of the data collection (Rosellini and Brown, 2021). First, we targeted domain experts to ensure the survey's relevance, clarity, and appropriateness. Before launching, we conducted a pilot study (Shakir and ur Rahman, 2022) with five independent trauma clinicians, incorporating their feedback to refine the clarity of the question and the response options. In Section 4, we applied triangulation by comparing our findings with other data sources, such as quantitative studies, and literature. Finally, to ensure consistency in qualitative analysis, multiple researchers independently reviewed the coded responses.

3. Best practices for IML with missing values at test time

This section presents findings aligned with the three main objectives (Figure 1): clustering participants by their attitudes toward AI/ML in healthcare, summarizing clinician interactions with current IML models for handling test-time missingness, and outlining requirements for future IML research.

3.1. Finding 1: Clustering based on AI/ML attitudes and missing values practices

We received 55 survey responses, including 20 fully completed and 33 partially completed, all of which were analyzed, resulting in varying answer counts per question. The average participant age was 41.2

years, with a gender distribution of 22 females and 33 males. The majority of respondents were clinical doctors (83%), primarily in Anesthesiology/Intensive Care (84%), and 90% held at least a medical degree. Participant demographics are provided in Table 1.

We identified four distinct clusters among survey participants based on the FAMD results (see Section 2). This provides a structured understanding of how clinicians' attitudes toward AI/ML and their handling of missing values vary across different response groups (Figure 5 and 6 in the Appendix). Cluster 1 includes clinicians (16% of 55, 9 participants) with low understanding and use of risk scores, little familiarity with decision trees, and dissatisfaction with all three models when dealing with missing values. Despite this, they are comfortable using black-box models like MICE but prefer implicit imputation and surrogate splits. Cluster 2 consists of clinicians (22%, 12 people) who show high levels of fear of AI, blind trust in its capabilities, and concerns about job loss, alongside perceiving current AI solutions as low-quality. They are familiar with decision trees and favor zero imputation or surrogate splits. Cluster 3 (36%, 20 people) has significant missing data and may be excluded. Cluster 4 comprises clinicians (25%, 14 people) who use AI daily, are highly familiar with all three IML models, and feel confident explaining these models with or without missing values. They favor implicit imputation over black-box models and surrogate splits. These clusters highlight the diverse perspectives among clinicians, emphasizing the challenge of developing approaches that accommodate varying familiarity, trust, and confidence in handling missing data.

Current practices with missing values. Across all clusters, we identified various strategies to manage missing values in currently used clinical scores, e.g. RedFlag (Dutton et al., 2010), primarily relying on clinical intuition or other available clinical information. Many mentioned using plausible values based on clinical context or the patient's history to estimate missing values. Some prefer worst-case imputation to avoid underestimating risk, though one noted it may lead to overtraining and resource inefficiency. Others choose to partially calculate or omit scores if key data is missing, feeling that imputation could 'skew the score concept'. One clinician also suggested using alternatives, like thromboelastography when prothrombin time is unavailable.

3.2. Finding 2: Clinician preferences and interpretations from a use case with IML approaches and missing values

Next, we analyze clinicians' responses from the use case evaluation, focusing on their approach to handling missing patient data in IML. We summarize their interpretation patterns and preferences for imputation or native methods.

Confidence does not always align with accuracy when handling missing values. Out of 20 responses, 9 clinicians rated themselves as highly confident on a Likert scale in explaining risk scores, even with missing values, while 10 expressed the same level of confidence for decision trees. For logistic regression, however, confidence was lower and accompanied by uncertainty; only 6 participants were correct, with 4 out of 6 correct among those expressing confidence in its usefulness with missing values.

Low trust in zero and mean imputation. When clinicians received the mean hemoglobin value for the dataset, almost all 23 respondents reported that this additional information neither influenced their decisions nor supported their decision-making in all three IMLs. The main reasons included the irrelevance to individual cases: "I don't take it into account. Each patient is unique, and I don't see the relevance of giving him an average value." When presented with a prediction for a patient in which the missing value was imputed as zero, most clinicians indicated that they would ignore the feature in the risk score. Similarly, a machine learning model would also ignore the feature, effectively diminishing its influence on the prediction. While the representation of risk scores was still considered helpful by most, some clinicians preferred more accurate imputations for the linear model, with one remarking: "If imputation is used during testing, it should be clearly indicated, ideally with the associated uncertainty. This transparency allows for comparison between predictions with and without the imputed value."

Surrogate splits can be viable. For decision trees, using surrogate splits (the direction to take when the splitting feature is missing) was explored, but this method was less familiar to clinicians and caused some confusion. However, many clinicians noted that its usefulness depends on clearly indicating where the substitution occurs: "If the substitution is clearly indicated and can be taken into account by the clinician, yes, it could be a decision aid." De-

Table 1: Descriptive statistics of survey participants. Categorical variables have been simplified for brevity: The 'Other' category for current occupation included roles such as nurses, clinical studies technicians, hospital clinical trials coordinators, clinical manager assistants, and retired professionals. For medical specialization, it included fields like rehabilitation, pediatrics, surgery, anesthesiology, bioinformatics, pediatric orthopedic surgery, and clinical research associate.

Characteristic	**Total (N=55)**	**Complete answers (N=20)**
Age in years, mean (SE)	41.18 (1.13)	39.35 (1.55)
Female, n (%)	22 (40%)	7 (5%)
Male, n (%)	33 (60%)	13 (65%)
Highest received educational degree, n (%)		
University bachelor/master/postgraduate education	9 (16,36%)	3 (15%)
Medical doctoral degree	37 (67,27%)	12 (60%)
Associate Professor/Professor	9 (16,36%)	5 (25%)
Current level of occupation, n (%)		
Clinical doctor	46 (83,64%)	19 (95%)
Intern/Trainee/Non-fully graduated clinician	3 (5,45%)	0
Other	6 (10,91%)	1 (5%)
Medical specialty, n (%)		
Anesthesiology / Intensive Care	46 (83.64%)	16 (80%)
Emergency medicine	5 (9.09%)	2 (10%)
Other	4 (7,27%)	2 (10%)

spite this, some preferred other methods, one of them stating: "Decision trees are easy to understand, but unstable and too sensitive to values. That's why we prefer Random Forest algorithms. It's difficult to rely on a decision tree, especially if there are missing values that have been imputed Rather than imputing [...], there are algorithms that estimate missing values in relation to other parameters."

High trust in risk scores and decision trees despite missing values. Lastly, the survey results show varying levels of trust among the 23 clinicians in using the three IML models for diagnosing hemorrhagic shock. Risk Scores were the most trusted, with 15 clinicians indicating they 'Almost Trust' the model. However, concerns about over-reliance on numbers rather than clinical judgment were noted, as one clinician remarked, "I'm not dealing with numbers, but with a patient! I need a history, background treatment, a clinical history [...]" Linear models faced the most skepticism, with 15 clinicians 'Not Sure' or 'Rather not' trusting them, due to unfamiliarity and doubts about handling missing values. Decision trees had mixed feedback—13 clinicians 'Fully or Almost Trust' them for their clarity, but some raised concerns about potential limitations.

Comparison of imputation methods across IMLs. Clinicians were asked to rank commonly used machine learning imputation methods, comparing explicit approaches (e.g., zero, mean imputation) to implicit imputation, which relies on clinical intuition to infer missing values. The results indicate a slight preference for implicit imputation within risk score models, though explicit methods were also widely used. Notably, black-box imputation, such as MICE, was rarely chosen (<10% of cases), despite its frequent use in the medical literature. This may stem from its lack of clinical interpretability and limited exposure to medical training. Figure 3 visually summarizes clinician selections across imputation methods and IML models. Surrogate splits and mean imputation were the most favored approaches, particularly in risk score and decision tree models, chosen in about 35-40% of cases. We did not limit the selection to feasible combinations of IML and imputation methods. As a result, surrogate splits for risk scores and linear regression were selected, even though they lacked a sound methodological basis. Implicit impu-

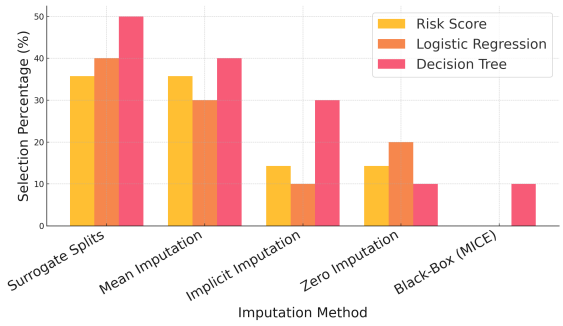


Figure 3: Clinician preferences for imputation methods across different IML models. We normalize by dividing the number who chose a combination by the total, as the total votes for a model can vary.

tation was more commonly selected for decision trees (∼30%), indicating its perceived alignment with tree-based models. Zero imputation, while moderately preferred (particularly in linear models, ∼25%), was ranked lower than other explicit methods. These results highlight a strong clinician preference for transparent imputation techniques.

Clinically reasoned preferences for missing values at test time. When choosing between implicit and explicit imputation methods, participants strongly preferred implicit imputation. This preference stems from the need for transparency, ease of understanding, and alignment with clinical reasoning, crucial in medical decision-making. One participant emphasized these factors, stating, "I prefer implicit imputation, as I'd find it hard to trust a black-box algorithm. Ease of understanding and communication are important, as well as from a 'medico-legal' point of view, to ultimately make a decision that involves my responsibility and the patient's life! [...] Even if the computer makes the decision, the responsibility [...] lies with me, and I must be able to understand and justify it." This highlights the importance of comprehensibility in high-stakes settings like healthcare. Another respondent noted, "I prefer implicit imputation because it seems simpler to reason with the values we have than to rely on a complex algorithm we don't understand." This reflects broader concerns among clinicians about the trustworthiness and transparency of black-box imputation. The practical aspects of clinical work were also highlighted by a participant who remarked, "Implicit imputation is more comprehensible and easier for clinicians to understand. Nevertheless, we need to see which technique is the best in terms of prediction, and I can't say what I prefer without knowing the relevance and PPV/VPn of each technique." This illustrates the balance clinicians must strike between understanding a method and ensuring its effectiveness. Finally, the intuitive nature of implicit imputation was underscored by another participant, who stated, "I prefer implicit imputation because, in my opinion, it's the closest thing to clinical reasoning: [...] it's the most intuitive and perhaps the safest approach, rather than an arbitrary rule." Some responses suggest sophisticated methods like MICE improve predictive performance by better preserving variable relationships than simpler methods like mean or zero imputation.

3.3. Finding 3: Requirements for designing IML models for clinical integration

Most respondents were open to using IML for diagnosis with missing values, though some were uncertain. A key concern was the need for *clear guidelines* on handling missing values. Clinicians emphasized the importance of structured explanations, with one suggesting that models should present multiple imputed values for intuitive selection, while another stressed the need for "a clear, simple, and rational explanation of how this missing data is treated, and its consequences on the test result." Without such transparency, trust in IML recommendations remains limited. Understanding and communicating *uncertainty* also emerged as a critical factor. Clinicians highlighted the challenge of making decisions with missing values and emphasized the importance of incorporating decision thresholds. One respondent noted that they could make a decision despite missing values but would "weight my decision with the patient's clinical data," underscoring the need for models to reflect confidence levels in their outputs. Beyond uncertainty, *model reliability* remains a fundamental requirement. Clinicians stressed the importance of ensuring that both models and imputation methods produce stable and trustworthy results across different clinical scenarios. Without this assurance, the adoption of IML in practice would be challenging. Equally important is the need for *interpretability and accessibility*, particularly for non-technical users. Some clinicians highlighted the value of *external val-*

idation and visual aids to clarify how missing values impact model outputs, making results easier to understand and act upon. Finally, *clinical intuition must remain central to decision-making.* Clinicians emphasized that IML should enhance, rather than replace, human expertise. One respondent stressed that "the human element and intuition based on experience and the CLINICAL EXAMINATION OF THE PATIENT!!!!" should continue to guide clinical decisions, highlighting the need for models that complement expert judgment rather than override it.

4. Related work

Our study aligns with previous findings that AI positively influences healthcare decision-making. van der Meijden et al. (2023) show in their study, that 97% of respondents were familiar with AI, and 86% believed it could aid clinical work. However, concerns remain about confirmation bias, over-reliance on models, and the effort needed for tool interaction (Wysocki et al., 2023). Our results suggest readiness for IML models, but successful implementation requires clinician participation in design to address needs like understanding missing data mechanisms and uncertainties (Lage et al., 2019a,b). While Janssen et al. (2010) argues that simple methods for handling missing data can lead to misleading results, others advocate for multiple imputation techniques as they enhance validity in clinical research (Austin et al., 2021). In contrast, our expert evaluation suggests implicit imputation, based on clinical intuition, is generally preferred at test time, while black-box models are the least favored due to interpretability issues. Notably, Read (2024) highlights that senior critical care physicians have traditionally relied on autonomy and clinical intuition. While this approach enables adaptability in handling complex and unpredictable scenarios, it also contributes to unwarranted variations in care delivery and potential harm. Following Tonekaboni et al. (2019), who emphasized evaluating clinical explainability in high-stakes settings, IML models must undergo rigorous validation to assess performance, interpretability, and legal implications (Wiens et al., 2019). Clinicians can build trust in IML models by understanding how they work and through positive experiences where outputs align with their domain knowledge (Kelly et al., 2019). Randomized controlled trials can validate models but are costly, time-consuming, and may lack generalizability (Shortliffe and Sepúlveda, 2018). Our findings suggest that successful deployment requires robust and purpose-built models, a clear presentation of uncertainties (Doshi-Velez and Kim, 2017), and training clinicians on the impact of missing values to integrate these tools into workflows (Topol, 2019; Maddox et al., 2019).

5. Discussion and conclusion

Our survey evaluates how clinicians interact with predictive models, particularly in handling missing values. Our findings reveal a disconnect between missing data handling in ML and retrospective research, such as MICE imputation, and clinical decision-making, where clinicians rely more on observed data and medical intuition. While imputation methods like zero, mean, or black-box approaches may aid model training, they are less favored at test time. Worst-case imputation, though clinically cautious, can introduce statistical bias and misrepresent patient risk levels.

To better align with clinical needs, future methods should either reduce reliance on imputation or adopt more transparent strategies. Native handling of missing values, as seen in tree-based models, can, for instance, be implemented using "Missingness Incorporated in Attributes" (MIA), which treats missing values as a distinct category Twala et al. (2008); Kapelner and Bleich (2015); Josse et al. (2024). This approach is similar to adding missingness indicators in linear models, although increasing the number of features can reduce interpretability. Stempfle and Johansson (2024), a generalized linear rule model minimizing reliance on missing values, avoids the need for imputation during training or testing, offering an intriguing direction for future research.

The diversity of clinician attitudes toward AI highlights the challenge of aligning models with clinical intuition and user preferences. Clinicians in Clusters 1 and 2, unfamiliar with IML, fearful of AI, or reliant on black-box models, would benefit from training on interpretable models, handling missing data, and AI's supportive role in practice. In contrast, Cluster 4, comprising highly engaged AI users, is well-suited for co-designing IMLs and refining missing value strategies. Capturing clinician intuition (Read, 2024) by integrating preferred practices in a way similar to computer programming holds great potential, particularly in complex or evidence-limited scenarios. Without clinician involvement, the adoption of these tools is unlikely, underscoring the need for IML methods that enhance clinician interaction with predictive models, particularly for handling missing values.

This study lays the foundation for understanding the challenges in deploying clinical prediction models with missing values. However, it is subject to self-selection bias due to clinicians' demanding workloads and limited availability. Future work could explore how these findings apply to other learning paradigms, like reinforcement learning, and different clinical settings, such as outpatient care,emphasising on the interaction between IMLs and clinicians. Expanding the study beyond the size of clinicians in France, we could also extend the study to include other healthcare systems for broader insights.

References

Peter C Austin, Ian R White, Douglas S Lee, and Stef van Buuren. Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331, 2021.

Leo Breiman. *Classification and regression trees*. Routledge, 2017.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Zhi Chen, Sarah Tan, Urszula Chajewska, Cynthia Rudin, and Rich Caruna. Missing values and imputation in healthcare data: Can interpretable machine learning help? In *Conference on Health, Inference, and Learning*, pages 86–99. PMLR, 2023.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

Richard P Dutton, Lynn G Stansbury, Susan Leone, Elizabeth Kramer, John R Hess, and Thomas M Scalea. Trauma mortality in mature trauma systems: are we doing better? an analysis of trauma mortality patterns, 1997–2008. *Journal of Trauma and Acute Care Surgery*, 69(3):620–626, 2010.

Ad Feelders. Handling missing data in trees: Surrogate splits or statistical imputation? In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 329–334. Springer, 1999.

Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of rule learning*. Springer Science & Business Media, 2012.

Sophie Rym Hamada, Anne Rosa, Tobias Gauss, Jean-Philippe Desclefs, Mathieu Raux, Anatole Harrois, Arnaud Follin, Fabrice Cook, Mathieu Boutonnet, Traumabase® Group, et al. Development and validation of a pre-hospital "red flag" alert for activation of intra-hospital haemorrhage control response in blunt trauma. *Critical Care*, 22:1–12, 2018.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Arthur James, Paer-Selim Abback, Pierre Pasquier, Sylvain Ausset, Jacques Duranteau, Clément Hoffmann, Tobias Gauss, Sophie Rym Hamada, and Traumabase Group. The conundrum of the definition of haemorrhagic shock: a pragmatic exploration based on a scoping review, experts' survey and a cohort analysis. *European Journal of Trauma and Emergency Surgery*, 48(6):4639–4649, 2022.

Kristel JM Janssen, A Rogier T Donders, Frank E Harrell Jr, Yvonne Vergouwe, Qingxia Chen, Diederick E Grobbee, and Karel GM Moons. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*, 63(7):721–727, 2010.

Xiaoyu Jin, Fu Xiao, Chong Zhang, and Ao Li. Gein: An interpretable benchmarking framework towards all building types based on machine learning. *Energy and Buildings*, 260:111909, 2022.

Julie Josse and François Husson. Handlin missing values in exploratory multivariate data analysis methods. *Journal de la Societe Française de Statistique*, 153(2):79–99, 2012.

Julie Josse and François Husson. missmda: a package for handling missing values in multivariate data analysis. *Journal of statistical software*, 70:1–31, 2016.

Julie Josse, Jacob M. Chen, Nicolas Prost, Gaël Varoquaux, and Erwan Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479, 2024.

Adam Kapelner and Justin Bleich. Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2):224–239, 2015.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019a.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67, 2019b.

Sébastien Lê, Julie Josse, and François Husson. Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25:1–18, 2008.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gael Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3165–3174. PMLR, 26–28 Aug 2020.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Yuan Luo. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1):bbab489, 2022.

Thomas M Maddox, John S Rumsfeld, and Philip RO Payne. Questions for artificial intelligence in health care. *Jama*, 321(1):31–32, 2019.

Vincent Margot and George Luta. A new method to compare the interpretability of rule-based algorithms. *AI*, 2(4):621–635, 2021.

Christoph Molnar. *Mol*. Lulu. com, 2020.

T.C. Nunez, I.V. Voskresensky, L.A. Dossett, R. Shinall, W.D. Dutton, and B.A. Cotton. Early prediction of massive transfusion in trauma: simple as abc (assessment of blood consumption)? *Journal of Trauma*, 66(2):346–352, Feb 2009. doi: 10.1097/TA.0b013e3181961c35.

M.S. Read. Algorithms that might replace clinical intuition in critical care. In D.W. Crippen, editor, *Transformations of Medical Education and Practice Impacting Critical Care in the New Millennium*. Springer, Cham, 2024. ISBN 978-3-031-69685-5. doi: 10.1007/978-3-031-69686-2_14.

Anthony J Rosellini and Timothy A Brown. Developing and validating clinical questionnaires. *Annual review of clinical psychology*, 17(1):55–81, 2021.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Cátia M Salgado, Carlos Azevedo, Hugo Proença, and Susana M Vieira. Missing data. *Secondary analysis of electronic health records*, pages 143–162, 2016.

Mahrukh Shakir and Atteq ur Rahman. Conducting pilot study in a qualitative inquiry: Learning some useful lessons. *Journal of Positive School Psychology*, 6(10):1620–1624, 2022.

Edward H Shortliffe and Martin J Sepúlveda. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200, 2018.

Lena Stempfle and Fredrik Johansson. Minty: Rule-based models that minimize the need for imputing features with missing values. In *International Conference on Artificial Intelligence and Statistics*, pages 964–972. PMLR, 2024.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Bheki ETH Twala, MC Jones, and David J Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019a.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019b.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Siri L van der Meijden, Anne AH de Hond, Patrick J Thoral, Ewout W Steyerberg, Ilse MJ Kant, Giovanni Cinà, and M Sesmu Arbous. Intensive care unit physicians' perspectives on artificial intelligence–based clinical decision support tools: Preimplementation survey study. *JMIR human factors*, 10:e39114, 2023.

Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 39 (2):519–581, 2023.

B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC)*, 1(3):1035, Dec 17 2013. doi: 10.13063/2327-9214.1035.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Oskar Wysocki, Jessica Katharine Davies, Markel Vigo, Anne Caroline Armstrong, Dónal Landers, Rebecca Lee, and André Freitas. Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making. *Artificial Intelligence*, 316:103839, 2023.

Xu Zhang, Lin Liu, Minxuan Lan, Guangwen Song, Luzi Xiao, and Jianguo Chen. Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94: 101789, 2022.

Appendix A. Survey details

Survey setup LimeSurvey is an open-source online survey tool that allows users to create, manage, and analyze surveys. It provides an easy-to-access interface where participants can complete surveys via a web link. For your usage, participating clinics could access LimeSurvey to fill out surveys or collect data in a structured format. The tool offers customizable survey templates, various question types, and features like branching logic to tailor the survey experience, making it user-friendly and efficient for collecting data from clinical settings.

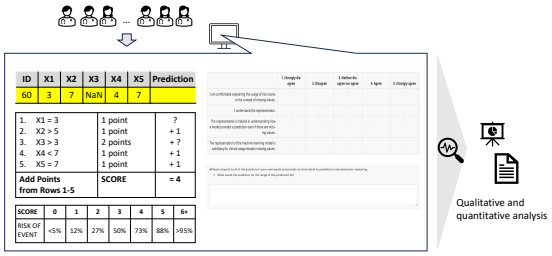


Figure 4: Experimental setup. Clinicians are shown a data entry of a patient with 5 features whereas one feature is missing in an interpretable machine-learning model along with the questions. After the answers are gathered qualitative and quantitative methods are used to analyze the results.

Use Case of Hemorrhagic Shock Diagnosing hemorrhagic shock is often complex, as there is no universally accepted minimal set of criteria.

Research, including a study by the Traumabase teams (James et al., 2022), has highlighted the coexistence of multiple definitions of shock. In clinical practice, the diagnosis is based on a combination of factors indicating decreased cardiac output and hemorrhage. These include patient history (e.g., potential sources of bleeding), clinical examination (e.g., hypotension, tachycardia, visible bleeding), and paraclinical tests (e.g., active bleeding on CT scan or ultrasound findings). The survey use case was designed with input from trauma specialists, and we can clarify this further in the paper.

Missing values in FAMD The primary goal of the Factor Analysis for Mixed Data (FAMD) (Josse and Husson, 2012) analysis and clustering is descriptive, allowing us to characterize clinicians from a multivariate perspective, rather than relying on just one or two variables to uncover patterns and group-specific insights. FAMD is the counterpart of PCA for mixed data, both continuous and categorical and is well-suited for describing results from questionnaires. Excluding incomplete responses could introduce bias if the missing data affects a specific subpopulation. Therefore, we compute the principal components using an EM algorithm, which maximizes the observed likelihood. In low-rank approximation methods like PCA or FAMD, estimation is equivalent to missing values and imputation. This means that minimizing the weighted least squares criterion or maximizing the observed likelihood yields the same results. The algorithms in missMDA (Josse and Husson, 2016) use an iterative SVD (Hastie et al., 2015) process: impute, compute SVD, update the imputation, and repeat. Thus, the imputation is "neutral" because the SVD on the imputed dataset produces the same results as on the dataset with missing values.

Appendix B. Survey questions

The survey contained along with the demographic questions four domains. The domains are: Physicians' Current Decision-Making Behavior in Diagnosing Hemorrhagic Shock in Trauma Patients (Current Decision-Making), Interpretation of ML models in the context of predicting with missing values at test time for a clinical use case (Interpretation of ML models and missing values), Preferences for using IML prediction model in daily workflows (Preferences in clinical care), Willingness to incorporate IML for prediction tasks in the context of missing values for diagnosing hemorrhagic shock and organ failure in clinical practice (Willingness)

In French: Les domaines sont les suivants: Comportement décisionnel actuel des médecins concernant le diagnostic du choc hémorragique chez les patients traumatisés (Opinion de l'Intelligence artificielle/Machine learning), Interprétation des modèles d'apprentissage automatique dans le contexte de la prédiction avec des valeurs manquantes au moment du test (Interprétation des modèles d'apprentissage automatique et des valeurs manquantes), Préférences pour l'utilisation du modèle de prédiction de l'IML dans les flux de travail quotidiens (Préférences de soins cliniques), Volonté d'incorporer l'IML pour les tâches de prédiction dans le contexte des valeurs manquantes pour le diagnostic du choc hémorragique et de la défaillance d'organe dans la pratique clinique (Volonté).

Appendix C. Additional results

This section presents further findings, including figures on the attitudes of different cohorts and the characteristics of identified clusters within these cohorts.

Table 2: All questions in survey in the English language

Domain	**Question**
Current Decision-Making Behavior	Q: Are you currently relying on decision support tools, such as risk scores, for diagnosing hemorrhagic shock in trauma patients? If yes, which ones and how do they support you? Q: Are you currently making use of any risk scores, e.g., RedFlag or ABC Score to support you in the diagnosis of hemorrhagic shock for trauma patients? If yes, please name the risk score and how you interpret the risk score when applying it to patients with missing values. Q: For which group of patients is the diagnosis of hemorrhagic shock the most difficult? Q: Which factors/features are most important when diagnosing hemorrhagic shock in trauma patients?
IML models and missing values at test time	Q: Please respond to all of the questions in your own words and provide as much detail as possible to understand your reasoning. What would your prediction (or the range of the prediction) be? Q: The average hemoglobin value in this data set for females is 12.70 g/dl. Does that knowledge affect your prediction assuming the patient is a woman? Q: Assuming the prediction made for the patient (no. 32450) was made by adding a zero for the missing value, which resulted in a positive prediction indicating they will have hemorrhagic shock, is the representation of risk scores despite the missing value helpful for your decision-making? Q: Do you trust the risk score? Motivate your answer briefly. Q: Please respond to all of the questions in your own words and provide as much detail as possible to understand your reasoning. What would the prediction (or the range of the prediction) be? Q: The average hemoglobin value in this data set for females is 12.70 g/dl. Does that knowledge affect your prediction assuming the patient is a woman? Q: Assuming that the coefficients were derived from a dataset completed using mean imputation, a positive prediction was then made for the patient. Would you trust that the patient has a hemorrhagic shock? Q: Do you trust the linear model? Motivate your answer briefly. Q: Please respond to all of the questions in your own words and provide as much detail as possible to understand your reasoning. What would the prediction (or the range of the prediction) be? Q: The average hemoglobin value in this data set for females is 12.70 g/dl. Does that knowledge affect your prediction assuming the patient is a woman? Q: Assuming the prediction made for the patient (no. 32450) was made by replacing the missing value using a surrogate split, a common approach in decision trees to handle missing values, which resulted in a positive prediction indicating they will have hemorrhagic shock. In a surrogate split, an alternative value is used to decide whether to go left or right in the tree. Is the representation of risk scores despite the missing value helpful for your decision-making? Q: Do you trust the decision tree? Motivate your answer briefly.
Preferences in clinical care	Q: Missing values in decision trees and risk scores can be handled by implicit imputation, which means that a value for the missing one can be imputed by considering the values of the other features. This might result in individual values depending on the clinician. Rank which method—decision tree or risk score—you think implicit imputation is more suitable for? Q: Another alternative is to use so-called black box models that perform the imputation by using methods that use complex, often non-transparent algorithms to fill in missing values in a dataset. Examples might be using Multivariate Imputation by Chained Equations (MICE). For which interpretable model would you most likely choose the black box imputation? Rank them according to your preference. Q: Do you prefer implicit or explicit imputation and why? You can consider aspects such as ease of understanding, usefulness in communicating your prediction, or intuitiveness. Q: Which missingness handling methods do you like best for each IML? Mark 3 in total with an x and add 0 in the other cells.
Willingness	Q: What is lacking in interpretable ML models for effective clinical use when dealing with missing values?

Table 3: Toutes les questions de l'enquête sont en français.

Domaine	**Question**
Opinion de l'IA/ML	Q: Utilisez vous actuellements des outils d'aide à la décision, tels que les scores de risque, pour diagnostiquer le choc hémorragique chez les patients traumatisés? Si oui, lesquels et comment vous aident-ils? Q: Utilisez-vous actuellement des scores de risque, par exemple RedFlag ou ABC Score, pour vous aider à diagnostiquer le choc hémorragique chez les patients traumatisés ? Dans l'affirmative, veuillez indiquer le score de risque et comment vous l'interprétez lorsque vous l'appliquezà des patients dont les valeurs sont manquantes. Q: Pour quel groupe de patients le diagnostic de choc hémorragique est-il le plus difficile? Q: Quels facteurs/caractéristiques sont les plus importants lors du diagnostic du choc hémorragique chez les patients traumatisés?
Interprétation des modèles IML	Q: Veuillez répondre à toutes les questions avec vos propres mots et fournir autant de détails que possible pour comprendre votre raisonnement. Quelle serait votre prédiction (ou la plage de prédiction)? Q: La valeur moyenne d'hémoglobine dans cet ensemble de données pour les femmes est de 12,70 g/dl. Cette connaissance affecte-t-elle votre prédiction en supposant que le patient est une femme? Q: En supposant que la prédiction faite pour le patient (n° 32450) a été réalisée en remplaçant la valeur manquante par un zéro, ce qui a abouti à une prédiction positive indiquant qu'il y aura un choc hémorragique, la représentation des scores de risque malgré la valeur manquante est-elle utile pour votre décision? Motivez brièvement votre réponse. Q: Faites-vous confiance au score de risque? Motivez brièvement votre réponse. Q: Veuillez répondre à toutes les questions avec vos propres mots et fournir autant de détails que possible pour comprendre votre raisonnement. Quelle serait la prédiction (ou la plage de prédiction)? Q: La valeur moyenne d'hémoglobine dans cet ensemble de données pour les femmes est de 12,70 g/dl. Cette connaissance affecte-t-elle votre prédiction en supposant que le patient est une femme? Q: En supposant que les coefficients étaient dérivés d'un jeu de données complété par imputation moyenne, une prédiction positive a alors été faite pour le patient. Seriez-vous convaincu que le patient présente un choc hémorragique? Q: Faites-vous confiance au modèle linéaire? Q: Veuillez répondre à toutes les questions avec vos propres mots et fournir autant de détails que possible pour comprendre votre raisonnement. Quelle serait la prédiction (ou la plage de prédiction)? Q: La valeur moyenne d'hémoglobine dans cet ensemble de données pour les femmes est de 12,70 g/dl. Cette connaissance affecte-t-elle votre prédiction en supposant que le patient est une femme? Q: En supposant que la prédiction pour le patient (no. 32450) a été faite en utilisant une technique appelée division de substitution pour traiter la valeur manquante, cette méthode est en effet couramment employée dans les arbres de décision. Dans une division de substitution, une autre valeur est utilisée pour décider du chemin dans l'arbre (gauche ou droite). Imaginons qu'une division de substitution ait été effectuée et qu'elle ait conduit à une prédiction positive indiquant un choc hémorragique. La représentation des arbres décisionnels avec la division de substitution pour traiter la valeur manquante vous aide-t-elle dans votre prise de décision? Q: Faites-vous confiance à l'arbre de décision? Motivez brièvement votre réponse.
Préférences de soins cliniques	Q: Les valeurs manquantes dans les arbres de décision et les scores de risque peuvent être traitées par imputation implicite, ce qui signifie qu'une valeur pour la valeur manquante peut être imputée en considérant les valeurs des autres caractéristiques. Cela peut entraîner des valeurs individuelles en fonction du clinicien. Classez quelle méthode (arbre de décision ou score de risque) pour laquelle vous pensez que l'imputation implicite est la plus adaptée? Q: Une autre alternative consiste à utiliser des modèles dits de boîte noire qui effectuent l'imputation en utilisant algorithmes plus complexes, souvent non transparents, pour combler les valeurs manquantes dans un ensemble de données. Des exemples pourraient être l'utilisation de l'imputation multivariée par équations chaînées (MICE). Pour quel modèle interprétable choisiriez-vous le plus probablement l'imputation par boîte noire? Classez-les selon vos préférences. Q: Préférez-vous l'imputation implicite ou explicite et pourquoi? Pouvez-vous prendre en compte des aspects tels que la facilité de compréhension, l'utilité de communiquer votre prédiction à un collègue ou l'intuitivité? Q: Quelles méthodes de gestion des absences préférez-vous pour chaque modèle interprétable? Marquez vos 3 favoris d'un x et ajoutez 0 dans les cellules.
Volonté	Q: Que manque-t-il aux modèles ML interprétables pour une utilisation clinique efficace lorsqu'il s'agit de valeurs manquantes?

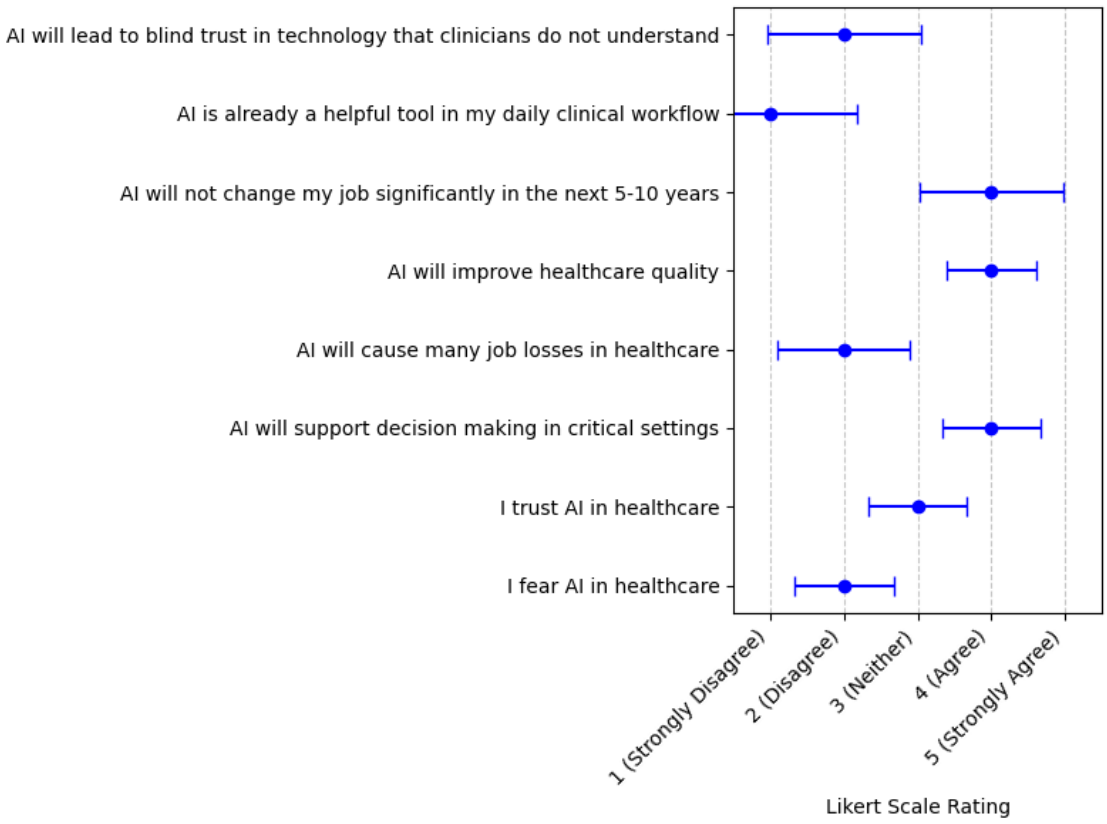


Figure 5: Cohort's attitude on AI/ML. We show the mode and standard deviation for eight statements exploring perspectives on AI/ML tools in healthcare, covering familiarity with these technologies, beliefs regarding their potential to replace physicians, expectations about their added value and support, and whether these tools adequately represent physicians' work to be useful. Missing values were not yet discussed.

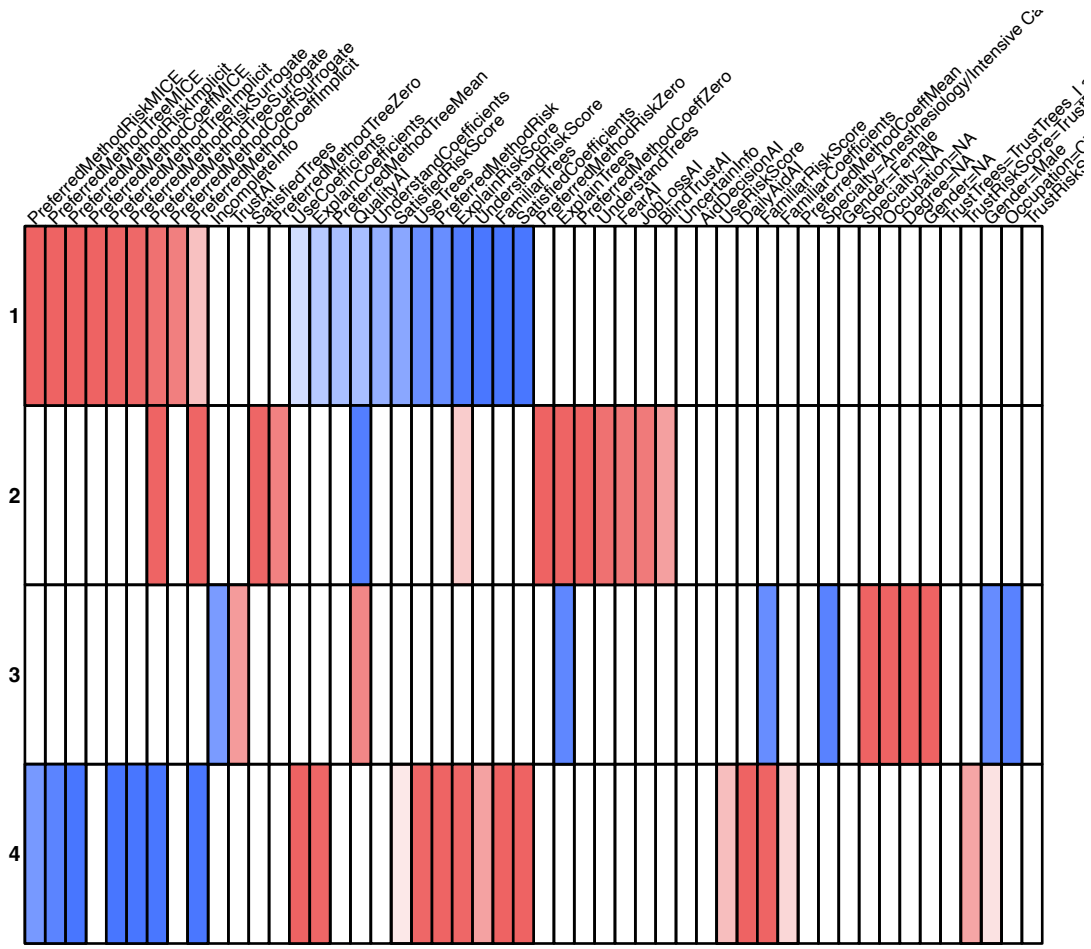


Figure 6: The cohort is divided into four clusters, each reflecting different attitudes towards AI/ML, varying levels of familiarity with IML, both within and outside the context of missing values and imputation preferences along with general trust in IML. In the color coding, blue indicates that the mean for this cluster is significantly lower than the global mean, while red indicates that the mean is significantly higher. White or light shades of blue and red suggest no significant difference within the group regarding the variable. Demographical features were only added to describe the clusters.