International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	PATIENT TRUST IN AI-ASSISTED DIAGNOSIS AND TRIAGE: A STRUCTURED NARRATIVE REVIEW OF ETHICAL, SOCIAL, AND IMPLEMENTATION CONDITIONS
DOI	https://doi.org/10.31435/ijitss.3(51).2026.6008
RECEIVED	19 May 2026
ACCEPTED	04 September 2026
PUBLISHED	08 September 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

PATIENT TRUST IN AI-ASSISTED DIAGNOSIS AND TRIAGE: A STRUCTURED NARRATIVE REVIEW OF ETHICAL, SOCIAL, AND IMPLEMENTATION CONDITIONS

OliwerMuller *(Corresponding Author, Email: oliwermuller@gmail.com)*
Kielce Hospital of St. Aleksandra Sp. z o.o., Kielce, Poland
ORCID ID: 0009-0005-6197-8461

Jagoda Pałubska
Independent Public Health Care Institution of the Ministry of the Interior and Administration in Kielce named after St. John Paul II, Kielce, Poland
ORCID ID: 0009-0000-3833-7977

Zuzanna Rafałowska
University Clinical Hospital in Wrocław Borowska 213, 50-556 Wrocław, Poland
ORCID ID: 0009-0003-6699-3121

Stanisław Rutz
Wroclaw Medical University, wybrzeże Ludwika Pasteura 1, 50-367 Wrocław, Poland
ORCID ID: 009-0007-7918-9625

Anna Wojtala
Szpital Uniwersytecki nr 2 im. dr. Jana Biziela, ul. Ujejskiego 75, 85-168 Bydgoszcz, Poland
ORCID ID: 0009-0008-4082-0835

Julia Kret
Szpital Specjalistyczny im. F. Ceynowy, ul. dr. A. Jagalskiego 10, 84-200 Wejherowo, Poland
ORCID ID: 0009-0003-2780-8448

Aleksandra Wyrostkiewicz
Plac Hirszfelda 12, 53-413 Wrocław, Poland
ORCID ID: 0009-0005-3679-7570

Aleksander Sadkowski
Collegium Medicum w Bydgoszcz Uniwersytet Mikołaja Kopernika w Toruniu, Jagielońska 15, 85-067 Bydgoszcz, Poland
ORCID ID: 0009-0008-3300-2447

Paulina Bigda
Wojewódzki Szpital Zespolony im. Ludwika Rydygiera ul. Św. Józefa 53-59, 87-100 Toruń, Poland
ORCID ID: 0009-0005-5022-6184

Jan Gdula
Collegium Medicum w Bydgoszczy Uniwersytetu Mikołaja Kopernika w Toruniu, Jagiellońska 15, 85-067 Bydgoszcz, Poland
ORCID ID: 0009-0007-9987-3758

Lilianna Zielińska
Plac Hirszfelda 12, 53-413 Wrocław, Poland
ORCID ID: 0009-0004-2471-841X

Michał Szpunar
Collegium Medicum w Bydgoszcz Uniwersytet Mikołaja Kopernika w Toruniu, Poland
ORCID ID: 0009-0000-5981-3696

Adrianna Zabiegałowska
Wojewódzki Szpital Zespolony im. Ludwika Rydygiera ul. Św. Józefa 53-59, 87-100 Toruń, Poland
ORCID ID: 0009-0005-3048-0097

Magda Buśko
Collegium Medicum w Bydgoszczy Uniwersytetu Mikołaja Kopernika w Toruniu, Poland
ORCID ID: 0009-0005-6451-8612

ABSTRACT

Artificial intelligence is increasingly used in medical imaging, clinical decision support, and digital triage. However, technical performance alone does not determine whether such tools will be accepted in practice. Their implementation also depends on whether patients perceive them as understandable, safe, fair, and compatible with person-centred care. This review examines the main factors that shape patient trust in AI-assisted diagnosis and triage and discusses the social and ethical conditions that support responsible adoption. A structured narrative review design was used to synthesize recent literature on patient attitudes toward clinical AI, with particular attention to radiology, diagnostic support, communication, explainability, accountability, and emerging triage applications. The literature shows a consistent pattern of conditional acceptance. Patients are often willing to accept AI when it is used as a support tool under clinician supervision, but they are much more hesitant when AI is framed as replacing doctors in final decision-making. Trust is influenced not only by explainability, but also by validation, data quality, fairness, privacy, communication, and visible human accountability. Patients repeatedly express the wish to be informed when AI is involved in their care and to retain access to a clinician who can interpret results, answer questions, and assume responsibility. The review argues that trust in medical AI should be understood as calibrated confidence rather than simple approval or rejection. From a social-science perspective, AI in diagnosis and triage is a sociotechnical arrangement whose legitimacy depends on governance, workflow design, and the continued protection of the doctor-patient relationship.

KEYWORDS

Artificial Intelligence, Patient Trust, Diagnosis, Triage, Explainability, Healthcare Ethics

CITATION

Muller, O., Pałubska, J., Rafałowska, Z., Rutz, S., Wojtala, A., Kret, J., Wyrostkiewicz, A., Sadkowski, A., Bigda, P., Gdula, J., Zielińska, L., Szpunar, M., Zabiegałowska, A., & Buśko, M. (2026). Patient trust in AI-assisted diagnosis and triage: A structured narrative review of ethical, social, and implementation conditions. *International Journal of Innovative Technologies in Social Science*, 3(51). https://doi.org/10.31435/ijitss.3(51).2026.6008

Introduction

Artificial intelligence (AI) has become one of the most discussed innovations in contemporary healthcare. It is now routinely associated with medical imaging, clinical decision support, risk prediction, automated documentation, and digital pathways that help sort patients before they are seen by a clinician. In public and professional debate, these tools are often described in terms of accuracy, speed, efficiency, and scalability. That language is understandable, but it captures only part of the issue. In healthcare, a diagnosis is not merely a technical classification. It is also a human event shaped by uncertainty, vulnerability, communication, and responsibility. Because of that, the introduction of AI into diagnosis and triage cannot be assessed only through technical performance. It must also be assessed through the question of trust.

Trust matters because even highly accurate systems can fail socially if patients do not see them as legitimate, understandable, or accountable. In medicine, trust has never depended on technical competence alone. Patients also judge whether they are being respected, whether information is communicated honestly, and whether someone remains answerable for decisions that affect their lives. This makes patient trust in AI an especially important topic for a journal concerned with innovative technologies in social science. Medical AI is not only a computational development. It is also a change in social practice, institutional authority, and the relationship between expertise and care.

Recent literature suggests that patients are not simply either enthusiastic or resistant toward AI. Their attitudes are more nuanced. Across surveys, interviews, and reviews, patients often express openness to AI when it is described as assisting clinicians, improving efficiency, or reducing diagnostic error. However, they become more cautious when AI is framed as replacing human judgment or operating in ways that are not explained clearly. Several studies in radiology have found that patients are more comfortable with a collaborative model in which AI supports physicians but does not act independently in final decision-making (Ongena et al., 2020; Ibba et al., 2023; Xuereb & Portelli, 2024). Similar patterns appear in broader research

on public and patient attitudes, where people tend to accept the use of AI more readily when human oversight remains visible and responsibility is clearly retained by a clinician (Khullar et al., 2022; Young et al., 2021).

This preference for supervision rather than substitution is not trivial. It points to the deeper meaning of diagnosis in clinical practice. Patients do not see diagnosis as a purely data-driven output. They expect someone to interpret the finding in context, explain what it means, discuss uncertainty, and take responsibility for consequences. AI may be capable of identifying patterns, but for many patients this does not make it a sufficient substitute for a clinician who can integrate technical findings with the patient's history, values, emotional state, and questions. The issue is therefore not only whether AI can improve diagnosis, but whether it can be introduced without weakening the relational and ethical foundations of care.

Explainability has become a central concept in this discussion. At first glance, the solution seems straightforward: if AI systems are explainable, patients will trust them more. Yet the literature shows that the matter is more complicated. Patients do not necessarily require a deep technical explanation of algorithmic architecture. What they often want is something simpler and more practical: to know whether AI is being used, what role it plays, how much weight is given to its output, what its limitations are, and who is responsible if it fails. In this sense, practical intelligibility may matter more than technical transparency alone (Amann et al., 2020; Kostick-Quenet et al., 2024). At the same time, scholars have warned that explainability should not be treated as a substitute for rigorous validation, representative data, and external testing. An explanation that sounds persuasive cannot compensate for poor performance or hidden bias (Ghassemi et al., 2021; Fehr et al., 2022).

A second recurring theme is disclosure. Patients frequently want to be informed when AI is involved in diagnosis or decision support, especially when the technology shapes an interpretation that may affect treatment or access to care. This desire is closely tied to autonomy and dignity. Patients do not want AI to become an invisible actor in medicine. They may not insist on formal consent for every form of automation, but they often expect openness, especially in higher-stakes situations. Studies in imaging and broader healthcare contexts show that patients value being informed, being able to ask questions, and understanding how AI enters the clinical process (Ongena et al., 2020; Adus et al., 2023; Xuereb & Portelli, 2024).

The question becomes even more important in triage. Unlike many back-office systems, triage tools often interact directly or indirectly with patients at a moment of uncertainty. They influence urgency classification, referral pathways, waiting times, and sometimes the speed with which a person gains access to further care. Trust in this setting can be fragile because triage is experienced under pressure. Although the patient-centered literature on AI-based triage is less developed than the literature on radiology and diagnostic AI, newer work already suggests that trust in triage depends on accurate information, alignment with clinical expertise, meaningful supervision, and clear communication about how the system should be used (Steerling et al., 2025). This makes triage an especially important case for thinking about the social legitimacy of healthcare AI.

The present review examines patient trust in AI-assisted diagnosis and triage from an ethical and social-science perspective. Its aim is not to determine whether AI is good or bad in general. Instead, it asks a more useful question: under what conditions do patients regard clinical AI as worthy of trust? The article brings together literature on patient perceptions, explainability, accountability, person-centred care, and implementation in order to show that trust is shaped by a combination of technical, relational, and institutional factors. In doing so, it also illustrates why the successful integration of medical AI is not just a matter of software design, but also of governance, communication, and professional responsibility.

Methodology

This paper uses a structured narrative review design. The objective was not to conduct a formal meta-analysis or a fully exhaustive systematic review, but to build a coherent synthesis of key peer-reviewed literature relevant to patient trust in AI-assisted diagnosis and triage. A narrative review was chosen because the available literature is methodologically heterogeneous. It includes qualitative interview studies, questionnaire development studies, cross-sectional surveys, focus groups, conceptual ethics papers, and systematic reviews. Given this diversity, thematic synthesis was more appropriate than quantitative pooling.

The review strategy prioritized literature that addressed one or more of the following domains: patient attitudes toward AI in healthcare, patient trust in diagnostic AI, perceptions of AI in radiology or medical imaging, explainability and transparency in patient-facing AI, accountability and human oversight in clinical decision-making, and emerging evidence on trust in AI-based triage. The core body of literature was concentrated in recent years, especially between 2020 and 2025, reflecting the rapid development of medical AI and the expansion of patient-attitude research during this period. Earlier or more theoretical sources were included selectively when they remained conceptually important for patient-centred decision-making or relational care.

The main search orientation for a real review of this type would include databases such as PubMed and Scopus, using combinations of terms including *artificial intelligence*, *machine learning*, *diagnosis*, *triage*, *patient trust*, *patient attitudes*, *radiology*, *explainability*, *transparency*, *accountability*, and *doctor-patient relationship*. For the purposes of this model article, the literature base was constructed around recurring and well-cited studies in the field, especially those that directly investigated patient perceptions rather than only technical model performance.

Inclusion was based on thematic relevance. Papers were retained when they provided empirical or conceptual insight into how patients understand, accept, question, or resist AI in clinical settings. Particular attention was given to diagnostic contexts because these are the areas in which patient-centered research is currently most developed. Studies on radiology were therefore especially prominent. This is not because radiology is the only field that matters, but because it has become a central site for patient discussions about AI and therefore offers useful evidence for broader questions of trust. Triage-related literature was included where available, especially when it addressed actual trust in use rather than purely speculative attitudes.

Purely technical benchmarking papers were not a central focus unless they were cited in relation to validation, transparency, or implementation. Likewise, papers that discussed AI in very broad or purely futuristic terms without engaging patient perspectives were treated with caution. The intention was to keep the analysis close to patient-facing issues: information, oversight, communication, acceptance, fairness, and accountability.

The selected literature was read comparatively and grouped by repeated themes. During this process, several issues appeared consistently across study types and settings. These included patient preference for AI as support rather than replacement, the importance of understandable disclosure, the central role of human accountability, the continuing significance of the doctor-patient relationship, and concerns about fairness, privacy, and inclusion. These themes became the analytical basis of the review.

This approach has clear limitations. First, much of the evidence comes from hypothetical scenarios, questionnaire items, or early-stage implementation rather than mature long-term use. Second, diagnostic imaging is overrepresented relative to triage. Third, patient attitudes may vary across countries, institutions, medical conditions, and levels of digital familiarity, and the literature has not yet mapped that variation fully. For these reasons, the conclusions of this article should be read as a synthesis of trust conditions rather than a universal statement about all patients in all settings. Even so, the convergence across multiple studies is strong enough to support a meaningful interpretation of how trust in medical AI is currently being formed.

Results

The literature suggests that patient trust in AI-assisted diagnosis and triage is neither automatic nor absent. Instead, it is conditional. Patients are often willing to accept AI in healthcare, but their acceptance is shaped by how the technology is described, who remains in control, what information is provided, and whether the system appears aligned with clinical and moral expectations of care. Across the reviewed literature, six broad themes recur.

Table 1. Major factors shaping patient trust in AI-assisted diagnosis and triage

Factor	What the literature suggests	Practical implication
AI as support, not replacement	Patients are more comfortable with AI that assists clinicians than with AI acting alone.	Present AI as augmentation and preserve clinician review.
Understandable disclosure	Patients want to know when AI is used and what role it plays.	Use plain-language explanation rather than technical jargon.
Human accountability	Trust weakens when no clear person remains responsible.	Keep final responsibility and escalation pathways explicit.
Validation and safety	Trust depends on evidence, not only on claims of innovation.	Communicate performance, limits, and uncertainty honestly.
Relationship and communication	Patients value dialogue, empathy, and the ability to ask questions.	Ensure AI does not reduce meaningful human contact.
Fairness, privacy, and inclusion	Concerns about bias, data use, and exclusion affect legitimacy.	Audit data practices and maintain equitable access pathways.

AI is more acceptable as support than as replacement

The strongest and most consistent finding across the patient literature is that people generally prefer AI to function as a support tool rather than an autonomous decision-maker. This pattern appears in radiology, general healthcare surveys, and studies of public attitudes toward AI in medicine. Patients tend to welcome AI when it is presented as improving efficiency, reducing error, or helping doctors interpret findings. However, they become notably more cautious when AI is framed as fully replacing physicians or making independent final decisions (Ongena et al., 2020; Yakar et al., 2022; Khullar et al., 2022; Young et al., 2021).

This preference reflects how patients understand medical care. A diagnosis is not viewed as an isolated computational output. Patients expect the clinician to interpret test results in light of history, symptoms, uncertainty, and individual circumstances. They also expect explanation, reassurance, and the opportunity to ask follow-up questions. Even where patients see benefits in automation, they often do not view those benefits as a sufficient reason to remove the clinician from the final interpretive role.

Radiology studies show this especially clearly. Ongena et al. (2020) found that patients' views clustered around issues such as distrust and accountability, procedural knowledge, personal interaction, and being informed. In later studies, similar preferences appeared again. Ibba et al. (2023) reported that most respondents supported the use of AI to support diagnosis but did not feel comfortable with diagnosis made by AI alone. Xuereb and Portelli (2024) likewise found that patients were broadly open to AI in medical imaging but still preferred that doctors retain final decision-making authority. Baghdadi et al. (2024) also showed that patients favored personal interaction and physician involvement over full replacement in radiology settings.

The same logic appears in broader, non-radiology studies. Khullar et al. (2022) described a nationally representative survey in which patients expressed interest in AI's use in healthcare but remained cautious about its role in core decisions. Robertson et al. (2023) found that people were not categorically opposed to AI diagnostics, but uptake improved when the technology was presented as more accurate, more attentive, and better integrated into a physician-guided pathway. These results do not suggest technophobia. Rather, they show that patients often accept AI under conditions of hybrid intelligence, where machine capability is paired with human judgment.

Patients value understandable explanation and disclosure

A second major theme concerns explanation. Patients often want to understand how AI enters the care process, but that does not necessarily mean they want a highly technical account of model architecture. In practice, many patients seek practical intelligibility rather than technical detail. They want to know whether AI is being used, what kind of task it performs, how much the doctor relies on it, what the system can and cannot do, and what happens if there is disagreement or uncertainty.

This distinction matters. Explainability in technical AI discourse often refers to model interpretability, saliency maps, or other computational methods designed to make outputs more transparent. Yet the patient literature suggests that trust is rarely built by technical explanation alone. What matters more is whether patients feel informed in a way that supports autonomy, comprehension, and participation. Amann et al. (2020) emphasize that explainability in healthcare has technological, ethical, legal, and patient-facing dimensions. Their central point is that explainability cannot be treated as a narrow engineering problem because it is tied to informed consent, liability, and patient understanding.

Empirical patient studies support this broader view. Many respondents in radiology studies indicate that they want disclosure when AI is involved in diagnosis or treatment planning (Ongena et al., 2020; Ibba et al., 2023; Xuereb & Portelli, 2024). This preference appears to reflect respect as much as information needs. Patients may not demand to know how an algorithm was trained, but they often want acknowledgement that it was part of the diagnostic pathway. That disclosure signals that the patient is being treated as an active participant rather than a passive data source.

Recent work on patient engagement and implementation strengthens this point. Adus et al. (2023) found that patients wanted meaningful involvement in AI development and emphasized that education is important for making such engagement real rather than symbolic. In a 2025 focus group study, Foresman et al. reported that patients valued transparency, communication, privacy, and human oversight across different AI scenarios related to diagnosis and communication. Together, these studies suggest that explanation should not be reduced to model output explanation. It should be understood as a broader communicative practice that helps patients understand where AI sits within care.

Trust depends on validation, data quality, and evidence of safety

Although explanation matters, the literature repeatedly shows that explanation alone does not guarantee trust. Patients may not always speak in technical research language, but many of their concerns point toward a deeper issue: is the system actually reliable, and has it been evaluated responsibly? This introduces a crucial distinction between *appearing* transparent and *being* trustworthy.

Ghassemi et al. (2021) argue that current approaches to explainable AI often create false hope in high-stakes settings such as medicine. Their argument is that explanation methods are not sufficient substitutes for rigorous internal and external validation when tools are used for patient-level decisions. Fehr et al. (2022) similarly show that transparency and trustworthiness are related but not identical. A tool may provide partial information about its training or intended use and still fall short of what is needed to judge its trustworthiness.

From a patient perspective, this matters because trust is not only built through communication. It is also built through institutional assurance that the tool has been tested properly, works across relevant populations, and is being monitored in real-world settings. Kostick-Quenet et al. (2024) make this point especially clearly in their work on trust criteria for AI in health. They found that patients and physicians often ground trust in epistemic factors such as accuracy, validity, relevance of data, and the integrity of the training set. This is important because it suggests that users do not simply care about whether the system can explain itself; they also care about whether the knowledge behind it is sound.

In practical terms, this means that responsible implementation should communicate not only the existence of an AI tool, but also its scope, limits, and degree of evidence. A patient may be more willing to accept AI support if they know that it has been validated in settings similar to their own and that clinicians understand where it performs well and where caution is required. By contrast, vague institutional confidence or marketing-style claims about innovation may fail to generate trust, especially in sensitive clinical contexts.

Human accountability remains central

The literature also shows that patients strongly associate trustworthy AI with clear human responsibility. When a clinician remains visibly answerable for the final interpretation or decision, AI is more acceptable. When responsibility becomes diffuse or ambiguous, trust weakens.

This pattern is visible across empirical and conceptual studies. Patients often say that they prefer doctors to make the final call even if AI contributes to the analysis (Ongena et al., 2020; Xuereb & Portelli, 2024). In part, this reflects beliefs about expertise, but it also reflects beliefs about moral accountability. Patients want to know that if something goes wrong, someone remains responsible for reviewing the case, explaining the outcome, and responding to harm.

Sung (2023) argues that responsibility and accountability in AI-assisted medicine must be clarified before harm occurs, not afterward. The point is not merely legal. It is social and ethical. Patients are unlikely to trust a system when responsibility is spread across developers, hospitals, vendors, and clinicians in a way that makes direct accountability hard to locate. Sauerbrei et al. (2023) reach a similar conclusion from the perspective of person-centred care, emphasizing that the use of AI should remain assistive if the relationship of trust between doctor and patient is to be preserved.

This issue is particularly important in triage. Triage is not just advisory; it shapes access, urgency, and the sequence of care. For that reason, human oversight in AI-based triage cannot be symbolic. It needs to be operational. There must be clear escalation procedures, clear rules for overruling system outputs, and clear mechanisms for what happens when a patient's description does not fit the system's assumptions.

The emerging triage literature supports this. In a qualitative study from Sweden, Steerling et al. (2025) found that both patients and healthcare professionals linked trust in AI-based triage to accurate patient information, alignment with clinical expertise, and supervision of patient health and safety. Participants also emphasized constructive dialogue and clear instructions about how information should be used and stored. This is significant because it shows that, in triage settings, trust depends on the system's fit with professional judgment and patient safety, not merely on efficiency.

The doctor-patient relationship remains a key site of trust

Another recurring finding is that AI can either support or weaken the doctor-patient relationship depending on how it is implemented. Patients do not necessarily assume that AI will damage care, but they do worry that it might reduce empathy, explanation, or personal interaction.

Sauerbrei et al. (2023) argue that person-centred care depends on values such as empathy, compassion, trust, and shared decision-making. They suggest that AI can have a positive impact if it is used in an assistive role and if healthcare education adapts to support good human-AI interaction. Their point is important because

it moves beyond a simplistic assumption that technology always either saves or destroys the relationship. Instead, the impact depends on use.

Patients themselves often express this relational concern indirectly. In imaging and survey studies, many prefer the presence of a physician not only because of technical trust but because physicians can explain results, answer questions, and provide reassurance (Ongena et al., 2020; Ibba et al., 2023). Robertson et al. (2023) also show that acceptance improves when AI pathways are framed as more attentive to the patient's unique situation. This suggests that trust is linked not just to the accuracy of the tool, but also to whether the care pathway still feels human.

Bjerring and Busch (2021) discuss this issue at the level of patient-centred decision-making. They argue that even if AI becomes extremely accurate, this does not automatically justify replacing the participatory and value-sensitive elements of medical decision-making. Decisions in medicine are not reducible to technical optimization. Patients bring preferences, values, fears, and interpretations that matter to what counts as a good decision. The more a system obscures or sidelines that process, the more difficult trust may become.

For implementation, the implication is clear: healthcare systems should evaluate AI not only in terms of clinical performance and workflow efficiency, but also in terms of how it affects the relational texture of care. A system that saves time but reduces explanation, limits access to clinicians, or makes patients feel unheard may still fail in a patient-centred sense.

Fairness, privacy, and inclusion shape legitimacy

Trust also extends beyond the immediate diagnostic encounter to the wider data and governance environment in which AI operates. Patients care about how their data are collected, stored, and used. They also care whether AI works fairly across different social groups.

This concern appears in both empirical and conceptual literature. Fehr et al. (2022) note that ethical and legal considerations form part of transparency and trustworthiness assessments. Sung (2023) likewise argues that inclusiveness, autonomy, and dignity must be preserved if AI is to gain trust and acceptance. In practice, patients may accept AI in principle but still reject a specific implementation if they believe data protection is weak, bias is possible, or the technology may work better for some groups than others.

Studies of patient attitudes also show that acceptance is not evenly distributed. Fritsch et al. (2022) found that older respondents, women, and people with lower educational levels or lower technical affinity were more cautious about AI in healthcare. Robertson et al. (2023) also found meaningful differences across demographic and attitudinal lines. This does not mean that skepticism is irrational. It suggests that trust is socially situated. People assess AI in light of prior experiences, cultural beliefs, technological familiarity, and broader trust in institutions.

These findings have direct implications for equity. A health system that introduces AI without considering language accessibility, digital literacy, data representativeness, and the availability of human alternatives risks reproducing or intensifying existing inequalities. Trustworthy AI, therefore, is not simply accurate AI. It is also inclusive AI.

Implications for triage specifically

Although patient-centered literature on triage is still smaller than the literature on radiology, the available evidence suggests several important principles. First, patients need clarity about what the triage system actually does. Is it offering advice, sorting urgency, or making a binding decision about access to care? Trust is harder to sustain when the system's authority is unclear.

Second, triage systems need routes to human review. Because triage often occurs when symptoms are uncertain or distressing, patients may be especially sensitive to the fear of being dismissed by an automated pathway. Third, communication around triage tools needs to include uncertainty. A system that presents a recommendation too confidently may appear efficient but can undermine trust if patients feel their case is more complex than the tool allows for.

The Swedish qualitative study by Steerling et al. (2025) is particularly valuable here because it moves beyond hypothetical acceptability into actual trust influences among both professionals and patients. The study suggests that trust in AI-based triage is strengthened when the system supports accurate information, aligns with clinical expertise, and maintains supervision over patient health and safety. These are not minor conditions. They imply that trustworthy triage is inseparable from real human oversight and from a design that helps, rather than obstructs, patient-clinician interaction.

Discussion

The literature reviewed in this article supports a clear conclusion: patient trust in AI-assisted diagnosis and triage is best understood as **calibrated confidence**. Patients are not simply deciding whether they are "for" or "against" AI. Instead, they are judging whether a particular use of AI is appropriate, understandable, supervised, and worthy of reliance in a healthcare setting. This is a more demanding standard than simple acceptance, and it is also a more useful one.

Seeing trust as calibrated confidence helps explain why support for AI often coexists with strong insistence on human oversight. Patients may believe that AI can improve care while still believing that the clinician should remain responsible for final decisions. This is not a contradiction. It reflects a realistic view of medicine as both technical and relational. AI may improve pattern recognition, prioritization, or consistency, but patients still need someone who can interpret results, contextualize uncertainty, and remain answerable for what happens next.

This understanding also helps clarify why explainability has become such an important yet sometimes overused concept. In policy and public discourse, explainability is often presented as the key to trust. The literature reviewed here suggests a more cautious interpretation. Explainability matters, but what matters most for patients is not necessarily access to the internal logic of the model. It is whether the use of AI is intelligible in practice. Patients want to know that AI is present, what role it plays, what its limitations are, and how clinicians are using it. They also want assurance that the tool has been validated responsibly and that someone remains accountable when problems arise.

This distinction is important because it prevents the trust debate from becoming too narrow. A polished explanation interface cannot fix weak validation, biased data, or poor workflow design. Likewise, a technically strong model may still struggle socially if it is introduced without patient disclosure or without preserving access to a clinician. Trust therefore needs to be approached as a property of a broader sociotechnical system, not just of the algorithm itself.

From the perspective of implementation, several implications follow.

First, AI should be introduced as a support tool unless there is exceptional evidence and a robust ethical case for greater autonomy. The literature repeatedly shows that patients are far more comfortable with augmentation than replacement. Framing matters here, but so does actual practice. Systems should be designed so that clinician oversight is visible and operational rather than nominal.

Second, communication should be plain, proportionate, and honest. In higher-stakes contexts, patients should be informed when AI contributes to diagnosis or triage, especially when it may affect interpretation, risk categorization, or access. This does not require lengthy technical consent language in every case. It does require a clear, respectful explanation of the system's role.

Third, institutions should communicate not only benefits but also limitations. Trust is more durable when uncertainty is acknowledged rather than hidden. Patients are not likely to lose confidence merely because a clinician says that an AI tool is helpful but not infallible. They are more likely to lose confidence if the system appears overhyped or if its role is obscured.

Fourth, implementation should include governance measures that patients may never fully see but whose effects shape trust nonetheless. These include external validation, subgroup analysis, auditing for bias, clear escalation routes, secure data governance, and documentation of when human review overrides AI output. These institutional measures are part of what makes trust deserved.

Fifth, patient engagement should be taken more seriously at earlier stages of development. Adus et al. (2023) show that patients themselves want more meaningful involvement in AI design and implementation. This matters because trust is not produced simply by informing patients after a system is built. It is also produced by incorporating their expectations and concerns into the development process.

Triage deserves especially careful attention in future research and policy. Compared with radiology, the patient-perspective literature on triage is still limited. Yet triage tools are increasingly important because they affect where patients go, how fast they are seen, and how urgency is interpreted. These systems operate at the front edge of care, often at moments of stress. For that reason, the ethical margin for poor communication or unclear accountability may be even smaller than in other AI applications.

Several limitations of the present review should also be acknowledged. Much of the evidence remains concentrated in radiology and imaging. Many studies still rely on hypothetical vignettes or surveys rather than long-term observation of mature AI systems in practice. Cross-national differences, cultural expectations, and institutional contexts are not yet fully understood. Also, trust is not static. It may shift as patients gain more experience with AI in everyday healthcare.

Future research should therefore move in at least three directions. First, it should study actual use rather than only hypothetical acceptance. Second, it should include more work on triage, primary care, and digitally mediated access pathways. Third, it should pay closer attention to how trust differs across social groups, especially where digital literacy, structural disadvantage, or prior mistrust of institutions may affect how AI is experienced.

Overall, the evidence suggests that the success of AI in diagnosis and triage will not be determined by performance metrics alone. It will also depend on whether health systems can make these tools understandable, accountable, and aligned with the values of care. That is a technical task, but it is equally a social one.

Conclusions

Patient trust in AI-assisted diagnosis and triage does not arise automatically from innovation. It has to be earned through design, communication, governance, and clinical practice. The literature reviewed in this article shows that patients are generally not opposed to AI in healthcare. What they want is an AI that improves care without displacing responsibility, weakening communication, or eroding dignity.

Across multiple studies, patients appear more willing to accept AI when it supports clinicians rather than replaces them, when they are informed about its use, and when responsibility for final decisions remains clearly human. Trust is also shaped by validation, data quality, privacy, fairness, and the broader effect of AI on the doctor-patient relationship. These factors are especially important in triage, where AI may influence access and urgency at emotionally charged points of care.

For healthcare institutions, the practical lesson is that the key question is not whether AI can be inserted into diagnosis and triage, but how it should be introduced so that it deserves trust. For social-science research, the broader lesson is that trust in medical AI is not a side issue. It is part of the core legitimacy of technological change in healthcare.

Acknowledgments: No external funding was received for the preparation of this educational model manuscript.

Ethical Considerations: This article is a narrative review of published literature and does not involve original research with human participants, animals, or identifiable personal data. Ethics committee approval was therefore not required.

Conflicts of Interest: No conflicts of interest to declare.

REFERENCES

1. Adus, S., Macklin, J., & Pinto, A. (2023). Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. *BMC Health Services Research, 23*, 1163. https://doi.org/10.1186/s12913-023-10098-2
2. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*, 310. https://doi.org/10.1186/s12911-020-01332-6
3. Baghdadi, L. R., Mobeirek, A. A., Alhudaithi, D. R., Albenmousa, F. A., Alhadlaq, L. S., Alaql, M. S., & Alhamlan, S. A. (2024). Patients' attitudes toward the use of artificial intelligence as a diagnostic tool in radiology in Saudi Arabia: Cross-sectional study. *JMIR Human Factors, 11*, e53108. https://doi.org/10.2196/53108
4. Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology, 34*(2), 349-371. https://doi.org/10.1007/s13347-019-00391-6
5. Fehr, J., Jaramillo-Gutierrez, G., Oala, L., Gröschel, M. I., Bierwirth, M., Balachandran, P., Werneck-Leite, A., & Lippert, C. (2022). Piloting a survey-based assessment of transparency and trustworthiness with three medical AI tools. *Healthcare, 10*(10), 1923. https://doi.org/10.3390/healthcare10101923
6. Fritsch, S. J., Blankenheim, A., Wahl, A., Hetfeld, P., Maassen, O., Deffge, S., Kunze, J., Rossaint, R., Riedel, M., Marx, G., & Bickenbach, J. (2022). Attitudes and perception of artificial intelligence in healthcare: A cross-sectional survey among patients. *Digital Health, 8*, 20552076221116772. https://doi.org/10.1177/20552076221116772
7. Foresman, G., Biro, J., Tran, A., MacRae, K., Kazi, S., Schubel, L., Visconti, A., Gallagher, W., Smith, K. M., Giardina, T., Haskell, H., & Miller, K. (2025). Patient perspectives on artificial intelligence in health care: Focus group study for diagnostic communication and tool implementation. *Journal of Participatory Medicine, 17*, e69564. https://doi.org/10.2196/69564
8. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

9. Ibba, S., Tancredi, C., Fantesini, A., Cellina, M., Presta, R., Montanari, R., Papa, S., & Alì, M. (2023). How do patients perceive the AI-radiologists interaction? Results of a survey on 2119 responders. *European Journal of Radiology, 165*, 110917. https://doi.org/10.1016/j.ejrad.2023.110917
10. Khullar, D., Casalino, L. P., Qian, Y., Lu, Y., Krumholz, H. M., & Aneja, S. (2022). Perspectives of patients about artificial intelligence in health care. *JAMA Network Open, 5*(5), e2210309. https://doi.org/10.1001/jamanetworkopen.2022.10309
11. Kostick-Quenet, K., Lang, B. H., Smith, J., Hurley, M., & Blumenthal-Barby, J. (2024). Trust criteria for artificial intelligence in health: Normative and epistemic considerations. *Journal of Medical Ethics, 50*(8), 544-551. https://doi.org/10.1136/jme-2023-109338
12. Ongena, Y. P., Haan, M., Yakar, D., & Kwee, T. C. (2020). Patients' views on the implementation of artificial intelligence in radiology: Development and validation of a standardized questionnaire. *European Radiology, 30*(2), 1033–1040. https://doi.org/10.1007/s00330-019-06486-0
13. Robertson, C., Woods, A., Bergstrand, K., Findley, J., Balser, C., & Slepian, M. J. (2023). Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. *PLOS Digital Health, 2*(5), e0000237. https://doi.org/10.1371/journal.pdig.0000237
14. Sauerbrei, A., Kerasidou, A., Lucivero, F., & Hallowell, N. (2023). The impact of artificial intelligence on the person-centred, doctor-patient relationship: Some problems and solutions. *BMC Medical Informatics and Decision Making, 23*, 73. https://doi.org/10.1186/s12911-023-02162-y
15. Steerling, E., Svedberg, P., Nilsen, P., Siira, E., & Nygren, J. (2025). Influences on trust in the use of AI-based triage-an interview study with primary healthcare professionals and patients in Sweden. *Frontiers in Digital Health, 7*, 1565080. https://doi.org/10.3389/fdgth.2025.1565080
16. Sung, J. (2023). Artificial intelligence in medicine: Ethical, social and legal perspectives. *Annals of the Academy of Medicine, Singapore, 52*(12), 695–699. https://doi.org/10.47102/annals-acadmedsg.2023103
17. Xuereb, F., & Portelli, J. L. (2024). The knowledge and perception of patients in Malta towards artificial intelligence in medical imaging. *Journal of Medical Imaging and Radiation Sciences, 55*(4), 101743. https://doi.org/10.1016/j.jmir.2024.101743
18. Yakar, D., Ongena, Y. P., Kwee, T. C., & Haan, M. (2022). Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. *Value in Health, 25*(3), 374-381. https://doi.org/10.1016/j.jval.2021.09.004
19. Young, A. T., Amara, D., Bhattacharya, A., & Wei, M. L. (2021). Patient and general public attitudes towards clinical artificial intelligence: A mixed methods systematic review. *The Lancet Digital Health, 3*(9), e599–e611. https://doi.org/10.1016/S2589-7500(21)00132-1