

Toward a Common Language for Trustworthy AI in Healthcare

Michael S. Ramirez Campos

ramirm7@mcmaster.ca

McMaster University

Rebecca Elliott

McMaster University

Thomas E. Doyle

McMaster University

Research Article

Keywords: Artificial Intelligence, Bias, Explainability, Fairness, Healthcare, Interpretability, Reliability, Robustness, Transparency, Trustworthy AI, Validity

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11078759/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Toward a Common Language for Trustworthy AI in Healthcare

Michael S. Ramirez Campos^{1*}, Rebecca Elliott¹ and
Thomas E. Doyle^{1,2,3*}

¹School of Biomedical Engineering, McMaster University, Hamilton,
ON, Canada.

²Department of Electrical and Computer Engineering, McMaster
University, Hamilton, ON, Canada.

³Department of Medical Imaging, McMaster University, Hamilton, ON,
Canada.

*Corresponding author(s). E-mail(s): ramirm7@mcmaster.ca;
doylet@mcmaster.ca;
Contributing authors: ellior@mcmaster.ca;

Abstract

Artificial Intelligence (AI) is increasingly integrated into healthcare, supporting tasks involving diagnosis, prognosis, treatment, and monitoring. As these systems become more complex, ensuring that they are trustworthy has become a central research priority. Trustworthy AI has emerged to address these requirements and promote trust among stakeholders. Yet, an important question remains: are the foundations on which trustworthy AI in healthcare is being built sufficiently well-defined? To address this question, we conducted a scoping review of 675 articles published between 2015 and 2025, including 41 review articles and 634 research articles. We examined how eight concepts central to trustworthy AI *Interpretability*, *Explainability*, *Transparency*, *Fairness*, *Bias*, *Robustness*, *Reliability*, and *Validity* have been defined and studied using thematic synthesis and semantic analysis. We found substantial conceptual overlap and inconsistency across the literature. In particular, *Interpretability*, *Explainability*, and *Transparency* were frequently used interchangeably, while contradictions and variations were also identified in the definitions of other concepts. Beyond these conceptual challenges, we characterized the current landscape of trustworthy AI in healthcare, revealing sustained growth in research across clinical domains, data types, AI approaches, and methods for *Interpretability*, *Explainability*, validation, *Fairness*, and *Bias*. However, the evidence remains unequally distributed,

with classical Machine Learning (ML), Electronic Health Records (EHR), neurology and oncology, SHAP, and hold-out validation predominating within their respective category groups. Based on these findings, we propose a framework that unifies definitions and clarifies relationships among the concepts examined. We further identify key research gaps in conceptual standardization, validation methods, clinical and data coverage, and AI approaches and tasks implementation. We conclude by identifying priorities for future research and proposing recommendations to strengthen the conceptual and evidentiary foundations of trustworthy AI in healthcare.

Keywords: Artificial Intelligence, Bias, Explainability, Fairness, Healthcare, Interpretability, Reliability, Robustness, Transparency, Trustworthy AI, Validity

1 Introduction

Artificial intelligence (AI) is defined as the ability of machines to simulate human intelligence [1, 2]. The use of AI in healthcare has been steadily increasing, with researchers continually developing new models for various healthcare applications [3–7]. Despite the exponential growth in the number of publications related to the development of AI models in healthcare, only a small fraction of these studies proceeds to real-world evaluation [8, 9]. In a review of 172 studies proposing AI models for intensive care units (ICUs), Fleuren et al. found that 93% were limited to prototyping or model development stages. In contrast, only 1% assessed the models using clinically relevant outcomes. Notably, none of the reviewed studies had integrated the proposed models into the actual workflows of clinical institutions [10]. Several studies have discussed the various barriers that hinder the integration of AI technologies into clinical practice. Berardi et al. proposes a division of the mental health system into four levels: macro (health policy), meso (healthcare organizations), micro (health professionals), and nano (patients). At the macro and meso levels, the main barriers identified include limitations in infrastructure, investment, coverage, lack of interdisciplinary collaboration, incentives for researchers, and legal regulations [11]. At the micro level, this and related studies report that insufficient knowledge, education, or training among health professionals represents key obstacles to the effective adoption of AI technologies [11–13]. Finally, at the nano level, challenges are associated with the limited adaptability of AI models to individual patients, biases in models, and a lack of information on vulnerable or underrepresented populations [11, 14–16]. To address these limitations, mainly those present at the micro and nano levels, which are inherently human-centered, the field of trustworthy AI has emerged. For an AI system to be considered trustworthy, it must not only be functional, efficient, and technically robust, but also adhere to ethical principles, ensure safety and transparency, prioritize human-centric design, and provide outputs and processes that are understandable to human users [17–19].

1.1 Trustworthy AI common challenges

The most common and visible challenge in this context is that most AI models, regardless of their type, tend to operate as black boxes. These are systems based on highly complex mathematical functions that are difficult to interpret and explain, even by domain experts [20]. As a result, understanding how these models function, why they produce specific outputs, and under what circumstances they may fail is often unclear [21, 22].

In addition, the concepts that form the basis of this field have not been systematically collected, analyzed, and unified. Our findings show that concepts such as *Explainability*, *Interpretability*, and *Transparency* are used interchangeably. Also, the definitions of *Bias*, *Fairness*, *Reliability*, and *Validity* proposed by different authors present contradictions and inconsistencies. This indicates that there is no consensus on the fundamentals of trustworthy AI. Although these concepts and their definitions will be presented and discussed in later sections, it is important to mention them here because inconsistencies in these fundamental concepts represent a significant obstacle to the implementation of trustworthy AI in real-life settings. Furthermore, to the best of our knowledge, no article has addressed this problem as comprehensively as this review.

In this sense, this article has three main objectives: 1) Analyze the patterns, similarities, and differences that emerge in the definitions and use of concepts related to trustworthy AI in the literature; 2) Propose a conceptual framework using unified definitions of the concepts related to trustworthy AI; 3) Characterize the current landscape of trustworthy AI in healthcare and identify knowledge gaps in the existing literature. Through a comprehensive analysis of key concepts and methodological approaches, the review addresses the following research questions:

- RQ1. How have trustworthy AI concepts been studied and implemented in healthcare contexts?
- RQ2. What are the current research trends, methodological characteristics, and knowledge gaps in trustworthy AI for healthcare?

To address the research questions, a systematic search was conducted to identify relevant review and research articles, select those aligned with key topics, extract data and relevant results, propose a conceptual definition framework using the review articles, and report the findings of the research articles.

1.2 Historical context

In 1943, McCulloch and Pitts proposed the first mathematical model of an artificial neural network based on the structure of biological neurons [23]. Later, in 1950, Alan Turing introduced the idea of machine intelligence through the Imitation Game (now known as the Turing Test), in which he posed the famous question, “*Can machines think?*” [24]. The term *Artificial Intelligence* was first introduced in 1955 when McCarthy et al. proposed the Dartmouth Summer Research Project on Artificial Intelligence for the summer of 1956. The proposal was based on the conjecture that “every aspect of learning or any other feature of intelligence can in principle be

114 so precisely described that a machine can be made to simulate it” [25]. Since then, the
115 field has evolved rapidly, giving rise to a wide range of AI methods and applications
116 across numerous domains.

117 Although the evolution of trustworthy AI is closely linked to that of AI, it was
118 not until 1978 that MYCIN, the first rule-based expert system capable of provid-
119 ing explanations, was proposed [26, 27]. In the early 1990s, as part of the Strategic
120 Computing Initiative funded by the Defense Advanced Research Projects Agency
121 (DARPA) [28], Swartout et al. introduced one of the first design frameworks for
122 explainable expert systems, proposing explanation mechanisms that extended beyond
123 simple rule tracing [29]. Later, in 2004, Van Lent et al. presented one of the earliest
124 documented uses of the acronym *XAI* (eXplainable Artificial Intelligence), referring
125 to an explainable AI feature integrated into the Full Spectrum Command system [30].
126 The acronym became widely adopted following the launch of DARPA’s Explainable
127 Artificial Intelligence (XAI) program in 2015, whose objective was to develop expla-
128 nation techniques that enable end users to understand, trust, and effectively manage
129 AI systems [31, 32]. This program marked the beginning of the modern resurgence
130 of XAI research. Subsequently, widely adopted techniques such as Local Interpretable
131 Model-agnostic Explanations (LIME) (2016) [33] and SHapley Additive exPlanations
132 (SHAP) (2017) [34] were introduced. Furthermore, in 2016, the *IEEE Global Initiative*
133 *for Ethical Considerations in Artificial Intelligence and Autonomous Systems* aimed to
134 promote AI systems that prioritize human well-being and adhere to ethical principles,
135 including Human Benefit, Responsibility, Transparency, and Education and Awareness
136 [35]. At the governmental level, Canada launched the Pan-Canadian Artificial Intel-
137 ligence Strategy (PCAIS) in 2017, investing approximately CAD\$125 million in six
138 strategic areas: (1) Talent and Research, (2) National AI Institutes, (3) Commercial-
139 ization and Adoption, (4) Standards and Conformity, (5) Compute and Infrastructure,
140 and (6) Responsible AI and Ethics [36]. At the international level, the European
141 Union’s General Data Protection Regulation (GDPR), which became enforceable in
142 2018, established a legal framework for protecting personal data by requiring orga-
143 nizations to comply with principles such as lawfulness, fairness, transparency, data
144 integrity, confidentiality, and accountability [37].

145 1.3 Background

146 A growing body of research has focused on developing explanation methods that aim
147 to increase transparency and user trust. These techniques typically identify the input
148 features or attributes that contribute most to a model’s predictions, either through fea-
149 ture importance measures or, in the case of neural networks, gradient-based approaches
150 [38–40]. XAI methods can generally be categorized into model-agnostic and model-
151 specific approaches. Model-agnostic methods can be applied to a wide range of AI
152 models because they explain predictions by analyzing the relationship between inputs
153 and outputs without requiring access to the model’s internal mechanisms. In contrast,
154 model-specific methods are tailored to a particular model architecture and leverage its
155 internal structure to provide insights into the model’s predictions and, in some cases,
156 its internal representations [41]. XAI methods can also be categorized as ante-hoc
157 or post-hoc approaches. Ante-hoc approaches, often referred to as glass-box models,

are inherently interpretable and explainable. A decision tree is a common example, as its internal structure can be directly examined to understand the rules underlying its predictions. In contrast, post-hoc approaches aim to explain the predictions of already trained models. For example, methods such as SHAP estimate the contribution of each input feature to a prediction by analyzing how changes in the input affect the model’s output [42]. In this section, we provide an overview of the XAI methods most frequently identified in our review. We strongly recommend to check the book by Christoph Molnar, ”Interpretable Machine Learning: A Guide for Making Black Box Models Explainable”, to get a deeper understanding of the concepts and methods explained here [43].

1.3.1 SHAP

SHAP is a model-agnostic post-hoc technique based on game theory that aims to quantify the contribution of each feature to a model’s prediction for a specific input instance. The calculation of the contribution of a feature j involves considering all possible subsets of features that do not include j . The trained model is then used to generate predictions for each subset, both with and without feature j . The difference between these predictions represents the marginal contribution of j (i.e., its marginal contribution to the prediction, indicating whether the feature increases or decreases the model output) across all possible feature combinations [34]. This contribution is calculated using the Shapley value formulation (Eq. (1)), introduced by Lloyd Shapley in 1953 [44].

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (V(S \cup \{j\}) - V(S)) \quad (1)$$

In this equation, ϕ_j represents the contribution value of feature j ; S represents a subset of features in a given combination; N is the total number of features; $V(S)$ represents the model’s prediction using the set of features in S ; $V(S \cup j)$ represents the model’s prediction when feature j is added to set S ; $(V(S \cup j) - V(S))$ represents the marginal contribution of feature j ; $\sum_{S \subseteq N \setminus j}$ represents the summation over all possible feature subsets that do not include j ; and $\frac{|S|!(|N| - |S| - 1)!}{|N|!}$ represents the weight assigned to each feature combination. An example of SHAP implementation can be found in the tutorial by Ponce-Bobadilla et al. [45].

1.3.2 LIME

LIME is a model-agnostic post-hoc technique that explains the prediction of a specific input instance by constructing a locally interpretable approximation of the original model. This approximation is typically a simple linear model. To build it, LIME generates perturbed versions of the input instance and obtains the corresponding predictions from the original model. The resulting surrogate model estimates the contribution of each feature to the prediction through its coefficients, providing an interpretable approximation of the model’s local behavior [33].

From a mathematical perspective, LIME can be described using three main equations. Considering x as the original input instance and z as a perturbed instance,

the first step consists of calculating an exponential kernel with width σ^2 , which incorporates a distance function $D(x, z)$ to measure the similarity between both instances (Equation (2)).

$$\pi_x(z) = \exp\left(-\frac{D(x, z)^2}{\sigma^2}\right) \quad (2)$$

Equation (3) calculates the local fidelity of the local model. Here, z' is the representation of z used by the local model. Similarly, $\pi_x(z)$ represents the proximity weight assigned to the perturbed instance; $f(z)$ represents the prediction generated by the original model; $g(z')$ represents the prediction generated by the local model; and $\sum_{z, z' \in Z}$ represents the summation over all perturbed instances within the set Z .

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) (f(z) - g(z'))^2 \quad (3)$$

Finally, Equation (4) defines the optimization objective used to identify the optimal local model g that minimizes the combination of $L(f, g, \pi_x)$ (local fidelity) and the model complexity term $\Omega(g)$. Here, G represents the set of possible local models considered during the optimization process.

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (4)$$

1.3.3 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a model-specific post-hoc technique that explains and visualizes the predictions of deep learning models, particularly Convolutional Neural Networks (CNNs) [46, 47]. Grad-CAM computes the importance of each feature map in the last convolutional layer by using the gradients of the pre-softmax class score (logit) of a target class with respect to the activations of that layer. Equation (5) computes the importance weight α_k^c associated with feature map k for class c . Here, A^k denotes the k -th feature map of size $i \times j$; K is the total number of feature maps; Z is the number of elements in A^k ; and $\frac{\partial y^c}{\partial A_{ij}^k}$ represents the gradient of the pre-softmax class score y^c with respect to the activation at position (i, j) in feature map A^k .

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (5)$$

The final class activation map $L_{\text{Grad-CAM}}^c$ is obtained by applying a ReLU function to the weighted combination of feature maps $\sum_k \alpha_k^c A^k$ (Equation. (6)). The ReLU operation discards negative values and retains only the regions with positive contributions to the target class score, indicating the areas of the input representation that most strongly support the model's prediction for that specific class.

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (6)$$

1.3.4 Gini Importance

Gini Importance is a model-specific post-hoc method used for decision tree-based models, such as random forests and gradient boosting. It estimates feature relevance based on the reduction in Gini Impurity resulting from splits involving each feature. Gini Impurity is computed at each node where the model uses a feature to define a split and make a decision [48, 49]. Equation (7) defines the Gini Impurity $G(a)$ for a node a . Here, N is the total number of samples at node a , and N_k is the number of samples belonging to class k .

$$G(a) = 1 - \frac{1}{N^2} \sum_k^K N_k^2 \quad (7)$$

When a split is generated, the initial node a is referred to as the parent node. Equation (8) defines the reduction in Gini Impurity resulting from a split, where $G(a_L)$ and $G(a_R)$ are the Gini Impurity values of the resulting child nodes, and N_L and N_R are the numbers of samples in each child node, respectively.

$$\Delta G(a) = G(a) - \frac{1}{N} (N_L G(a_L) + N_R G(a_R)) \quad (8)$$

The Gini importance of a feature j in a single tree t is defined as the sum of the Gini impurity reductions $\Delta G(a)$ over the set of nodes S at which feature j is used to perform a split (Equation (9)).

$$I_{tj} = \sum_{a \in S_j} \Delta G(a) \quad (9)$$

Equation (10) defines the normalized Gini feature importance for an ensemble of decision trees. The term $\sum_{t=1}^{\tau} I_{tj}$ represents the sum of the Gini importance of feature j across all τ trees. The denominator, $\sum_{m=1}^M \sum_{t=1}^{\tau} I_{tm}$, represents the total Gini importance across all M features and τ trees.

$$FI_j = \frac{\sum_{t=1}^{\tau} I_{tj}}{\sum_{m=1}^M \sum_{t=1}^{\tau} I_{tm}} \quad (10)$$

1.3.5 Permutation importance

Permutation importance is a model-agnostic, post-hoc method that quantifies the effect of randomly shuffling the values of a specific feature on model performance [43]. Permutation importance was originally introduced by Leo Breiman (2001) as a tool for Random Forests[50]. It was later mathematically formalized and generalized into a model-agnostic framework known as Model Reliance by Fisher, Rudin, and Dominici (2019) [51]. To calculate the permutation importance of a feature j , we first compute the model's error e_{orig} when making predictions on the original test set. Equation (11)

252 represents the calculation of the original error e_{orig} , where N_{test} is the total number of
 253 samples in the test set, and $\sum_{i=1}^{N_{test}} L(y^{(i)}, f(\hat{x}^{(i)}))$ represents the sum of an arbitrary
 254 loss function L evaluated between the true label $y^{(i)}$ and the model prediction $f(\hat{x}^{(i)})$.

$$e_{orig} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} L(y^{(i)}, f(\hat{x}^{(i)})) \quad (11)$$

255 Equation (12) shows the calculation of the error $e_{perm,j}$ for the matrix $X_{perm,j}$,
 256 where the values of j are permuted.

$$e_{perm,j} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} L(y^{(i)}, f(\hat{x}_{perm,j}^{(i)})) \quad (12)$$

257 Then, the original error e_{orig} is compared with the error $e_{perm,j}$ corresponding to
 258 the predictions using the matrix $X_{perm,j}$. Equations (13) and (14) show two ways to
 259 calculate the final feature importance.

$$FI_j = \frac{e_{perm,j}}{e_{orig}} \quad (13)$$

$$FI_j = e_{perm,j} - e_{orig} \quad (14)$$

260 2 Result

261 A total of 675 articles were included, of which 41 were review articles and 634 were
 262 research articles. Table 1 presents the unified definitions proposed for the concepts
 263 of interest, while Figure 1 illustrates the relationships among these concepts using a
 264 Venn diagram. Similarly, Section 2.2 section presents the most relevant data extracted
 265 from the included research articles.

266 2.1 Definitions proposal

267 This section presents the intermediate findings that led to the unified definitions pro-
 268 posed in Table 1. The relationships among these concepts are illustrated in Figure 1.
 269 These findings are the result of a mixed-methods analysis combining thematic synthesis
 270 (Sections 2.1.1 and 5.4.1) and semantic analysis (Sections 2.1.2 and 5.4.2).

271 2.1.1 Thematic synthesis

272 The thematic synthesis identified 108 definitions, from which 25 codes were generated.
 273 The codes with the highest number of associated definitions were “*Explainability*” (27
 274 definitions), “*Interpretability*” (26 definitions), “*Transparency*” (24 definitions), and
 275 “*Explainable–Interpretable*” (19 definitions). The latter code was created to identify
 276 studies that treated *Explainability* and *Interpretability* as interchangeable concepts
 277 or synonyms. The complete distribution of definitions across all codes is presented in
 278 Supplementary Figure S1. Furthermore, this synthesis revealed several differences and
 279 contradictions among the definitions proposed for the same concept:

- *Reliability*: [52, 53] define a reliable model as one that produces consistent results under similar conditions. In contrast, [54–56] define reliability as the ability to produce consistent results under different conditions, such as across datasets or clinical settings.
- *Bias*: [55, 57–60] define bias as a systematic, non-random error/behavior that may generate discrimination, harm, or inequity. Alternatively, [61–66] emphasize that bias can originate from both the data and the algorithm, making it a potential source of systematic errors.
- *Validity*: [67] define validity as the ability of a model to maintain its performance on external datasets (different from the training dataset). Meanwhile, [68] states that this is defined by the end user based on the model’s accuracy.

2.1.2 Semantic analysis

All review articles that provided definitions of the concepts of interest were included, regardless of whether those definitions were originally proposed by the authors or adopted from previous studies, several source works were cited repeatedly across the included articles. Among the most frequently cited were the review by Arrieta et al. (4 citations), which defines *Explainability*, *Interpretability*, *Transparency*, and *Fairness* [69]; Miller T. (3 citations), who discusses *Robustness* and *Interpretability*, and states that *Interpretability* and *Explainability* are equivalent concepts [70]; Lipton ZC (2 citations), who focuses on *Interpretability* and also defines *Transparency* [71]; and the OECD recommendations (2 citations), which, although they do not provide explicit definitions, were used by several studies as a conceptual basis [72]. We consider these studies as source articles. However, we did not find many articles that consistently cited the same reference for a given definition.

In addition, the definitions reported by [66] and [58] showed a high degree of similarity for *Explainability*, *Fairness*, and *Robustness*, likely because both publications originated from the same research group. To avoid redundancy in the thematic synthesis, we retained the definitions from the earlier publication [66]. The most recent publication was also included because it introduced a definition of *Bias* that was not present in the earlier work. The complete set of extracted definitions, together with their corresponding thematic labels, is provided in the Supplementary Material.

2.2 Research articles

This section summarizes the data extracted from the 634 research articles included in this study. In some figures, the sum of the category counts exceeds the total number of articles because a single study could be assigned to multiple categories within the same variable. For example, one article may have used signals, medical images, and EHRs to train its models, while also applying both ML and deep learning (DL) methods for classification and regression tasks. Section 2.2.1 presents general information about the included articles. Section 2.2.2 compares the concepts addressed by the authors with the definitions proposed in Table 1. Section 2.2.3 summarizes the AI tasks performed and the models implemented. Section 2.2.4 presents the XAI techniques used. Finally, Sections 2.2.5 and 2.2.6 describe the validation techniques and the methods used to

Concept	Definition
Explainability	Explainability is the ability of a model to provide a human-understandable justification of model behavior and predictions. This is achieved by offering insights into the decision-making processes in a clear way, enabling users to comprehend the reasoning behind decisions. Explainability focuses on describing the mechanisms that drive decisions, revealing the system’s strengths and weaknesses, and giving clear guidance to users so they can recognize why and how (as long as they have basic knowledge of the field) decisions are made.
Interpretability	Interpretability is the ability to make the model’s behavior, decision-making, and reasoning human-understandable through clear insights. Focusing on ensuring that the decisions are easy for users (whether experts or non-experts) to comprehend. Interpretability addresses why decisions are made, but not necessarily how they are made, as this is covered by explainability, which provides more technical details.
Transparency	Transparency is the property of a model to disclose information about its structure and functionality, including the data used for training, the training process, and the decision-making processes, in a manner that is understandable to stakeholders (patients, healthcare professionals, caregivers, and policymakers). It provides information to answer why and how decisions are made, while also helping to identify data and model biases. Transparency serves as the foundation for model accessibility and trust among stakeholders, representing the first step toward achieving genuine patient–clinician understanding.
Bias	Bias is a model’s tendency to exhibit specific decision-making behavior due to the use of data from particular characteristics and/or populations during the training process, potentially leading to systematic non-random errors and, consequently, discriminatory results or misdiagnoses. Bias is an inherent property of the model and may originate from both the data and the coding process.
Fairness	Fairness is the property of a model related to equitable decision-making, aiming to address and reduce algorithmic biases and discriminatory outcomes. The underrepresentation of marginalized user populations due to systemic inequalities in training data can lead to unfavorable models that contribute to healthcare disparities. Fairness is closely related to model robustness. However, it is more specific, as it concerns the model’s performance not across different conditions, but across populations with particular characteristics.
Robustness	Robustness is the property of a model to maintain consistent and stable performance despite variations, noise, and differences in input data. This property also applies to techniques designed to provide explainability, which should generate accurate explanations despite such variations.
Reliability	Reliability is the property of a model to maintain consistent and stable performance under the same conditions or scenarios. In other words, the model should yield consistent outputs when exposed to the same inputs and environment.
Validity	Validity is the property of a model to perform as intended, which depends on both robustness and reliability, and it is not achieved if either component is absent.

Table 1 Definitions proposal of trustworthy AI concepts

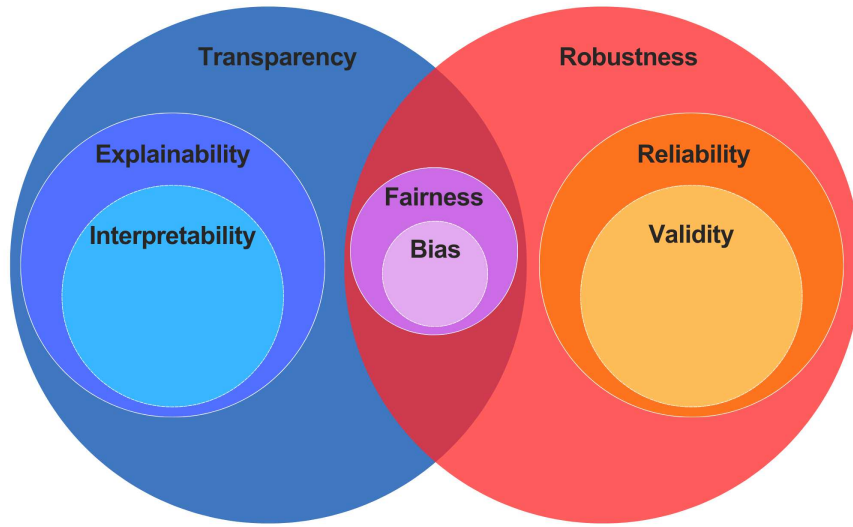


Fig. 1 Venn diagram that represents the relationship between the proposed definitions.

address fairness, respectively. Due to the large number of categories in some of the topics, it was decided to group some of them. The table containing the extracted data can be found in the supplementary material.

2.2.1 General information

The general information extracted from each study included the year of publication, the geographical region of the authors' affiliations, the clinical domain, and the types of data used to train the AI models. Figure 2 summarizes the distribution of the categories within each topic. The review included studies published between January 1, 2015, and November 2025. Because relatively few studies were published between 2015 and 2019, these years were grouped into a single category.

Regarding the geographical region, studies with at least one author affiliated with an institution outside the predominant region were classified as "International". In addition, the "Others" category comprises "Oceania", "Sub-Saharan Africa", "Latin America and the Caribbean", and "Eastern Europe and Russia".

The clinical domains were defined to encompass the broadest possible range of medical specialties. The "Not specified" category includes studies whose primary focus was not a specific aspect of human health but rather healthcare processes or systems, such as emergency department readmissions. In contrast, "General / Multisystem" refers to studies that were not limited to a particular organ, physiological system, or disease, including, for example, studies that analyzed medical images from multiple anatomical regions or predicted mortality.

343 Finally, the data used to train the AI models were grouped into six main cate-
 344 gories. Demographic information was included within the “Electronic Health Records”
 345 category.

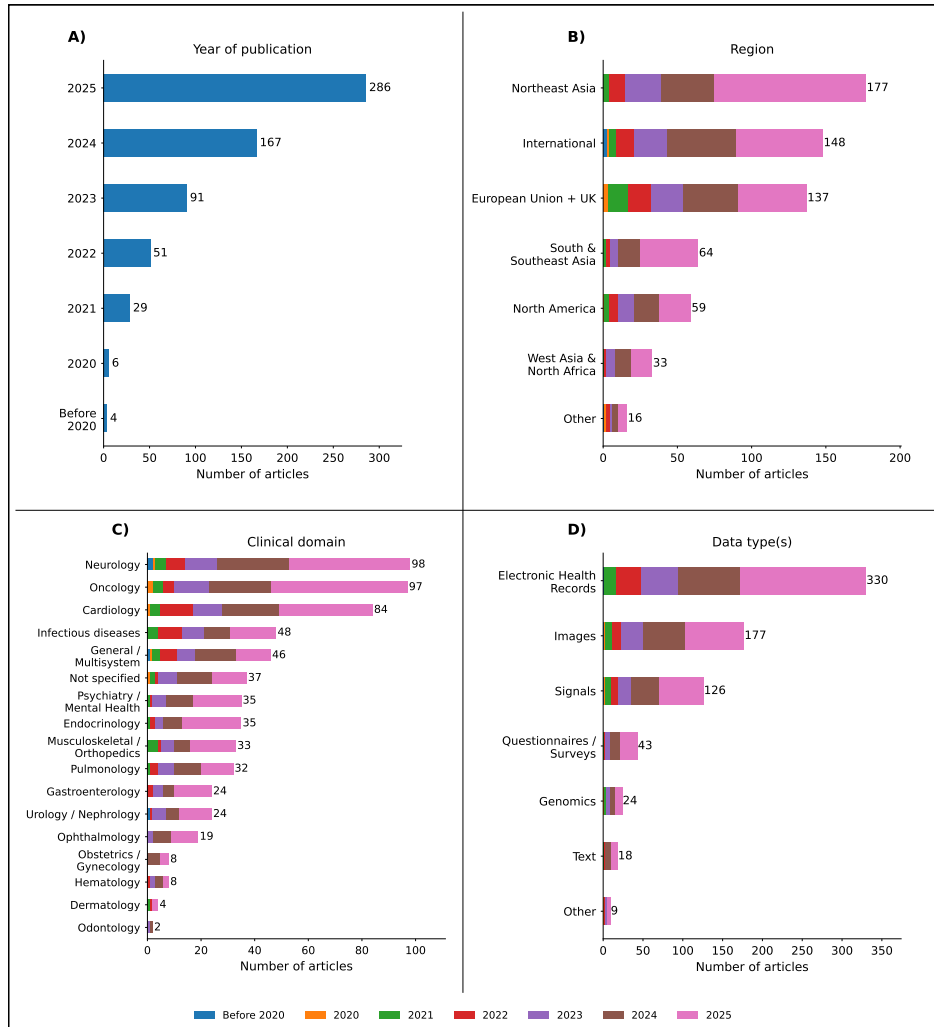


Fig. 2 Overview of the general information extracted from the included studies. (A) Number of publications by year (2015–2025). (B) Geographical regions of the authors' affiliations. (C) Clinical domains addressed in the included studies. (D) Distribution of the data types used to train the AI models.

346 2.2.2 Concepts addressed

347 The concepts explicitly discussed in each study, together with literal mentions, were
 348 used to identify the concepts addressed according to the authors. In contrast, the XAI

349 techniques employed, the use of external datasets for validation, the methods used
 350 to address bias, and the analyses performed on the developed models were used to
 351 identify the concepts addressed according to the definitions proposed in this study.
 352 **Figure 3** compares these two perspectives.

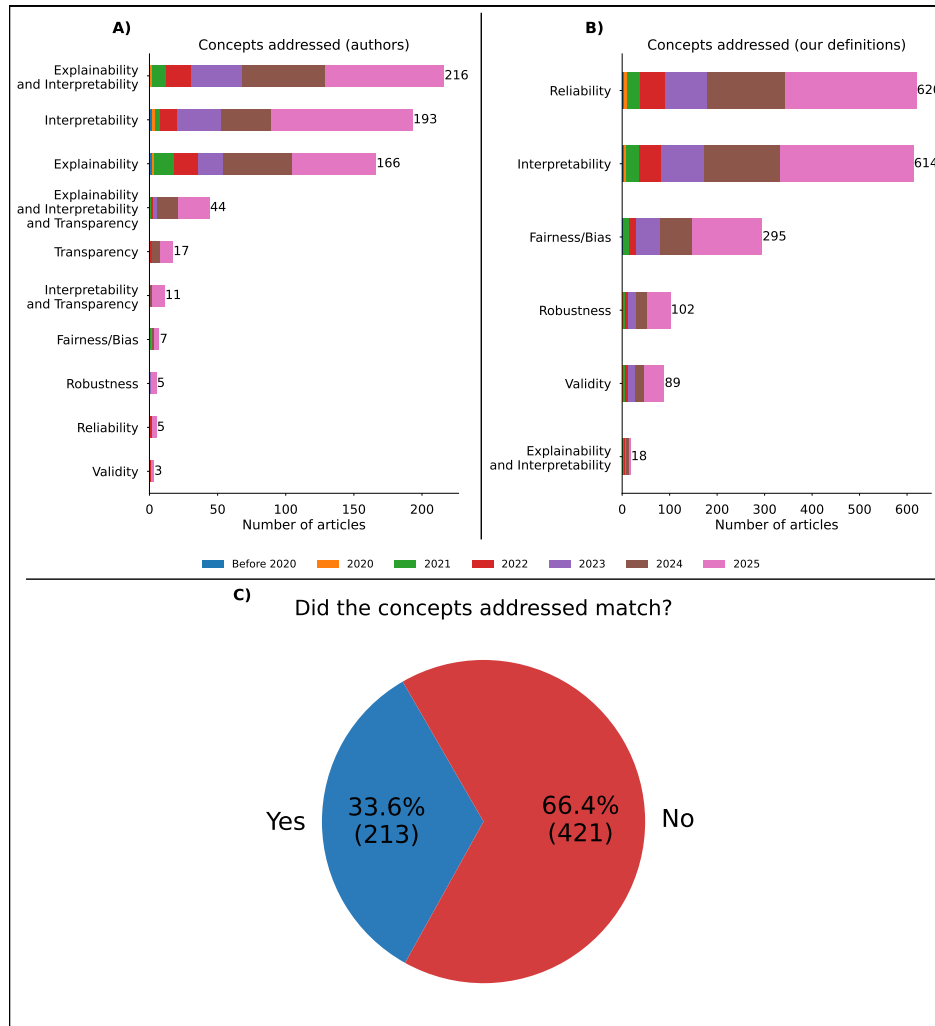


Fig. 3 Comparison of the concepts addressed in the included studies according to the authors and the definitions proposed in this study. (A) Distribution of the concepts addressed according to the authors. (B) Distribution of the concepts addressed according to the definitions proposed in this study. (C) Percentage and number of studies showing agreement and disagreement between both perspectives.

2.2.3 XAI techniques

Figure 4 summarizes the XAI techniques used in the analyzed studies. The “Feature importance” category groups together methods for assessing feature relevance, including Gini importance and permutation importance (Sections 1.3.4, 1.3.5). However, it is not feasible to list all the techniques included in this category because several studies proposed their own methods for estimating feature importance. Similarly, the “Others” category comprises less frequently used methods, such as Testing with Concept Activation Vectors (TCAV). The findings presented in this section were used to support the comparisons described in Section 2.2.2.

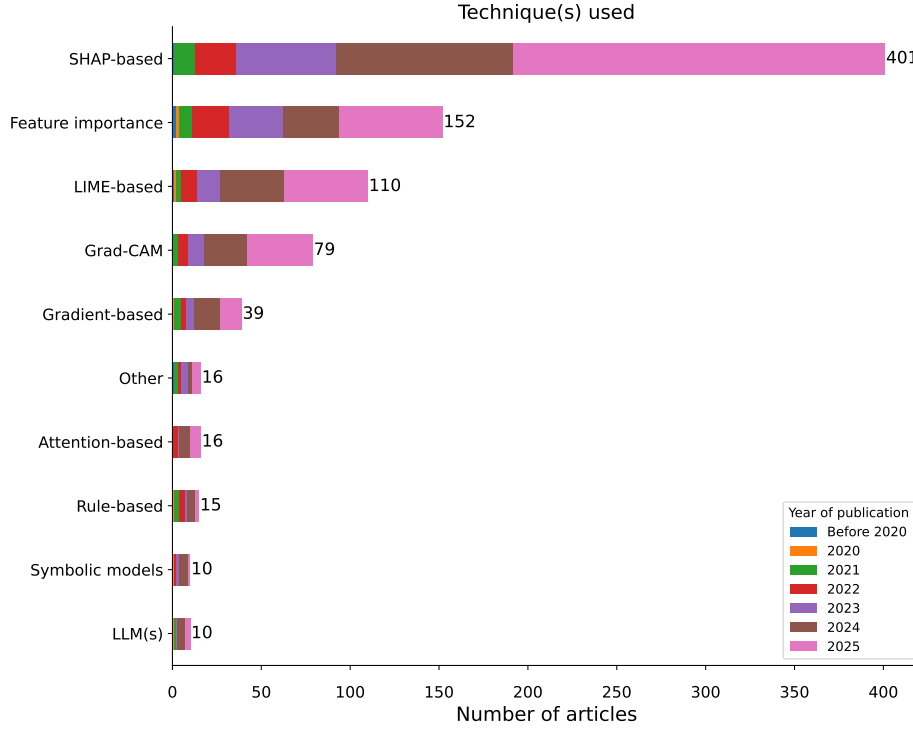


Fig. 4 Distribution of XAI techniques used in the analyzed articles.

2.2.4 Implementations

Figure 5 summarizes the AI models implemented in the included studies. The model categories comprise the most frequently used AI model types, as well as hybrid approaches combining DL models for feature extraction with ML models for prediction. In this context, the “Other” category includes less frequently used models, such as those based on swarm intelligence. Similarly, the summary of model families shows that most studies trained multiple models for comparison. Regarding the tasks performed, “Classification” and “Regression” were the predominant categories.

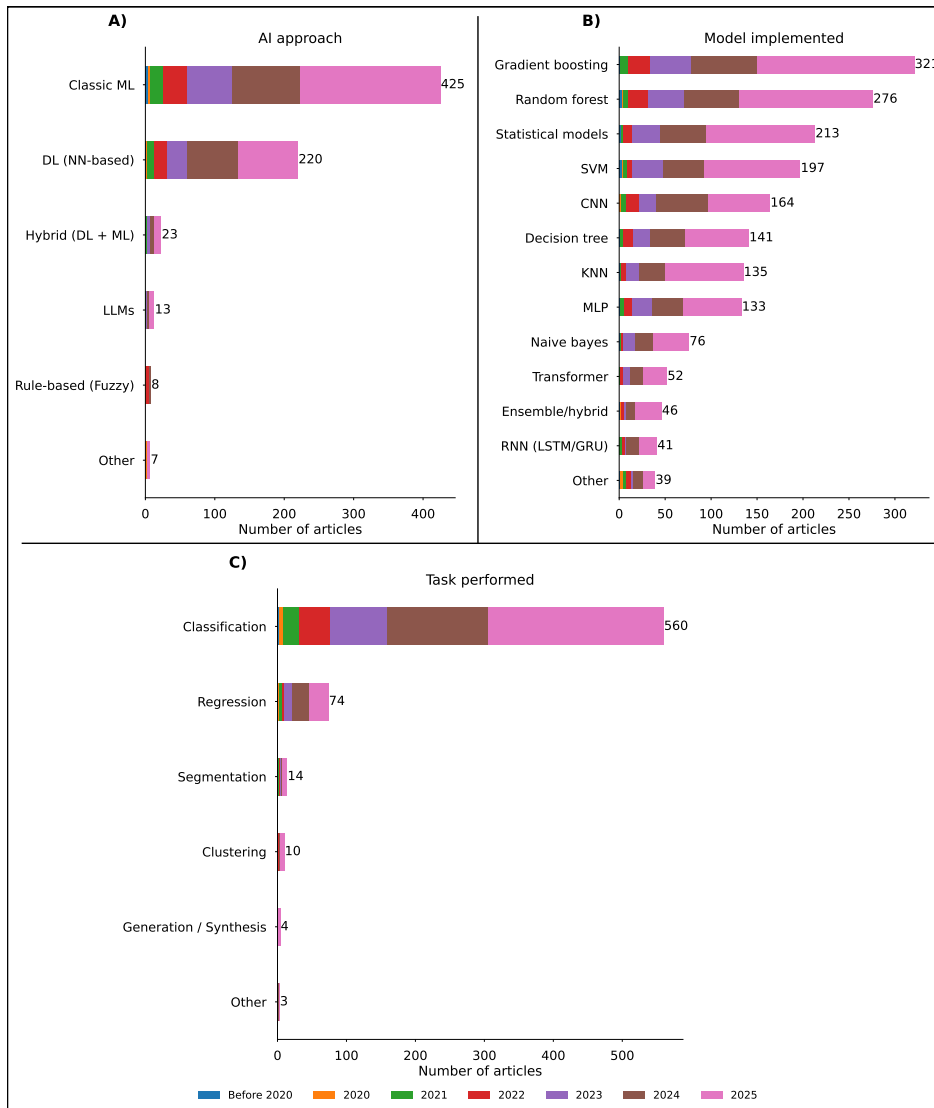


Fig. 5 Overview of the models used in the analyzed studies. A) Types of models used; the “Hybrid” category refers to a combination of models for feature extraction (usually DL) and subsequent prediction; B) Families of models used; C) Task performed by the models used.

2.2.5 Validation techniques

Validation techniques provide evidence to assess the *Robustness*, *Reliability*, and *Validity* of AI models according to the definitions proposed in Table 1. Figure 6 summarizes the distribution of the validation methods used in the included studies.

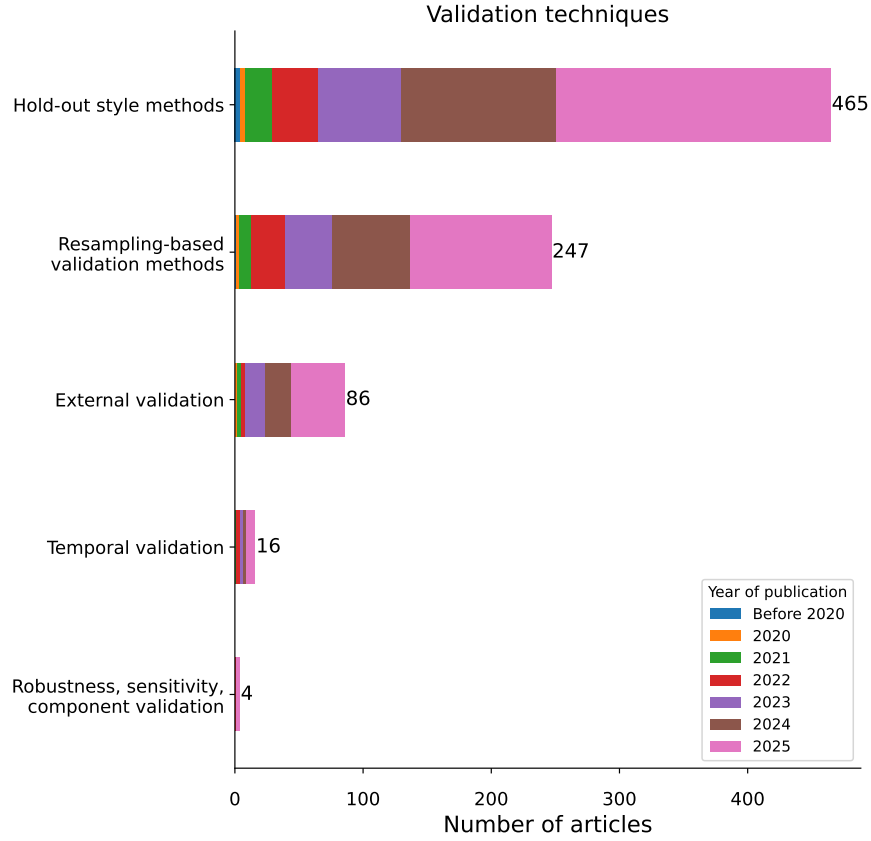


Fig. 6 Frequency of the validation methods identified in the included studies.

2.2.6 Fairness addressing techniques

The techniques used here are closely related to bias mitigation, which, according to Table 1, is the model's tendency to exhibit a specific behavior. Figure 7 summarizes the frequency of the methods used in the included studies to address this issue. Most of these methods mitigate bias by modifying the data distribution, either by balancing the number of samples across classes or by assigning class weights during training to reduce the model's tendency to favor a particular class. The category "No method applied" indicates that the study did not report any strategy for addressing bias.

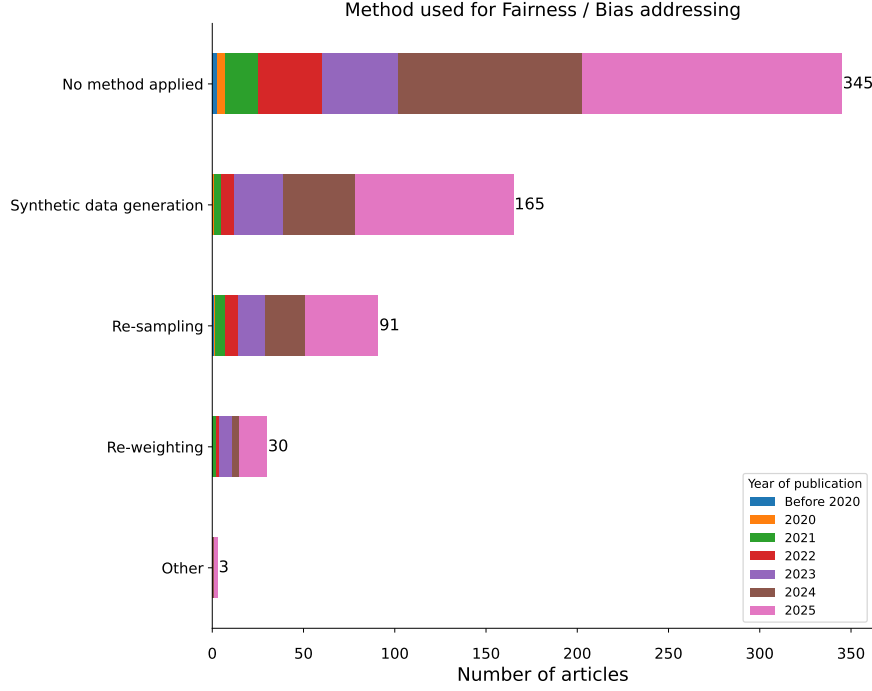


Fig. 7 Distribution of the bias mitigation methods reported in the included studies.

3 Discussion

Trustworthy AI is a field that has emerged to address limitations related to users' lack of trust and understanding of how AI models work, which have contributed to the limited implementation of AI models in clinical settings compared with the number of models published. However, some fundamental concepts have been defined in different ways, and there is no consensus regarding their meaning, leading to contradictions and discrepancies among definitions proposed in the literature. The definitions in [Table 1](#) represent a proposal to unify concepts identified in review articles published between 2020 and 2025 and included in this study. We analyzed 41 review articles and 634 research articles, providing a substantial body of evidence from which to characterize the current state of trustworthy AI research in healthcare across the topics, methodologies, and data types examined. Together, these findings characterize an intersection of conceptual and methodological aspects that may contribute to advancing the field of trustworthy AI in healthcare.

3.1 Definitions relationship

[Figure 1](#) illustrates the relationships among the concepts defined in [Table 1](#). *Transparency* is proposed as the overarching concept encompassing aspects related to understanding how a model operates and the rationale underlying its decision-making.

Within this framework, *Explainability* encompasses *Interpretability* because, according to our definition, it describes the mechanisms underlying a model’s decisions in a way that enables users with sufficient domain knowledge to understand why and how those decisions are made. *Interpretability* refers to the extent to which the model’s decisions can be easily understood by users, regardless of their level of domain knowledge. While *Interpretability* primarily addresses why a decision is made, it does not necessarily provide insight into how the model arrived at that decision. Therefore, a model may be explainable without being fully interpretable.

Similarly, *Robustness* encompasses aspects related to model performance, as it refers to the ability of a model to maintain consistent and stable performance despite variations in the input data. In contrast, *Reliability* refers to the property of a model to maintain consistent and stable performance under the same conditions or scenarios. Validation techniques used to assess *Reliability* include hold-out methods, such as train–test splitting, and resampling methods, such as K-fold cross-validation. In these approaches, the validation or test data are not used during training but are generally expected to represent conditions consistent with those used to train the model. Therefore, *Reliability* can be considered a specific case within the broader concept of *Robustness*, as it concerns model performance when evaluated under conditions consistent with those encountered during training. In contrast, *Validity* is related to both concepts because it refers to the ability of a model to perform as intended, and therefore depends on its ability to maintain appropriate performance across the conditions under which it is evaluated. Thus, *Validity* may be compromised when either *Robustness* or *Reliability* is insufficient.

Furthermore, *Fairness* encompasses *Bias* because, by definition, it seeks to identify and mitigate algorithmic and data-related biases to minimize discriminatory outcomes. Both concepts also lie at the intersection of *Transparency* and *Robustness*. Addressing *Bias* requires an understanding of the specific characteristics of the populations represented in the data, the properties of the data itself, and the model’s behavior when evaluated across different populations or samples. That understanding requires sufficient *Transparency* to examine the data, model properties and mechanism, and decision-making processes. At the same time, *Fairness* is closely related to *Robustness* because it requires assessing whether model performance remains consistent across populations with different characteristics. Thus, evaluating *Fairness* involves examining how variations in population characteristics may affect model behavior and performance.

In summary, the relationships proposed in [Figure 1](#) indicate that a model cannot be considered valid, interpretable, or capable of making fair decisions without sufficient *Transparency* to provide the information needed to formulate understandable explanations, and identify and address *Bias*. At the same time, these properties depend on the model behaving as expected when exposed to variations in its inputs and producing consistent results under the same conditions. Thus, *Transparency*, *Explainability*, *Interpretability*, *Fairness*, *Bias*, *Robustness*, *Reliability*, and *Validity* should not be considered isolated concepts, but rather interconnected components of trustworthy AI.

3.2 Thematic synthesis and semantic analysis

The fact that most of the definitions were related to *Explainability*, *Interpretability*, and *Transparency* suggests that the field has primarily focused on developing ways to analyze, understand, and explain how AI models work and make decisions. However, the 19 definitions associated with the “*Explainable-Interpretable*” code highlight a lack of clarity regarding the use of concepts that may be closely related. Additionally, the proposed definitions help address the contradictions and discrepancies identified in the literature, specifically in the definitions of *Reliability*, *Bias*, and *Validity*. These findings highlight the importance of reaching a consensus on the definitions of fundamental concepts in trustworthy AI, as the use of consistent terminology can facilitate communication, knowledge sharing, and further scientific progress in the field.

Furthermore, the fact that we did not identify a widely cited source article may suggest that there is not yet a work that has systematically compiled, unified, and proposed definitions that could serve as a basis for broader discussion within the research community and, ultimately, consensus.

3.3 Research articles

In general terms, the extracted data revealed several trends in the field of trustworthy AI in healthcare: (1) the number of articles published increased substantially from 2015 to 2025; (2) Northeast Asia was the most represented geographical region, followed by international collaborations; (3) Neurology, oncology, and cardiology were the clinical domains with the highest number of related studies; (4) EHRs, medical images, and physiological signals were the most commonly used data types; (5) classical ML and DL were the predominant AI approaches; (6) among ML models, decision-tree-based models, particularly gradient boosting and random forests, were the most commonly used, while CNNs were the most common DL models; (7) classification was the most frequently performed task; (8) among XAI techniques, SHAP was the most frequently used, followed by feature importance methods, such as Gini Importance and Permutation Importance, LIME, and Grad-CAM; (9) hold-out and resampling-based validation techniques were the most frequently used; and (10) fewer than half of the analyzed studies implemented techniques to address bias or fairness, with synthetic data generation being the most frequently used approach.

The growth in the number of publications and the diversity of researchers working in this area indicate that trustworthy AI in healthcare is an expanding field that is attracting increasing attention from the scientific community worldwide. Similarly, the distribution of studies across clinical domains is highly variable. Although neurology, oncology, and cardiology account for a relatively large number of publications, these values represent high-frequency categories within the overall distribution rather than evidence of a predominant focus. In fact, most studies are distributed across the remaining 14 clinical domains. This suggests that research in trustworthy AI is not concentrated exclusively on a specific clinical domain, but is being investigated across a broad range of healthcare areas. Similarly, it can be seen that for most clinical domains, the number of publications has increased over time, with 2025 being

the year in which the most articles were published, except for the categories “General / Multisystem”, “Not specified”, “Obstetrics / Gynecology”, “Hematology”, and “Odontology”.

Similar to the trend observed across data types, where the use of all data types increased over time and 2025 had the highest number of articles using each type, more than half of the studies used EHRs. Among these studies, more than 10% used some version of the Medical Information Mart for Intensive Care (MIMIC) [73, 74]. Supplementary Figure S2 shows the top 10 most used datasets by the studies included. The multimodal approaches represented approximately 14%(91) of the included articles (Supplementary Figure S3). Signals were the most frequently used data type in multimodal approaches, commonly combined with EHRs, medical images, questionnaires, or other signals collected from different sources. Multimodal AI refers to the use of data collected from different sources(modality) to train an AI model[75, 76]. The widespread use of EHRs may be partly attributed to their relatively lower storage and data-collection requirements. In contrast, medical images, physiological signals, and genomic data often require greater storage capacity and specialized equipment or protocols for data acquisition, which may increase the technical and logistical complexity of their use. Similarly, the large number of samples available in established datasets such as MIMIC may facilitate access to EHR data and encourage researchers to focus their efforts on these datasets.

3.3.1 Concepts addressed

As noted earlier, the number of publications in this area has increased over time. This is also reflected in the study of specific concepts. In Figure 3, it can be observed that, according to the proposed definitions, most of the articles address *Interpretability*. This finding is supported and complemented by the results presented in Figure 4, which show that SHAP, LIME, Grad-CAM, and feature importance methods were among the most widely implemented XAI techniques. Most of these techniques highlight the relevance of features in the model’s decision-making process, which can help users understand why a particular decision was made—a key aspect of *Interpretability*. However, they do not necessarily provide insight into how the underlying mechanisms of the model operate or contribute to its decision-making process, which is a key aspect of *Explainability*. Consequently, the articles that reported addressing “*Explainability* and *Interpretability*,” “*Explainability*, *Interpretability*, and *Transparency*,” or “*Interpretability* and *Transparency*” may have done so because these concepts were used interchangeably rather than as distinct concepts with clearly differentiated meanings.

Likewise, *Reliability* was the most frequently addressed concept according to our definitions (Figure 3), as almost all articles used validation techniques that assessed model performance using samples from the original dataset, that is, samples acquired under conditions similar to those used for training (Figure 6). In contrast, *Robustness* was assessed by 102 articles through external and temporal validation, as well as direct robustness tests, such as introducing artificial variations into the test data. External validation was used in 86 articles, of which 69 used one external dataset, 14 used two, and 3 used three or more (Supplementary Fig. S4). This finding may suggest that the availability of datasets containing sufficient and reliable information about data

collection protocols and study participants is likely to be limited. This is particularly important when selecting an external dataset because, although the goal is to assess model *Robustness*, a certain degree of consistency in data collection is necessary to ensure that differences in model performance can be meaningfully attributed to variations in the data rather than to differences in the data collection process. Similarly, only 89 studies ($\sim 14\%$) addressed *Validity* using validation techniques related to *Reliability* and *Robustness*. The relatively low number of studies addressing *Robustness* and *Validity* suggests a potential gap in evaluating model performance under conditions beyond those represented in the training data. Together, these findings highlight an opportunity to develop and use datasets acquired under well-documented and varied conditions, enabling more comprehensive assessment of AI models' *Robustness* and *Validity*.

The concepts of *Fairness* and *Bias* were addressed primarily through techniques for modifying the data distribution, including synthetic data generation, such as SMOTE [77], resampling (upsampling, downsampling, or random removal of samples), and reweighting (assigning class weights during training) (Figure 7). More than half of the analyzed articles did not report applying any such method (Figure 3). However, it is not possible to determine whether this was because the data had already been balanced or preprocessed by the authors or dataset providers, or because no approach to addressing *Bias* was undertaken. Among the analyzed studies, Ravindranath E et al. was the only study identified that directly addressed *Fairness* beyond applying a technique aimed at *Bias* mitigation [78]. This finding suggests an opportunity for further research into the development and evaluation of *Fairness* methods for AI models applied in healthcare.

Explainability, on the other hand, was addressed in only 18 articles ($< 3\%$), primarily in studies that implemented rule-based or symbolic techniques. The two techniques most commonly used in these studies were QLattice, which can generate mathematical expressions that provide insight into not only why but also how a model makes decisions [79], and Anchors, which uses conditional rules (IF-THEN) to help explain how a model generates predictions [80]. Furthermore, *Transparency* was not addressed by any article according to the definition proposed in this study, as none provided sufficient information about the model and its operation, the training and validation processes, the *Bias* present in the data and model, and other factors.

Finally, it is worth noting that in 33.6% (213) of the studies analyzed, the concepts identified by the authors coincided with those identified according to the definitions proposed in this study. Thus, while the proposed definitions align with the terminology used in a substantial proportion of studies, the disagreement observed in the remaining articles highlights the extent of the terminological inconsistency within the field.

3.3.2 Implementations

The predominant AI approaches were classical ML and DL. Among ML approaches, decision-tree-based and statistical models, such as linear regression, were the most commonly implemented, while CNNs were the most frequently used DL models. The prevalence of these models may be related to their lower complexity, which facilitates the interpretation and/or explanation of their predictions. For example, decision trees

573 have an inherently interpretable structure that can be visualized [81–84]. Similarly,
 574 CNNs are widely used in medical imaging, where XAI techniques such as Grad-CAM
 575 can be used to visualize the image regions that contribute most strongly to a partic-
 576 ular class prediction. Figure 5 shows an increasing trend in the use of ML models, as
 577 well as hybrid models and LLMs, whereas fuzzy approaches show a declining trend
 578 and the use of DL models appears relatively stable. The number of tasks performed
 579 also increased overall, with classification remaining the predominant task. These find-
 580 ings suggest that, within trustworthy AI research, well-established model families and
 581 tasks with clearly defined prediction targets continue to dominate. The limited use of
 582 alternative model architectures presents opportunities to explore a broader range of
 583 neural network architectures and hybrid approaches in trustworthy AI for healthcare,
 584 particularly for less explored tasks such as segmentation, clustering, and generative
 585 tasks.

586 3.4 Research gaps and recommendations

587 This section summarizes the specific research gaps identified in this review and pro-
 588 poses recommendations for future research. Although Table 1 and Figure 3 propose
 589 unified definitions and relationships among fundamental concepts as a basis for build-
 590 ing consensus in trustworthy AI for healthcare, further work is needed to determine
 591 whether other related concepts, such as *Accountability*, *Applicability*, *Generalizabil-*
 592 *ity*, *Inclusivity*, *Traceability*, *Understandability*, *Usability*, among others, should also
 593 be unified and clearly differentiated. Furthermore, there are currently no widely
 594 established guidelines for quantifying or evaluating *Interpretability*, *Explainability*, or
 595 *Transparency* in AI models. Several clinical domains, data types, model architec-
 596 tures, and AI tasks also remain under-explored. Most XAI methods identified in this
 597 review primarily provide information about *Interpretability*, with limited evidence of
 598 approaches that enable users to understand the underlying mechanisms of AI models.
 599 Similarly, there is a potential gap in the implementation of multimodal approaches and
 600 the evaluation of model performance under conditions beyond those represented in the
 601 training data, including evaluation across populations with different characteristics.

602 Based on these gaps, we propose the following recommendations:

- 603 • Adopt the definitions presented in Table 1 as a basis for analysis, discussion, and
 604 consensus-building within the trustworthy AI community.
- 605 • Identify additional concepts related to trustworthy AI in healthcare that may require
 606 standardization, further discussion, and future consensus.
- 607 • Develop standardized methods for evaluating *Interpretability*, *Explainability*, *Trans-*
 608 *parency*, and *Fairness*.
- 609 • Encourage research in under-studied clinical domains, such as dentistry, dermatol-
 610 ogy, and hematology, among others; promote the implementation of multimodal
 611 approaches and exploration of hybrid models (DL + ML), LLMs, and fuzzy models
 612 for less explored tasks, including segmentation, clustering, and generative tasks.
- 613 • Develop open-access databases and datasets with comprehensive information on
 614 data-collection protocols, population characteristics, and sample characteristics to

615 facilitate the use of diverse data types, more comprehensive *Robustness* validation,
616 and approaches addressing *Fairness* and *Bias*.

617 3.5 Limitations

618 The limitations of this study include the time required to analyze the articles and
619 extracted data, as well as to write, revise, and publish the manuscript. Given the rapid
620 growth of the field, some of the trends and distributions presented in the figures may
621 have changed by the time this article is consulted. (MORE TO BE ADDED)

622 4 Conclusion

623 This study analyzed a total of 675 articles published between 2015 and 2025 on trust-
624 worthy AI in healthcare. The findings show that the trustworthy AI concepts examined
625 in this study have been defined by multiple authors, without a consensus or highly
626 cited reference emerging from the literature. Furthermore, several definitions contained
627 contradictions and inconsistencies, while many studies used *Interpretability*, *Explain-*
628 *ability*, and *Transparency* interchangeably despite treating them as distinct concepts.
629 To address these inconsistencies, Table 1 proposes a set of unified definitions for
630 the concepts examined in this review. These definitions resulted from a methodology
631 combining qualitative and quantitative techniques to harmonize definitions previously
632 proposed, specifically within the healthcare domain. However, their applicability is
633 not necessarily limited to healthcare, as these concepts form part of the conceptual
634 foundations of trustworthy AI.

635 This is an evolving field, with a steady increase in the number of publications since
636 2015. According to the definitions proposed in this study, *Reliability* and *Interpretabil-*
637 *ity* were the concepts most frequently addressed. *Transparency* was not addressed
638 by any article, whereas *Explainability* was addressed by fewer than 10% of the arti-
639 cles analyzed. Moreover, several research gaps remain, including limited evaluation of
640 *Robustness*, *Validity*, and *Fairness*, as well as the limited exploration of certain clini-
641 cal domains, data types, model architectures, and AI tasks. To address these gaps, we
642 propose recommendations for future research, as described in Section 3.3.

643 5 Methodology

644 This scoping review followed the Preferred Reporting Items for Systematic Reviews
645 and Meta-Analyses (PRISMA) guidelines [85–87]. The included studies were classified
646 into two categories: review articles and research articles. To identify the trustworthy
647 AI concepts to be analyzed, unified, and defined, we adopted as a reference the liter-
648 ature review by Tucci et al., which analyzed 57 studies and identified 42 factors that
649 enhance end-user trust in medical AI systems [88]. Among these, the most frequently
650 investigated factors were:

- 651 • *Explainability*
- 652 • *Interpretability*
- 653 • *Transparency*

- 654 • *Reliability*
- 655 • *Fairness*
- 656 • *Robustness*
- 657 • *Validity*

658 Based on the importance of these factors in shaping end-user trust, we selected
 659 them as the primary concepts examined in this review. In addition, we included *Bias*
 660 because of its close conceptual relationship with *Fairness*.

661 5.1 Search strategy

662 A structured literature search was conducted in IEEE Xplore, PubMed/MEDLINE,
 663 Embase, CINAHL, PsycINFO, Scopus, and Web of Science between September and
 664 November 2025. This time frame refers to when the searches were performed. The
 665 search was limited to articles published from January 2015 onward. The starting year
 666 was selected to examine the evolution of research in trustworthy AI before, during,
 667 and after major milestones in the field, including the launch of the DARPA XAI
 668 program, the introduction of LIME and SHAP, and the publication of key ethical and
 669 regulatory initiatives by the IEEE and the EU (find more about the historical context
 670 in the section 1.2). Search filters included articles published in English from January
 671 1, 2015, onward.

672 The final search string was developed based on three core topics: Trustworthy
 673 AI, Healthcare and Real-world implementation. Keywords related to each one were
 674 combined using Boolean operators. The complete search string is provided below:

675 ("trustworthy ai" OR "trustworthy artificial intelligence" OR "responsible ai" OR "eth-
 676 ical ai" OR "reliable ai" OR "reliable artificial intelligence" OR "explainable ai" OR
 677 "interpretable ai" OR "explainable artificial intelligence" OR "interpretable artificial intel-
 678 ligence" OR "transparent ai" OR "robust ai" OR "safe ai" OR "secure ai" OR "fair ai"
 679 OR "accountable ai" OR "human-centered ai" OR "responsible machine learning" OR
 680 "explainable machine learning" OR "interpretable machine learning" OR "robust machine
 681 learning" OR "reliable machine learning") AND (healthcare OR "medical services" OR
 682 "medical diagnostic imaging" OR diseases OR illness OR "biological system modeling" OR
 683 "biomedical imaging" OR "medical diagnosis" OR "biomedical monitoring" OR "medical
 684 imaging" OR hospitals OR sensors OR "public healthcare") AND ("real-time systems"
 685 OR monitoring OR "technological innovation" OR industries OR "solid modeling" OR
 686 "market research" OR software) NOT ("IoT" OR "Internet of things" OR blockchain)

687 5.2 Study selection

688 For each, specific inclusion and exclusion criteria were established to ensure consistency
 689 in the selection process. The following inclusion and exclusion criteria were used to
 690 select relevant review articles:

- 691 I1. Provide specific definitions of the following concepts of interest presented in the first
 692 part of [section 5](#):
- 693 E1. Do not provide any definition of interest.
- 694 E2. It is not related to healthcare.
- 695 E3. Published before 2020.

696 The following inclusion and exclusion criteria were used to select research articles:

- 697 I1. Includes any technique related to interpretability or explainability (according to our
698 definitions).
- 699 I2. Implements external validation using one or more datasets or tested in real-world
700 scenarios.
- 701 I3. Proposes or uses methods related to bias/unbias.
- 702 I4. Original research articles from journals.
- 703 E1. Does not implement anything related to the concepts of interest. Just a designing,
704 framework, or general idea.
- 705 E2. It is not related to healthcare.

706 5.3 Screening

707 Study selection was conducted using Covidence systematic review software (Veritas
708 Health Innovation, Melbourne, Australia; www.covidence.org). Titles and abstracts
709 (first stage), and full-text (second stage) were screened independently by two reviewers
710 according to the predefined inclusion and exclusion criteria. Disagreements between
711 the two reviewers were resolved through discussion in both stages. Fig. 8 shows the
712 PRISMA[85–87] diagram where all the data of each step are illustrated.

713 5.4 Review articles data extraction and definitions proposal

714 5.4.1 Thematic synthesis

715 The thematic synthesis was conducted using ATLAS.ti qualitative data analysis
716 software (Scientific Software Development GmbH, www.atlasti.com). All definitions
717 presented in each of the analyzed review articles were identified using this software.
718 Some articles presented more than one definition; therefore, all definitions were initially
719 considered, regardless of whether or not they were proposed by the authors. Subse-
720 quently, different codes or labels were assigned to each definition based on its content
721 following the thematic synthesis methodology proposed by Thomas and Harden [89].
722 The goal was to compare the definitions related to each of the concepts of interest.

723 5.4.2 Semantic analysis

724 Once the definitions were grouped, semantic embeddings were generated for each one
725 using the all-mpnet-base-v2 model from the Sentence-Transformers framework [90].
726 Next, the cosine distance between the embeddings was calculated (this allows us to
727 determine how similar the vectors are to one another). Distance values range from 0
728 (similar) to 2 (opposite) (1 indicates that they are different). Using these distances, the
729 UPGMA algorithm [91] was applied to create a hierarchical cluster based on similarity.
730 In this case, a threshold of 0.5 was used to generate the clusters, since values below 1
731 indicate similarity. Clusters could only be formed if the distance between definitions
732 was 0.5 or less. Definitions that did not meet this criterion were not grouped together.

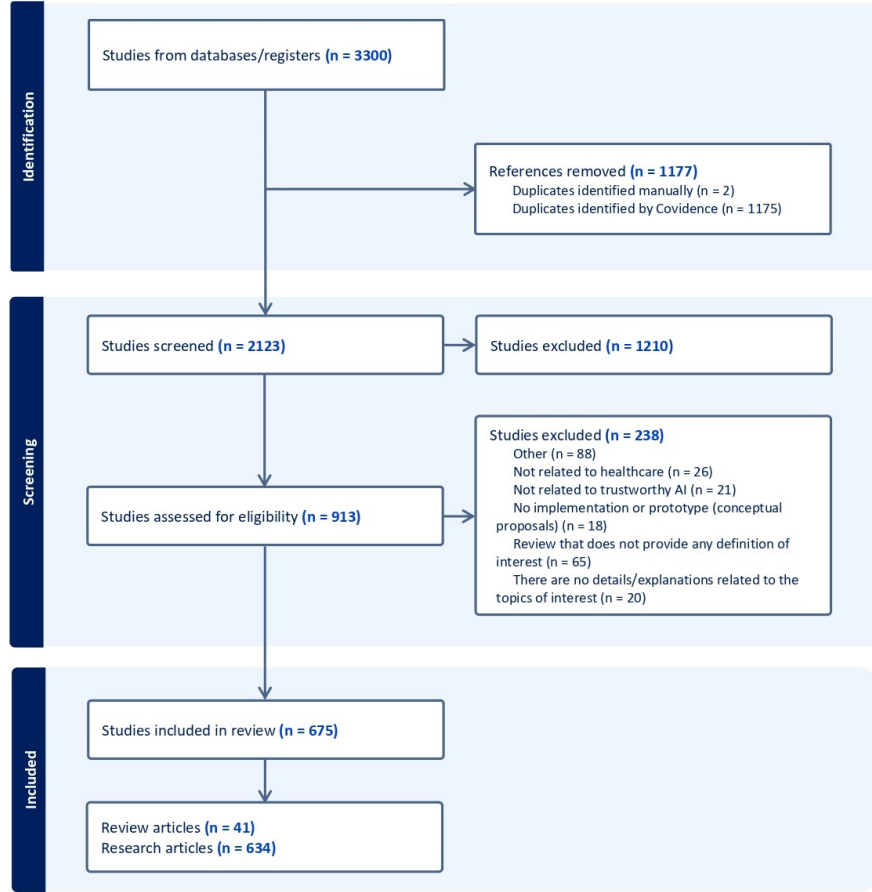


Fig. 8 Study selection and exclusion diagram according to the PRISMA guideline.

For each cluster, a similarity matrix was calculated using cosine similarity as the metric to identify repeated or highly redundant definitions. This is because cosine similarity allows us to quantify how similar two vectors are; in this case, the values range from -1 (opposite) to 1 (similar) (0 means they are different). This metric allows us to calculate the angle between the two vectors using the arccosine function, whereby angles closer to 0° indicate greater similarity. For this reason, to identify the aforementioned patterns, we decided to set a threshold of 30° (corresponding to a similarity of 0.86). All definitions that had an angle between them equal to or less than this value were reviewed independently—both at the definition level and at the level of the references used to construct them (if any)—in an effort to reduce the noise that repeated or redundant definitions might generate. Additionally, all definitions were reviewed individually to determine whether there were source articles (cited by multiple definitions) or whether, conversely, they proposed their own definition. An N-gram analysis was then performed, which consisted of determining the frequency of unigrams (one

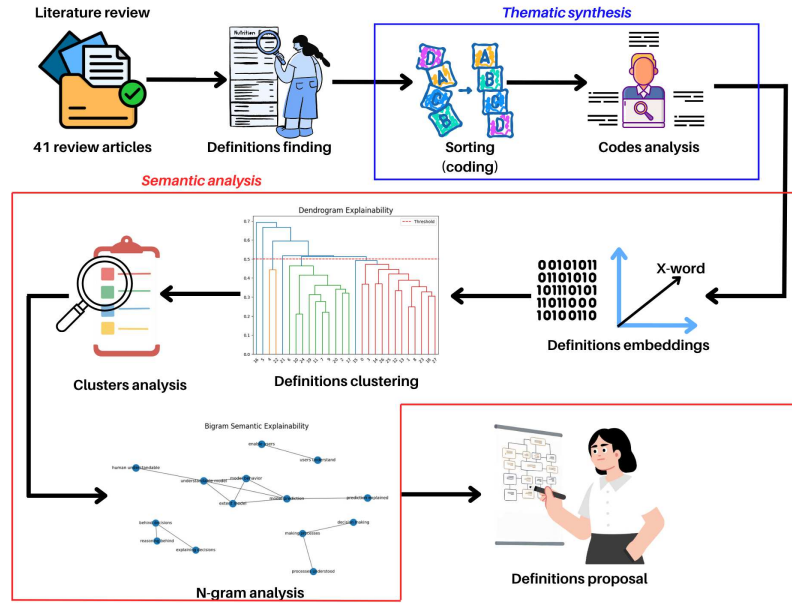


Fig. 9 Study selection and exclusion diagram according to the PRISMA guideline.

word) and bigrams (two words) in all definitions of each concept. Furthermore, the most frequent N-grams were used to construct graph representations of the semantic relationships between the bigrams, using the unigrams they have in common as connections. Finally, the definitions proposed in Table 1 were constructed using this semantic relationship.

5.5 Research articles data extraction

The extracted data from the selected articles included: Year of publication, Geographical region, Clinical domain addressed, Concepts addressed according to the authors, Task(s) performed, AI method, Model(s) implemented, Data type(s), It was a multimodal AI approach?, Technique(s) used for interpretability/explainability, Did the study use external databases used for validation?, Validation technique(s), there was any Fairness / Bias addressing?, Method(s) used for Fairness / Bias addressing, Concepts addressed according to our definitions, and Did the concepts addressed by the authors match with ours?

Competing interests. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Author contribution. This work was completed in the Biomedic.AI Lab at McMaster University. All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. The following is a breakdown of the individual contributions of each author. Conceptualization, M.S.R.C. and T.E.D.; studies screening, M.S.R.C. and R.E.; formal analysis, M.S.R.C.; writing—original draft preparation, M.S.R.C. and T.E.D.; writing—review and editing, M.S.R.C., R.E, T.E.D.; supervision, T.E.D. All authors have read and agreed to the published version of the manuscript.

Funding. This study was funded by National Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2022-05438.

Acknowledgements. The research team would like to acknowledge all individuals who took part in this study, thank you for making this possible.

References

- [1] Du-Harpur, X., Watt, F., Luscombe, N., Lynch, M.: What is ai? applications of artificial intelligence to dermatology. *British Journal of Dermatology* **183**(3), 423–430 (2020)
- [2] Janiesch, C., Zschech, P., Heinrich, K.: Machine learning and deep learning: C. janiesch et al. *Electronic markets* **31**(3), 685–695 (2021)
- [3] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., *et al.*: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education* **23**(1), 689 (2023)
- [4] Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal* **8**(2), 188–194 (2021)
- [5] Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: Ai in health and medicine. *Nature medicine* **28**(1), 31–38 (2022)
- [6] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.-Z.: Deep learning for health informatics. *IEEE journal of biomedical and health informatics* **21**(1), 4–21 (2016)
- [7] Yu, K.-H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. *Nature biomedical engineering* **2**(10), 719–731 (2018)
- [8] Schouten, B., Schinkel, M., Boerman, A.W., Pijkeren, P., Thodé, M., Beneden, M., Panday, R.N., Jonge, R., Wiersinga, W.J., Nanayakkara, P.W.: Implementing

- artificial intelligence in clinical practice: a mixed-method study of barriers and facilitators. *Journal of Medical Artificial Intelligence* **5** (2022)
- [9] Zhang, J., Whebell, S., Gallifant, J., Budhdeo, S., Mattie, H., Lertvittayakumjorn, P., Lopez, M.d.P.A., Tiangco, B.J., Gichoya, J.W., Ashrafi, H., *et al.*: An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. *The Lancet Digital Health* **4**(4), 212–213 (2022)
- [10] Fleuren, L.M., Thorat, P., Shillan, D., Ercole, A., Elbers, P.W., RJ, R.D.R.N.C.H.M.G.B.K.T.L.G.T.R.L.F.S.E.L.G.A.: Machine learning in intensive care medicine: ready for take-off? *Intensive care medicine* **46**(7), 1486–1488 (2020)
- [11] Berardi, C., Antonini, M., Jordan, Z., Wechtler, H., Paolucci, F., Hinwood, M.: Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. *BMC health services research* **24**(1), 243 (2024)
- [12] Ganapathy, A., Clough, B.A., Casey, L.M.: Organizational and policy barriers to the use of digital mental health by mental health professionals. *Telemedicine and e-Health* **27**(12), 1332–1343 (2021)
- [13] Kickbusch, I., Piselli, D., Agrawal, A., Balicer, R., Banner, O., Adelhardt, M., Capobianco, E., Fabian, C., Gill, A.S., Lupton, D., *et al.*: The lancet and financial times commission on governing health futures 2030: growing up in a digital world. *The Lancet* **398**(10312), 1727–1776 (2021)
- [14] Figueroa, C.A., Luo, T., Aguilera, A., Lyles, C.R.: The need for feminist intersectionality in digital health. *The Lancet Digital Health* **3**(8), 526–533 (2021)
- [15] Samarasekera, U.: The rise of racial minority health apps. *The Lancet Digital Health* **4**(4), 218–219 (2022)
- [16] Schueller, S.M., Hunter, J.F., Figueroa, C., Aguilera, A.: Use of digital mental health for marginalized and underserved populations. *Current Treatment Options in Psychiatry* **6**(3), 243–255 (2019)
- [17] Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)* **55**(2), 1–38 (2022)
- [18] Ai, H.: High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI* **6**(15.3), 2024 (2019)
- [19] Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence: S. thiebes *et al.* *Electronic Markets* **31**(2), 447–464 (2021)

- [20] Loyola-Gonzalez, O.: Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access* **7**, 154096–154113 (2019)
- [21] Bharati, S., Mondal, M.R.H., Podder, P.: A review on explainable artificial intelligence for healthcare: Why, how, and when? *IEEE Transactions on Artificial Intelligence* **5**(4), 1429–1442 (2023)
- [22] Joyce, D.W., Kormilitzin, A., Smith, K.A., Cipriani, A.: Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj digital medicine* **6**(1), 6 (2023)
- [23] McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5**(4), 115–133 (1943)
- [24] Turing, A.M.: Computing machinery and intelligence. In: *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, pp. 23–65. Springer, ??? (2007)
- [25] McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine* **27**(4), 12–12 (2006)
- [26] Shortliffe, E.H.: *Computer-Based Medical Consultations: Mycin Book*. Elsevier, ??? (1976)
- [27] Van Melle, W.: Mycin: a knowledge-based consultation program for infectious disease diagnosis. *International journal of man-machine studies* **10**(3), 313–322 (1978)
- [28] Computing, S.: *New generation computing technology: a strategic plan for its development and application to critical problems in defense*. Defense Advanced Research Projects Agency (1983)
- [29] Swartout, W., Paris, C., Moore, J.: Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert* **6**(3), 58–64 (1991)
- [30] Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 900–907 (2004). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999
- [31] Gunning, D., Aha, D.W.: Darpa’s explainable artificial intelligence program. *AI magazine* **40**(2), 44–58 (2019)
- [32] Gunning, D., Vorm, E., Wang, Y., Turek, M.: Darpa’s explainable ai (xai) program: A retrospective. *Authorea Preprints* (2021)

- [33] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
- [34] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017)
- [35] Association, I.S.: The ieee global initiative for ethical considerations in artificial intelligence and autonomous systems. See https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v1.pdf (2016)
- [36] CIFAR: Pan-Canadian AI Strategy - CIFAR — cifar.ca. <https://cifar.ca/ai/>. [Accessed 22-07-2026] (2017)
- [37] Union, E.: General Data Protection Regulation (GDPR) – Legal Text — gdpr-info.eu. <https://gdpr-info.eu/>. [Accessed 22-07-2026] (2018)
- [38] De Oliveira, A.C., Azevedo, J.P.C., Ruback, L., Moreira, R., Teixeira, S.S., Teles, A.S.: Effect of explainable artificial intelligence on trust of mental health professionals in an ai-based system for suicide prevention. *IEEE Access* (2025)
- [39] Kerz, E., Zanwar, S., Qiao, Y., Wiechmann, D.: Toward explainable ai (xai) for mental health detection based on language behavior. *Frontiers in psychiatry* **14**, 1219479 (2023)
- [40] Pendyala, V., Kim, H.: Assessing the reliability of machine learning models applied to the mental health domain using explainable ai. *Electronics* **13**(6), 1025 (2024)
- [41] Srinivasu, P.N., Sandhya, N., Jhaveri, R.H., Raut, R.: From blackbox to explainable ai in healthcare: existing tools and case studies. *Mobile Information Systems* **2022**(1), 8167821 (2022)
- [42] Retzlaff, C.O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H., Holzinger, A.: Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research* **86**, 101243 (2024)
- [43] Molnar, C.: *Interpretable Machine Learning*, 3rd edn. (2025). <https://christophm.github.io/interpretable-ml-book>
- [44] Shapley, L.S., et al.: *A value for n-person games*. Princeton University Press (1953)
- [45] Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S.: Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science* **17**(11), 70056

- (2024)
- [46] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
- [47] Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450 (2016)
- [48] Scornet, E.: Trees, forests, and impurity-based variable importance in regression. In: Annales de l’Institut Henri Poincaré (B) Probabilités et Statistiques, vol. 59, pp. 21–52 (2023). Institut Henri Poincaré
- [49] Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A.: A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics **10**(1), 213 (2009)
- [50] Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
- [51] Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. Journal of machine learning research **20**(177), 1–81 (2019)
- [52] Muhammad, D., Bendeche, M.: Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL **24**, 542–560 (2024) <https://doi.org/10.1016/j.csbj.2024.08.005>
- [53] FV, F., K, F., B, L., W, K., PK, D.: Explainable ai: A review of applications to neuroimaging data. Front Neurosci **16**, 906290 (2022) <https://doi.org/10.3389/fnins.2022.906290>
- [54] Band, S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A.T., Liang, H.-W.: Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked **40** (2023) <https://doi.org/10.1016/j.imu.2023.101286>
- [55] Ramos-Soto, O., Aranguren, I., Carrillo M, M., Oliva, D., Balderas-Mata, S.E.: Artificial intelligence in medical imaging diagnosis: are we ready for its clinical implementation? Journal of medical imaging (Bellingham, Wash.) **12**(6), 061405–061405 (2025) <https://doi.org/10.1117/1.JMI.12.6.061405>
- [56] Gorelik, A.J., Li, M., Hahne, J., Wang, J., Ren, Y., Yang, L., Zhang, X., Liu, X., Wang, X., Bogdan, R., Carpenter, B.D.: Ethics of ai in healthcare: a scoping

- review demonstrating applicability of a foundational framework. FRONTIERS IN DIGITAL HEALTH **7** (2025) <https://doi.org/10.3389/fdgth.2025.1662642>
- [57] Radanliev, P.: Privacy, ethics, transparency, and accountability in ai systems for wearable devices. FRONTIERS IN DIGITAL HEALTH **7** (2025) <https://doi.org/10.3389/fdgth.2025.1431246>
- [58] Kondylakis, H., Osuala, R., Puig-Bosch, X., Lazrak, N., Diaz, O., Kushibar, K., Chouvarda, I., Charalambous, S., Starmans, M.P., Colantonio, S., Tachos, N., Joshi, S., Woodruff, H.C., Salahuddin, Z., Tsakou, G., Ausso, S., Alberich, L.C., Papanikolaou, N., Lambin, P., Marias, K., Tsiknakis, M., Fotiadis, D.I., Marti-Bonmati, L., Lekadir, K.: A review of methods for trustworthy ai in medical imaging: The future-ai guidelines. IEEE journal of biomedical and health informatics **PP** (2025) <https://doi.org/10.1109/JBHI.2025.3614546> . Michael Stiven Ramirez Campos (2026-02-27 07:45:39)(Select): They defined interpretability in sec 7, fairness and bias in sec 2, and robustness in 6. ;
- [59] Souza, A.A., Stubbs, A.P., Hesselink, D.A., Baan, C.C., Boer, K.: Cherry on top or real need? a review of explainable machine learning in kidney transplantation. TRANSPLANTATION **109**(1), 123–132 (2025) <https://doi.org/10.1097/TP.0000000000005063>
- [60] Al-Ghuribi, S.M., Ahmed, A.A., Ibraheem, A.S., Hasan, M.K., Islam, S.: Navigating the ethical landscape of artificial intelligence: A comprehensive review. International Journal of Computing and Digital Systems **18**(1) (2025) <https://doi.org/10.12785/ijcds/1571038843>
- [61] Siala, H., Wang, Y.: Shifting artificial intelligence to be responsible in health-care: A systematic review. Social Science & Medicine **296**(Abramoff, M.D., Leng, T., Ting, D.S.W., Rhee, K., Horton, M.B., Brady, C.J., & Chiang, M.F. (2020). Automated and computer-assisted detection, classification, and diagnosis of diabetic retinopathy. Telemed. e-Health, 26(4), 544-550. Agarwal, R., Dugas, M.), 1–15 (2022) <https://doi.org/10.1016/j.socscimed.2022.114782>
- [62] Umer, L., Khan, M.H., Ayaz, Y.: Transforming healthcare with artificial intelligence in pakistan: A comprehensive overview. Pakistan Armed Forces Medical Journal **73**(4), 955–963 (2023) <https://doi.org/10.51253/pafmj.v73i4.10852>
- [63] S, N., DM, D., MU, A.: Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput Biol Med **156**, 106668 (2023) <https://doi.org/10.1016/j.compbimed.2023.106668>
- [64] Karim, M.R., Islam, T., Shajalal, M., Beyan, O., Lange, C., Cochez, M., Rebholz-Schuhmann, D., Decker, S.: Explainable ai for bioinformatics: Methods, tools and applications. BRIEFINGS IN BIOINFORMATICS **24**(5) (2023) <https://doi.org/10.1093/bib/bbad236>

- [65] Raposo, V.L.: How is ‘unexplainable’ and non-transparent ai affecting healthcare delivery? *Oslo Law Review* **11**(1) (2024) <https://doi.org/10.18261/olr.11.1.6>
- [66] Consortium, F.-A., Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., Weicken, E., Asselbergs, F.W., Prior, F., Collins, G.S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J.A., Kushibar, K., Riklund, K., Marias, K., Amugongo, L.M., Fromont, L.A., Maier-Hein, L., Cerda-Alberich, L., Marti-Bonmati, L., Cardoso, M.J., Bobowicz, M., Shabani, M., Tsiknakis, M., Zuluaga, M.A., Fritzsche, M.-C., Camacho, M., Linguraru, M.G., Wenzel, M., De Bruijne, M., Tolsgaard, M.G., Goisauf, M., Abadia, M.C., Papanikolaou, N., Lazrak, N., Pujol, O., Osuala, R., Napel, S., Colantoni, S., Joshi, S., Klein, S., Ausso, S., Rogers, W.A., Salahuddin, Z., Starmans, M.P.A.: Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ-BRITISH MEDICAL JOURNAL* **388** (2025) <https://doi.org/10.1136/bmj-2024-081554>
- [67] Khokhar, P.B., Gravino, C., Palomba, F.: Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. *Artificial intelligence in medicine* **164**, 103132–103132 (2025) <https://doi.org/10.1016/j.artmed.2025.103132>
- [68] Roy, S., Pal, D., Meena, T.: Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. *Network Modeling Analysis in Health Informatics and Bioinformatics* **13**(1) (2024) <https://doi.org/10.1007/s13721-023-00437-y>
- [69] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* **58**, 82–115 (2020)
- [70] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **267**, 1–38 (2019)
- [71] Lipton, Z.C.: The mythos of model interpretability. *Communications of the ACM* **61**(10), 36–43 (2018)
- [72] Instruments, O.L.: Recommendation of the council on artificial intelligence. Organization for Economic Cooperation and Development (2019)
- [73] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. *Scientific data* **3**(1), 160035 (2016)
- [74] Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., *et al.*: Mimic-iv, a freely accessible electronic health record dataset. *Scientific data* **10**(1), 1 (2023)

- 1019 [75] Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical
1020 ai. *Nature medicine* **28**(9), 1773–1784 (2022)
- 1021 [76] Ramirez Campos, M.S., Barati, K., Samavi, R., Pires, P., Duncan, L., Sassi, R.B.,
1022 Noseworthy, M.D., Gadsden, S.A., Doyle, T.E.: Multimodal artificial intelligence
1023 and online learning in youth mental health: a scoping review. *npj Mental Health*
1024 *Research* **5**(1), 26 (2026)
- 1025 [77] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic
1026 minority over-sampling technique. *Journal of artificial intelligence research* **16**,
1027 321–357 (2002)
- 1028 [78] Ravindranath, R., Stein, J.D., Hernandez-Boussard, T., Fisher, A.C., Wang, S.Y.,
1029 Amin, S., Edwards, P.A., Srikumaran, D., Woreta, F., Schultz, J.S., *et al.*: The
1030 impact of race, ethnicity, and sex on fairness in artificial intelligence for glaucoma
1031 prediction models. *Ophthalmology Science* **5**(1), 100596 (2025)
- 1032 [79] Broløs, K.R., Machado, M.V., Cave, C., Kasak, J., Stentoft-Hansen, V., Batanero,
1033 V.G., Jelen, T., Wilstrup, C.: An approach to symbolic regression using feyn.
1034 arXiv preprint arXiv:2104.05417 (2021)
- 1035 [80] Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic
1036 explanations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*,
1037 vol. 32 (2018)
- 1038 [81] Grazioli, S., Crippa, A., Rosi, E., Candelieri, A., Ceccarelli, S.B., Mauri, M.,
1039 Manzoni, M., Mauri, V., Trabatttoni, S., Molteni, M., Colombo, P., Nobile, M.:
1040 Exploring teliagnostic procedures in child neuropsychiatry: addressing adhd
1041 diagnosis and autism symptoms through supervised machine learning. *European*
1042 *Child & Adolescent Psychiatry* **33**(1), 139–149 (2024) [https://doi.org/10.1007/](https://doi.org/10.1007/s00787-023-02145-4)
1043 [s00787-023-02145-4](https://doi.org/10.1007/s00787-023-02145-4)
- 1044 [82] Gebran, A., Vapsi, A., Maurer, L.R., El Moheb, M., Naar, L., Thakur, S.S.,
1045 Sinyard, R., Daye, D., Velmahos, G.C., Bertsimas, D., Kaafarani, H.M.A.: Potter-
1046 icu: An artificial intelligence smartphone-accessible tool to predict the need for
1047 intensive care after emergency surgery. *SURGERY* **172**(1), 470–475 (2022) <https://doi.org/10.1016/j.surg.2022.03.023>
1048
- 1049 [83] X, G., S, A., P, S., F, D., N, K.: Interpretable machine learning models for hospital
1050 readmission prediction: a two-step extracted regression tree approach. *BMC Med*
1051 *Inform Decis Mak* **23**(1), 104 (2023) <https://doi.org/10.1186/s12911-023-02193-5>
- 1052 [84] De-La-Garza-Salazar, F., Egenriether, B.: Clinically interpretable electrovector-
1053 cardiographic machine learning criteria for the detection of echocardiographic left
1054 ventricular hypertrophy. *PLOS ONE* **20**(10 October) (2025) [https://doi.org/10.](https://doi.org/10.1371/journal.pone.0334829)
1055 [1371/journal.pone.0334829](https://doi.org/10.1371/journal.pone.0334829)

- 1056 [85] Liberati, A., et al.: The prisma statement for reporting systematic reviews and
1057 meta-analyses of studies that evaluate healthcare interventions: explanation and
1058 elaboration. *Bmj* **339** (2009)
- 1059 [86] Tricco, A.C., *et al.*: Prisma extension for scoping reviews (prisma-scr): checklist
1060 and explanation. *Annals of internal medicine* **169**(7), 467–473 (2018)
- 1061 [87] Page, M.J., et al.: The prisma 2020 statement: an updated guideline for reporting
1062 systematic reviews. *Bmj* **372** (2021)
- 1063 [88] Tucci, V., Saary, J., Doyle, T.E.: Factors influencing trust in medical artificial
1064 intelligence for healthcare professionals: a narrative review. *Journal of Medical
1065 Artificial Intelligence* **5** (2022)
- 1066 [89] Thomas, J., Harden, A.: Methods for the thematic synthesis of qualitative
1067 research in systematic reviews. *BMC medical research methodology* **8**(1), 45
1068 (2008)
- 1069 [90] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese
1070 bert-networks. In: *Proceedings of the 2019 Conference on Empirical Methods
1071 in Natural Language Processing*. Association for Computational Linguistics, ???
1072 (2019). <https://arxiv.org/abs/1908.10084>
- 1073 [91] Sokal, R.R., Michener, C.D., et al.: A statistical method for evaluating systematic
1074 relationships. *University of Kansas Science Bulletin* (1958)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [MultimodalapproachS3.png](#)
- [ExternaldatabasesusedforvalidationS4.png](#)
- [DatasetsS2.png](#)
- [CodesdistributionS1.png](#)
- [Researcharticlessummarysupplementary.csv](#)
- [Thematicanalysisquotations.xlsx](#)