American
Journal of
Advanced Technology and
Engineering Solutions

**Volume:** 01; **Issue:** 01
**Pages:** 01-39
**eISSN: 3067-5146**

# A Trust-Centered AI and Security Modeling Approach for Early Cancer Diagnosis, Population-Level Health Analysis, and Secure Deployment in U.S. Healthcare Infrastructure

**Md Fokhrul Alam[1]; Md Fardaus Alam[2]; Md Ashraful Alam[3];**

[1]. *Department of Computer Science, Bachelor of Science in Computer Science & Engineering, Southeast University, Dhaka, Bangladesh*
*Email: ashrafulinfo1234@gmail.com*

[2]. *Department of Science & Technology, Diploma in Computer Science and Application, Bangladesh Oper University, Gazipur, Bangladesh*
*Email: fardausdesh@gmail.com*

[3]. *Department of Computer Science, Bachelor of Science in Computer Science & Engineering, Southeast University, Dhaka, Bangladesh*
*Email: alaminfo121@gmail.com*

**Doi: 10.63125/zyvt7f56**

***Received:*** *09 September 2020;* ***Revised:*** *10 October 2020;* ***Accepted:*** *12 November 2020;* ***Published:*** *03 December 2020*

## Abstract

*This study addresses the persistent gap between high-performing cancer AI prototypes and real-world adoption by proposing and testing a trust-centered AI plus security-modeling blueprint for early cancer diagnosis and population-level health analysis within U.S. healthcare infrastructure. The purpose was to quantify how often published "enterprise-ready" capabilities actually co-occur with deploy ability pillars such as interpretability, robustness, security, and equity. Using a quantitative, cross-sectional, case-based review design, each included paper was treated as a case reflecting cloud and enterprise healthcare deployment contexts (for example, multi-site systems, integrated EHR and imaging stacks, or networked inference services). The sample comprised 45 cases (N = 45). Key variables were five Likert-scored readiness dimensions (1–5): clinical validation rigor, interpretability and communication support, robustness and generalization evidence, security and privacy modeling, and fairness and equity evidence, plus composite indicators such as trust-mechanism presence and a rubric-scaled Trust-Centered Deployment Readiness (TDR). The analysis plan applied descriptive statistics (counts, percentages, means), cross-tabs between trust-mechanism grouping and validation readiness, and a composite readiness summary. Headline findings show that radiology and pathology cases dominated (31/45, 68.9%), interpretability appeared in 28/45 (62.2%) but comprehensive interpretability was limited (12/45, 26.7%; mean M = 3.1/5), external validation or multi-site evaluation occurred in only 16/45 (35.6%), and explicit security or privacy-by-design elements were present in 14/45 (31.1%) with the lowest readiness mean (M = 2.4/5). Trust-mechanism studies (19/45, 42.2%) showed higher validation readiness (M = 3.6/5) and more external validation (11/19, 57.9%) than performance-only studies (5/26, 19.2%). Overall, only 8/45 (17.8%) met high composite readiness (≥0.75), while 21/45 (46.7%) were moderate and 16/45 (35.6%) low. Implications indicate that healthcare AI procurement and governance should prioritize a complete evidence package that couples external validation, calibrated trust cues, security controls across the lifecycle, and subgroup equity reporting, rather than selecting models based on accuracy alone.*

## Keywords

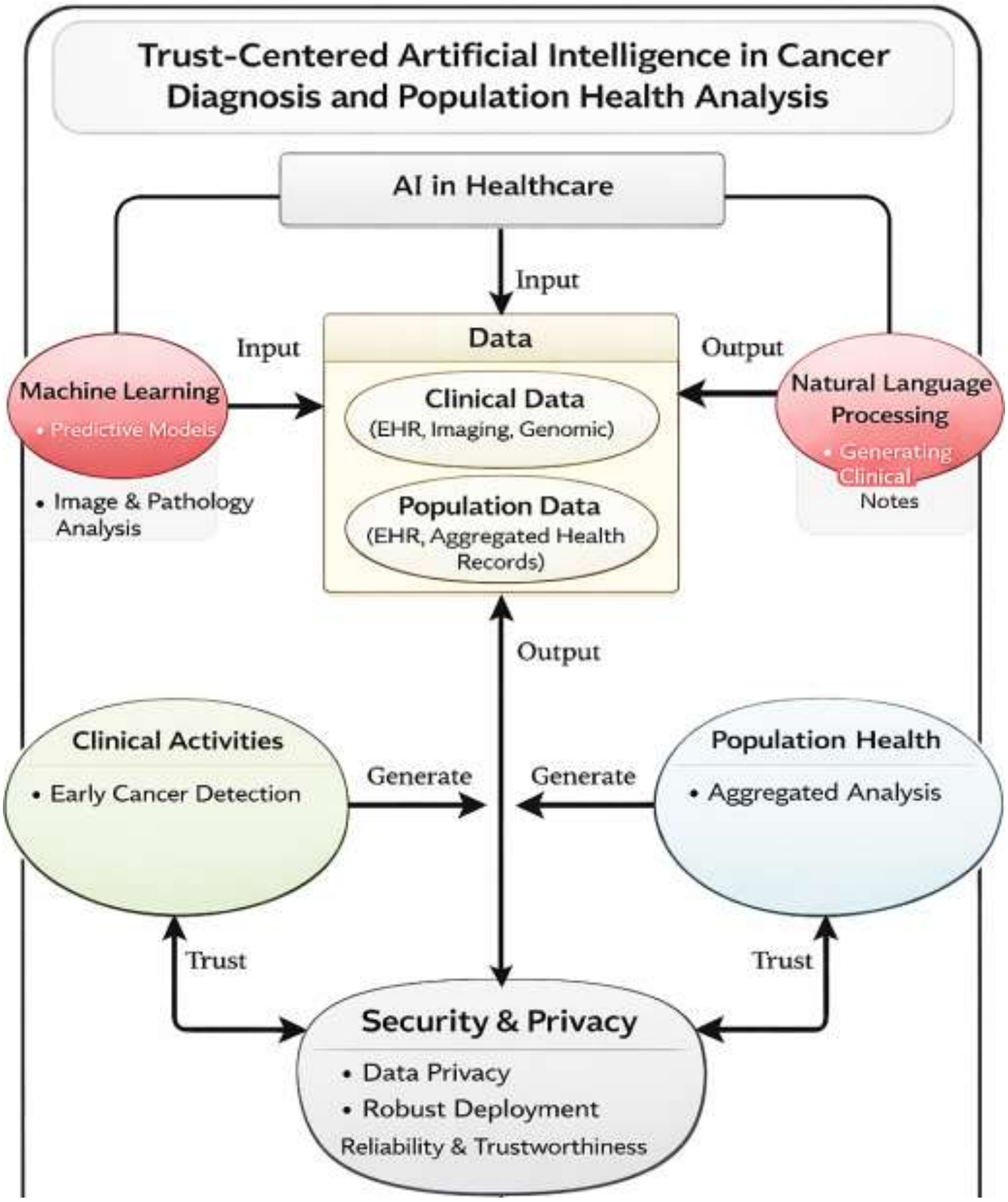
*Trust-Centered AI; Early Cancer Diagnosis; Security And Privacy Modeling; Population Health Analytics; Deployment Readiness.*

## INTRODUCTION

Artificial intelligence (AI) in healthcare can be defined as computational methods that perform tasks associated with human cognition—pattern recognition, reasoning, and learning—when applied to clinical data such as medical images, pathology slides, genomic measurements, and electronic health records (Abadi et al., 2016). Within AI, machine learning (ML) refers to algorithms that learn relationships from data, and deep learning (DL) describes multilayer neural networks that learn hierarchical representations capable of achieving high accuracy in perceptual tasks such as image classification and segmentation. In oncology, "early cancer diagnosis" commonly denotes the identification of malignancy at stages where intervention options and survival rates are substantially improved, often relying on screening programs and diagnostic workflows using radiology and pathology (Esteva et al., 2017). Population-level health analysis refers to the use of aggregated health data to describe disease patterns across communities and healthcare systems, including risk stratification, surveillance, and outcomes monitoring. Trust-centered AI, in this context, can be defined as AI development and deployment that foregrounds reliability, transparency, human oversight, and accountability so that clinical users can appropriately rely on outputs in settings characterized by uncertainty and vulnerability. Security modeling refers to the structured analysis of threats, attack surfaces, and controls across the AI lifecycle—including data collection, model training, inference, and system integration—so that confidentiality, integrity, and availability are protected when AI is deployed in operational healthcare environments (Litjens et al., 2017). Internationally, the significance of this research rests on shared challenges: cancer remains a leading cause of mortality worldwide, screening access is unequal across and within countries, and health systems increasingly depend on digitized infrastructure that creates both analytic opportunity and cyber risk. Within this global context, AI-enabled early detection is frequently positioned as a pathway to scalability and consistency in diagnostic decision support, particularly where specialist capacity is constrained, while the same digitization creates pressure for rigorous privacy and security practices to preserve public trust. These intersecting definitions anchor the need for a unified research framing that treats clinical accuracy, human trust, and secure deployment as structurally linked requirements for cancer AI used in clinical and public health settings (Machanavajjhala et al., 2007).

Research from 2005–2018 shows a steady shift from feature-engineered computer-aided detection toward data-driven learning, culminating in DL systems capable of extracting predictive features directly from medical images and digital pathology. Large-scale DL studies demonstrate high discriminative performance in clinically relevant classification tasks, including skin lesion assessment at dermatologist-comparable levels and lymph node metastasis detection in whole-slide pathology images within structured evaluation settings (Dwork, 2006). Parallel evidence in medical imaging for screening-related conditions supports the broader feasibility of DL in image-based diagnosis, exemplified by high-accuracy deep learning systems for diabetic retinopathy detection, which—while not cancer-specific—share similar screening logic. The consolidation of DL approaches across modalities is synthesized in major medical imaging reviews, which document rapid adoption of convolutional neural networks for classification, detection, and segmentation, along with recurring evaluation limitations such as dataset shift, inconsistent reporting practices, and constrained external validation. In oncology, workflow reality is shaped by heterogeneity in scanners, staining protocols, patient populations, and clinical decision thresholds; therefore, the literature commonly distinguishes internal testing from external validation and multi-site evaluation, with cross-case variation influencing claims of readiness for clinical translation (Gulshan et al., 2016). DL architecture advances that originate in general computer vision—such as residual networks—support medical performance gains by enabling deeper models and improved optimization, contributing indirectly to improved medical imaging performance and the feasibility of training high-capacity classifiers. At the same time, screening and early diagnosis in cancer requires more than high area-under-curve reporting; it also requires clinically interpretable error characterization, careful handling of false negatives, and operational performance when integrated into clinical pathways. The 2005–2018 literature includes growing acknowledgment that model behavior must be contextualized by clinical workflow constraints and decision support roles, particularly when AI outputs are used to triage cases, highlight suspicious regions, or assist time-constrained review (He et al., 2016). These patterns motivate a trust-centered orientation that treats "accuracy" as one part of a broader evidence package that includes generalization, clinician interaction, and system-level safety properties.

**Figure 1: Trust-Centered AI Framework for Cancer Diagnosis and Population Health**



Trust is a central construct in high-stakes automation because it shapes reliance decisions, including when users accept, ignore, or overuse automated outputs. In health AI, trust is anchored in both performance evidence and the human experience of interacting with uncertainty, error, and responsibility. Meta-analytic evidence on trust in human–robot interaction identifies measurable factors that influence perceived trust and reliance, including system performance, transparency cues, and contextual variables that shape expectations. A more focused synthesis of trust in automation organizes trust variability into dispositional, situational, and learned components, providing an empirical basis for understanding why identical model performance can produce different adoption outcomes across environments and users. Within clinical settings, "appropriate trust" can be conceptualized as calibrated reliance: the degree of reliance matches the system's validated competence under the current conditions (Hoff & Bashir, 2015). The 2005–2018 evidence base on medical AI highlights the difficulty of ensuring calibrated reliance when models are evaluated on limited distributions and then applied to more diverse or shifted conditions; therefore, clinical trust is linked to robustness evidence, external validation, and clear communication of model limitations. Trust is also influenced by interpretability and transparency mechanisms that help clinicians understand whether outputs are clinically coherent or potentially spurious. Visual explanation methods such as Grad-CAM link model predictions to salient image regions, supporting clinician sensemaking and enabling user studies that examine whether explanations help people differentiate stronger from weaker models (Bejnordi et al., 2017). Model-agnostic explanation frameworks such as LIME define explanation as local approximations that provide interpretable feature importance around individual predictions, explicitly framing explanation as a mechanism for trust and actionability. In medical imaging and pathology, explanation is particularly tied to spatial localization—identifying regions that justify a malignancy classification—because clinical reasoning often depends on identifiable morphological cues. The trust-centered literature during 2005–2018 therefore supports a framing where trust is not a subjective add-on but an empirically grounded mediator between evidence and reliance, directly influencing safety and effectiveness when AI supports early cancer diagnosis at scale.

Secure deployment in healthcare infrastructure requires attention to privacy-preserving data practices and the security of models and pipelines. Privacy in medical AI spans the protection of patient identity and sensitive attributes in data, alongside protection against leakage from trained models. Formal privacy frameworks such as differential privacy define privacy loss in rigorous probabilistic terms, enabling quantifiable guarantees when releasing statistics or training models on sensitive data. In healthcare contexts where patient datasets contain quasi-identifiers and sensitive clinical attributes, anonymization concepts beyond k-anonymity—such as ℓ-diversity and t-closeness—formalize protections against attribute disclosure by requiring diversity within equivalence classes and limiting distributional divergence. In the ML lifecycle, privacy threats extend beyond raw data releases: trained models can leak information about their training records (Fredrikson et al., 2015). Membership inference attacks formalize the risk that an adversary can determine whether a particular patient record was used in training, using only black-box access to model outputs; this matters directly for healthcare datasets where membership itself can be sensitive. Model inversion attacks show additional risk: confidence values and prediction interfaces can be exploited to reconstruct sensitive features, including in settings related to medical inference, which connects model design choices to privacy risk in deployment. Research also links privacy protection to learning procedures: deep learning with differential privacy operationalizes differential privacy in gradient-based training, demonstrating that privacy-preserving learning can be implemented within neural network training pipelines. These findings support a view of security modeling as a structured practice that evaluates threats not only to data storage but also to model interfaces, inference services, and confidence reporting. In healthcare infrastructure, these risks intersect with operational constraints and regulatory expectations, because clinical AI is often integrated into networked systems that handle protected health information and must maintain integrity and availability during clinical operations. A trust-centered approach therefore treats privacy and security not as external compliance steps but as intrinsic reliability conditions required for sustained public and clinical trust in early cancer diagnosis systems and population-level analytics (Hancock et al., 2011b).

Clinical reality introduces distributional complexity that challenges the stability of AI performance. In cancer diagnostics, small shifts in imaging acquisition, patient demographics, clinical prevalence, or labeling practices can change error profiles in ways not captured by narrow testing. Reviews of deep learning in medical image analysis emphasize recurring constraints: limited external validation, inconsistent dataset documentation, and underreporting of failure modes, all of which complicate the translation of reported metrics into operational expectations. Cancer-specific benchmarked evaluations illustrate both promise and challenge: lymph node metastasis detection competitions demonstrate that some deep learning systems can outperform pathologist panels under time-constrained simulation settings, while the broader evidence still depends on case definition, staining and scanner variability, and the role assigned to AI in workflow (Li et al., 2007). In skin cancer classification, dermatology-level performance emerges from large-scale supervised learning over diverse labeled lesion images, showing how scale and taxonomy design contribute to performance and how image-based classification tasks can approach expert baselines. Trust-centered evaluation requires mapping model performance to clinically meaningful questions: which errors occur, in which subgroups, and under which operating conditions. Explanation mechanisms contribute to this mapping by enabling clinicians and evaluators to inspect whether predicted malignancy is supported by clinically plausible regions or whether attention appears misdirected. Grad-CAM provides gradient-weighted localization for convolutional networks, enabling case-level visual inspection and user studies that connect explanations to user trust judgments. LIME offers local surrogate explanations for black-box predictions, positioning explanation as a means to assess trust at the level of individual cases and to identify problematic feature dependencies. These tools support qualitative synthesis approaches that treat trust not only as performance but also as interpretability, transparency, and the stability of decision logic across contexts. In practice, this aligns with a trust-centered model in which evidence for reliability and transparency is integrated with security and privacy requirements, because a clinically trustworthy system must remain reliable and safe while operating within real healthcare infrastructure constraints.

Population-level health analysis depends on aggregated data pipelines that link clinical records, imaging repositories, registries, and operational systems (Ribeiro et al., 2016). In the 2005–2018 literature, large-scale electronic health records are framed as an enabling substrate for predictive modeling and clinical risk stratification, with deep learning approaches demonstrating the ability to learn from raw records and to support diverse predictive tasks. When population analytics intersects with cancer, the same infrastructure can support surveillance, screening outreach optimization, and evaluation of diagnostic pathways, often requiring integration across multiple sites and data sources. Trust-centered AI in this setting requires credibility across the populations the system claims to represent, because population-level outputs can influence resource allocation and screening strategy decisions (Selvaraju et al., 2017). At the same time, population-scale data aggregation amplifies privacy and security exposure, including the risk that model training on sensitive cohorts or rare cancer subgroups increases re-identification or membership inference risk. Formal privacy approaches provide tools for managing this risk. Differential privacy offers a quantifiable privacy framework suitable for releasing statistics or training models, aligning naturally with population-level analysis where aggregate insights are emphasized. Anonymization criteria such as ℓ-diversity and t-closeness address attribute disclosure concerns in released microdata and demonstrate how privacy modeling evolves beyond simple k-anonymity assumptions, which matters when datasets include sensitive clinical attributes. Model-level privacy threats add additional complexity: model inversion shows that confidence outputs can leak sensitive training information and that interface design choices can materially affect privacy risk. Deep learning with differential privacy shows a concrete algorithmic route for training neural networks while bounding privacy loss, offering an operational method that aligns with large-scale healthcare datasets. Within a trust-centered framing, population health analysis therefore connects three requirements: analytic validity across heterogeneous populations, privacy preservation aligned with ethical and legal expectations, and security controls that protect the infrastructure and model interfaces used to deliver analytic outputs. The literature supports treating these as intertwined components of trustworthiness when AI supports early detection pipelines and population-level decision support (Rajkomar et al., 2018).

Deploying AI into U.S. healthcare infrastructure involves an end-to-end lifecycle that includes data governance, model development, validation, integration with clinical systems, and ongoing operational monitoring (Shokri et al., 2017). A trust-centered AI and security modeling approach treats this lifecycle as a unified system rather than a set of isolated technical steps. Trust synthesis in automation research provides empirical grounding for why trust varies across individuals and situations and why learned experience with system behavior shapes reliance over time, placing emphasis on designing for appropriate trust rather than maximal trust. In clinical AI for cancer diagnosis, the evidence base includes high-profile demonstrations of expert-level performance and competitive pathology results, supporting the feasibility of AI assistance when trained and evaluated under structured conditions (Dwork, 2006). Medical imaging reviews document that methodological consistency, external validation, and data diversity are recurring concerns that directly connect to generalization and trustworthiness claims. Interpretability methods provide mechanisms for auditing and communicating model behavior at the case level, enabling clinicians and evaluators to assess plausibility and detect anomalous reasoning patterns; Grad-CAM provides spatial localization for CNN decisions and LIME provides model-agnostic local explanations that support case-by-case trust judgments (Litjens et al., 2017). Security and privacy research demonstrates that model deployment introduces unique risks beyond data storage: membership inference can reveal whether an individual's data was used for training, and model inversion can reconstruct sensitive features from prediction confidence outputs. Formal privacy methods such as differential privacy provide a foundation for quantifying and controlling privacy leakage, while ℓ-diversity and t-closeness illustrate how privacy definitions address attribute disclosure risks in released datasets. Deep learning with differential privacy bridges privacy theory and neural network practice, showing implementable procedures that align with healthcare data sensitivity and large-scale training needs (Ribeiro et al., 2016). In parallel, large-scale EHR deep learning demonstrates how infrastructure standards and representations can support learning from raw clinical records, reinforcing the infrastructural dimension of trustworthy AI in health systems. Together, these strands define the conceptual space for a literature-review-based synthesis that frames early cancer diagnosis and population-level analysis as trust-dependent clinical functions that require security-modeled deployment conditions to maintain reliability, privacy, and operational integrity.

This study is structured around a set of objectives that collectively define a trust-centered AI and security modeling approach suitable for early cancer diagnosis, population-level health analysis, and secure deployment within U.S. healthcare infrastructure. The first objective is to systematically organize and classify prior scholarly evidence on AI-enabled early cancer diagnosis across the major clinical data modalities—radiology imaging, digital pathology, electronic health records, and multi-source fusion—so that the reviewed studies can be compared using consistent descriptors for clinical task type, dataset characteristics, validation design, and reported diagnostic performance. The second objective is to extract and synthesize the trust-oriented properties that are explicitly or implicitly evaluated in this literature, including interpretability mechanisms, calibration and uncertainty communication, robustness to data variability, human-in-the-loop workflow positioning, and the alignment of model outputs with clinician decision thresholds. The third objective is to analyze the extent to which reviewed studies address real-world clinical reliability through evidence of generalization, such as external validation, multi-site testing, subgroup analysis, and documented failure modes, and to develop a structured coding scheme that captures these indicators in a comparable manner across cases. The fourth objective is to consolidate security and privacy modeling practices reported across healthcare AI deployments by mapping threats, vulnerabilities, and controls to each stage of the AI lifecycle, including data governance, training pipelines, inference services, integration interfaces, and post-deployment monitoring. The fifth objective is to synthesize evidence related to fairness and bias in cancer AI and population-level analytics by assessing how studies report demographic representativeness, performance parity, bias mitigation strategies, and equity-centered evaluation logic, especially in contexts where outputs influence screening access, prioritization, and resource distribution. The final objective is to integrate the clinical trust evidence, security/privacy modeling insights, and equity considerations into a cohesive deployment blueprint that can be described as a conceptual framework, supported by cross-case patterns identified in the reviewed literature. This blueprint is intended to unify technical model evaluation, governance requirements, and operational safeguards into a single structure that can guide comparative assessment of existing approaches and provide a clear basis for organizing the findings in a manner consistent with a qualitative, cross-sectional, case-study–based literature review.

## LITERATURE REVIEW

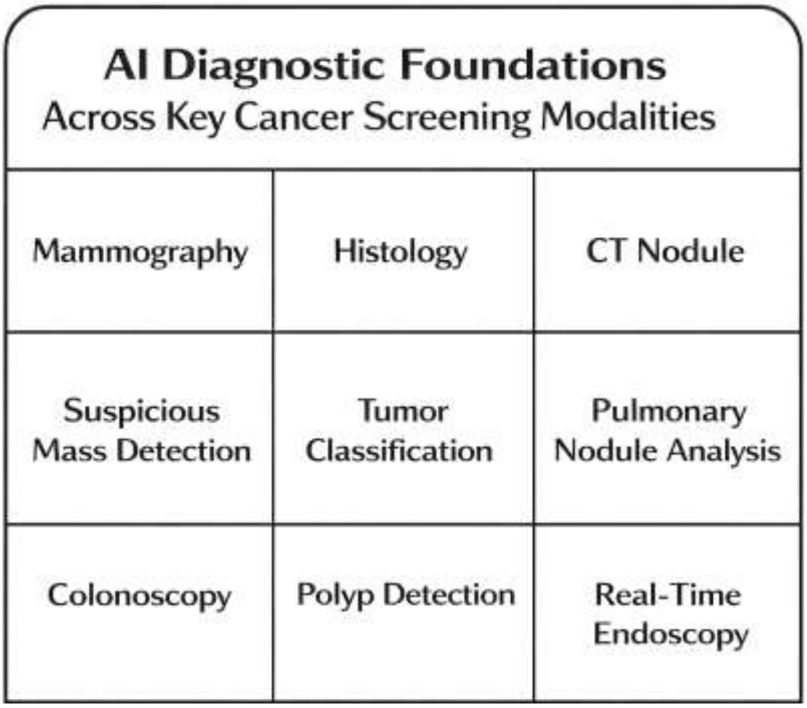
The literature on trust-centered artificial intelligence for early cancer diagnosis and population-level health analysis spans several intersecting domains, including medical imaging and digital pathology AI, clinical informatics and electronic health record modeling, human–AI trust and explainability research, and healthcare cybersecurity and privacy scholarship. Within cancer diagnostics, the core technical trajectory documented across studies moves from conventional computer-aided detection and handcrafted feature engineering toward data-driven machine learning and deep learning architectures capable of learning complex representations from radiology, pathology, and multi-omics data, with reported gains in sensitivity, specificity, and workflow support for screening and diagnostic decision-making. At the same time, this literature repeatedly emphasizes that strong predictive performance is not equivalent to real-world readiness, because clinical environments introduce heterogeneity in patient demographics, equipment, documentation practices, and disease prevalence, all of which can alter model reliability when systems are deployed outside controlled evaluation settings. As a result, a parallel strand of research focuses on trustworthiness as a multidimensional construct that includes interpretability, uncertainty awareness, transparency of decision logic, robustness under distribution shift, and the design of appropriate human-in-the-loop roles that support clinician judgment rather than replace it. A further layer of scholarship highlights the security and privacy implications of deploying AI systems in operational healthcare infrastructure, where protected health information, interconnected systems, and third-party integrations create exposure to data leakage, model inference risks, and service disruption, making security modeling and privacy-by-design essential to sustaining stakeholder confidence. When these technologies are applied at the population level, additional literature addresses the challenges of representativeness, bias, and equity, particularly because population analytics can influence screening prioritization, outreach strategy, and resource allocation across communities. Collectively, these research streams motivate an integrated framing in which trust-centered AI is treated as a lifecycle property spanning data governance, model development, validation, deployment, and monitoring, while security modeling and privacy controls are treated as foundational requirements that protect confidentiality, integrity, and availability without undermining clinical utility. This literature review therefore synthesizes evidence across technical performance studies, trust and interpretability research, and security and fairness scholarship to establish a coherent foundation for evaluating how early cancer diagnosis systems can be made reliable, accountable, and deployable within the complex realities of U.S. healthcare systems.

## AI Diagnostic Foundations Across Key Cancer Screening Modalities

Deep learning–enabled cancer screening research between 2015 and 2018 shows a clear shift from handcrafted feature pipelines toward end-to-end representation learning, with mammography and other image-intensive modalities serving as early proving grounds for clinical-grade pattern recognition. In mammography, the core diagnostic task is often framed as detection and classification of suspicious masses or calcifications under constraints of low contrast, heterogeneous breast density, and screening workflows that prioritize sensitivity while managing false positives. Large-scale training on curated lesion datasets supports convolutional architectures that learn discriminative visual cues directly from pixel data, enabling head-to-head comparisons against established computer-aided detection systems and demonstrating competitive or improved performance profiles across operating points (Kooi et al., 2017). Complementing these lesion-detection studies, representation learning work in mammography emphasizes that learned features can be combined with conventional descriptors or radiologist-guided information to yield practical gains, especially when datasets remain modest and label noise persists in real-world archives (Arevalo et al., 2016). These strands are important for a trust-centered framing because they anchor "trust" to concrete technical behaviors: how models generalize across scanners and sites, how they balance sensitivity-specificity tradeoffs, and how they can be audited via the stability of learned representations under clinically meaningful perturbations. They also define the baseline expectation that diagnostic AI must be evaluated not only as an algorithmic artifact but as a screening component that interacts with follow-up testing, biopsy pathways, and clinician decision thresholds. Within this study's scope, such modality-specific foundations provide the empirical substrate for later synthesis of robustness, uncertainty communication, and human-in-the-loop safety, while keeping attention on the operational realities that shape whether early cancer detection systems are dependable enough to support population-facing programs. Accordingly, modality evidence guides selection of evaluation metrics that align with screening impact, not only model accuracy. This anchors subsequent cross-case comparisons in shared clinical constraints.

Beyond radiographic screening, computational pathology provides another major pathway for early cancer diagnosis because histology encodes tumor architecture, grade, and microenvironmental signals that drive treatment decisions. In breast cancer histology, deep models have been used to classify hematoxylin-and-eosin–stained tissue images into clinically meaningful categories, reframing diagnostic support as a supervised learning problem over complex textures, glandular structures, and stromal patterns. A key implication for trust-centered modeling is that "ground truth" in pathology is often negotiated through expert interpretation and institutional practice, so model performance must be read alongside inter-rater variability, labeling protocols, and the representativeness of slide selection. Empirical work on histology classification shows that convolutional feature learning can reduce reliance on handcrafted descriptors and can perform competitively on public challenge data and curated datasets, while still reflecting sensitivity to staining variation and sample imbalance that are common in routine practice (Araujo et al., 2017). Parallel developments in thoracic computed tomography focus on pulmonary nodules, where early identification and risk stratification are central to lung cancer screening and surveillance pathways. Multi-scale and patch-based strategies treat nodules as localized candidates within noisy anatomical context, with learning objectives that approximate malignancy "suspiciousness" or related risk labels derived from radiologist consensus. The multi-crop pooling approach illustrates how architectural choices can encode clinically relevant scale variation without requiring explicit nodule segmentation, improving end-to-end learning and enabling interpretable links to semantic attributes such as margin or subtlety that clinicians already use in reporting (Shen et al., 2017). Together, histology and CT studies broaden the evidentiary base for early cancer AI by showing that trust is modality dependent: it is shaped by pre-analytic factors (fixation, staining, acquisition parameters), labeling regimes, and the stability of predictions across the range of presentations seen in heterogeneous patient populations. These findings motivate integrating diagnostic performance with uncertainty and context explicitly.

**Figure 2: AI Diagnostic Foundations Across Key Cancer Screening Modalities**



AI Diagnostic Foundations
Across Key Cancer Screening Modalities

| | | |
|---|---|---|
| Mammography | Histology | CT Nodule |
| Suspicious Mass Detection | Tumor Classification | Pulmonary Nodule Analysis |
| Colonoscopy | Polyp Detection | Real-Time Endoscopy |

Endoscopic imaging extends early cancer detection into procedural settings where real-time assistance can influence what lesions are noticed, biopsied, or removed, linking algorithmic perception directly to preventive outcomes. In colorectal screening, convolutional networks trained on large collections of annotated frames have been evaluated for their ability to localize and identify polyps within video streams under the latency constraints of live colonoscopy. This line of work operationalizes "trust" as a combination of high discriminative accuracy, bounded false alarms that do not distract clinicians, and stable behavior across diverse visual conditions such as bowel preparation quality, motion blur, and variable illumination. The reported performance of real-time polyp detection systems underscores that deep learning can be embedded into routine workflows on commodity hardware, creating a plausible path for scaling assistance across sites when paired with standardized quality metrics and careful validation designs (Urban et al., 2018). For a trust-centered AI and security modeling approach, the significance of endoscopy studies is not only technical; it is socio-technical because algorithm outputs become part of a human team's perceptual field, reshaping attention and potentially influencing documentation, follow-up intervals, and downstream treatment trajectories. When juxtaposed with screening mammography and CT nodule risk modeling, endoscopy highlights a shared set of methodological needs: reliable ground truth generation, transparent error characterization, and clinically meaningful thresholds that can be tuned to local practice. In turn, computational pathology adds a confirmatory layer that can reconcile imaging suspicions with tissue-level verification, supporting qualitative cross-case narratives about how evidence flows from screening to diagnosis. Taken together, these modality-spanning studies establish the diagnostic backbone for later sections of this review that examine robustness, bias, privacy and security controls, and deployment governance in U.S. healthcare infrastructure. They also clarify why a trust-centered model must bind technical evaluation to workflow fit and accountability structures across institutions consistently.

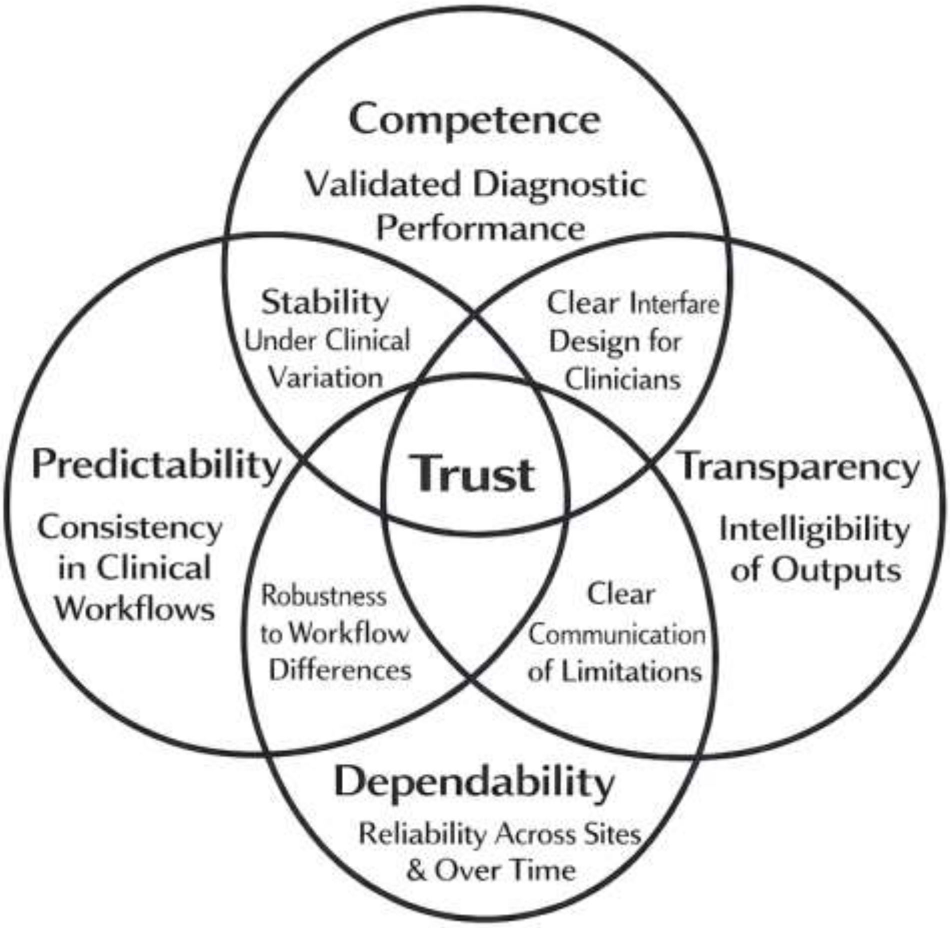
## Trust in Medical AI

Trust in medical artificial intelligence is commonly treated as a relational judgment that mediates how clinicians and organizations rely on algorithmic outputs under uncertainty, time pressure, and accountability constraints. In a trust-centered cancer AI context, trust is not identical to model accuracy; it functions as a multi-attribute belief about whether an AI system will behave competently, predictably, and safely within the user's operational environment. The literature on human–automation trust emphasizes that trust develops through both stable individual differences and experiences with system performance over time, which means two clinicians may react differently to the same AI evidence even when exposed to identical performance metrics. This perspective is useful for cancer screening and diagnosis because real-world adoption depends on how users perceive reliability, understand failure modes, and interpret the alignment between AI recommendations and clinical reasoning. Importantly, trust in this domain often concerns not only the algorithm, but the full socio-technical system that includes data provenance, workflow integration, and governance practices that define accountability for errors. Within this framing, trust is also influenced by whether the system supports appropriate reliance—neither disuse nor overuse—through cues that clarify capabilities, limitations, and expected behavior. Studies differentiating dispositional trust from history-based trust highlight how prior experiences with automation shape future reliance decisions, a pattern that translates directly into clinical settings where past exposure to false alarms or missed detections can affect willingness to use AI support. These principles underscore why trust-centered evaluation must document not only headline performance but also the conditions under which performance is stable, the nature of errors, and the user-facing signals that help calibrate reliance (Merritt & Ilgen, 2008).

A second strand of work clarifies trust as a measurable construct rather than a vague "acceptance" label, prompting researchers to define dimensions and corresponding indicators. In high-stakes automation, trust is often described through components such as perceived competence, predictability, dependability, transparency, and perceived intent or benevolence, with each component influencing reliance in different ways. For clinical AI, these dimensions map naturally to practical evaluation questions: competence relates to validated diagnostic performance across representative cases; predictability relates to stability under common clinical variations; dependability relates to consistency across sites and time; and transparency relates to the intelligibility of outputs and the visibility of system boundaries (de Visser et al., 2018). Meta-analytic evidence on trust in automation further shows that system capability and usability variables are strongly associated with trust development, suggesting that trust can be undermined by poor interface design, confusing outputs, or workflow friction even when the underlying model is strong. This matters for early cancer diagnosis because screening programs depend on rapid interpretation, low cognitive burden, and predictable system behavior across high volumes of cases. Trust is therefore shaped by the entire interaction cycle: how results are presented, how uncertainty is communicated, how clinicians can verify or contest outputs, and how the system responds to edge cases. From a measurement standpoint, this literature motivates including both subjective measures (e.g., trust ratings, perceived reliability, perceived understanding) and behavioral measures (e.g., reliance rates, override frequency, verification time) when synthesizing evidence across studies. By treating trust as a construct with identifiable antecedents and measurable consequences, a literature review can compare "trust evidence" across cancer AI studies even when they use different datasets and performance metrics (Schaefer et al., 2016).

Operationalizing trust for medical AI also requires linking it to interpretability and human oversight, because clinicians frequently seek reasons, not only predictions. In many healthcare settings, interpretability functions as a practical pathway for trust calibration: it enables clinicians to judge plausibility, recognize when the system might be reasoning from spurious cues, and decide when additional verification is required. The interpretability literature distinguishes between understanding the model globally (how it behaves in general) and explaining individual decisions (why a particular case was classified as suspicious), both of which matter in cancer diagnosis and screening where localized cues and subtle patterns can change clinical decisions. Interpretability is also tied to safety because it can support auditing, quality assurance, and error analysis, especially when combined with structured human-in-the-loop workflows that preserve clinician agency (Holzinger, 2016).

**Figure 3: Dimensions And Measurement Framework Of Trust In Medical AI**

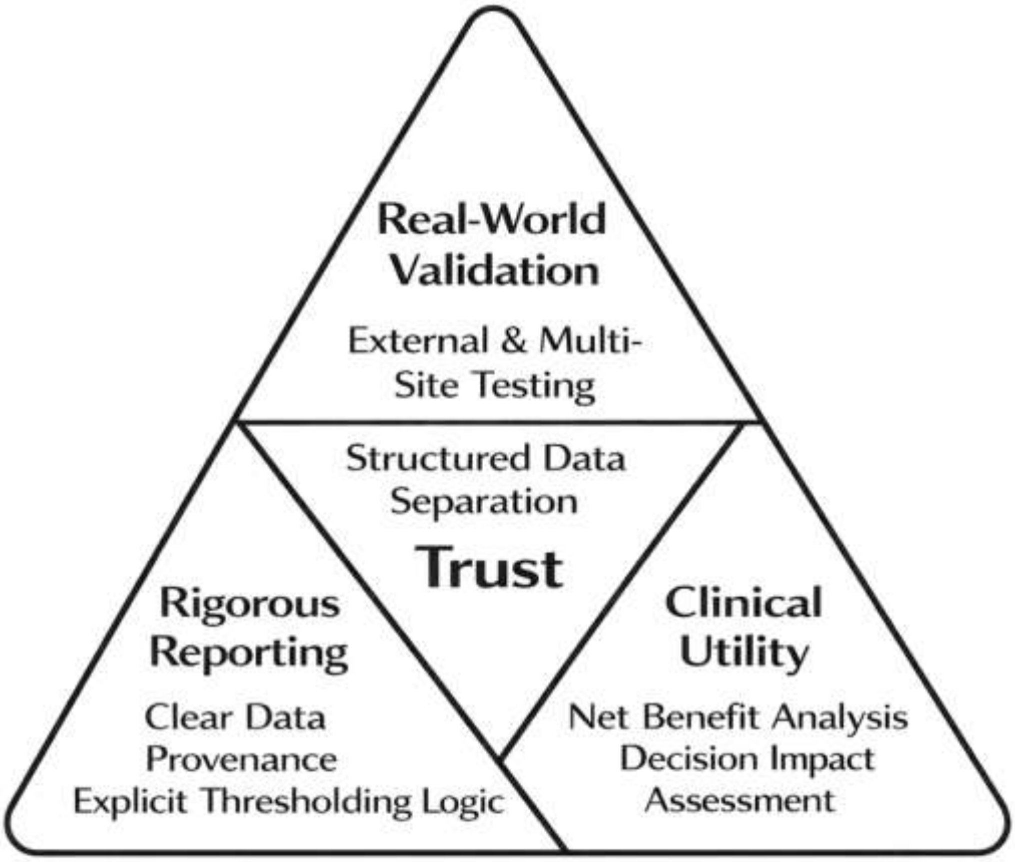

Competence
Validated Diagnostic Performance
Stability Under Clinical Variation
Clear Interfare Design for Clinicians
Predictability
Consistency in Clinical Workflows
Trust
Transparency
Intelligibility of Outputs
Robustness to Workflow Differences
Clear Communication of Limitations
Dependability
Reliability Across Sites & Over Time

Human-in-the-loop perspectives from health informatics highlight that clinical ML should be designed to leverage human expertise in labeling, validation, and deployment oversight rather than assuming fully autonomous operation, which aligns with trust-centered deployment models that prioritize appropriate reliance. Moreover, trust is dynamic: it can be strengthened or weakened by system performance over time, and it can require repair following errors, incidents, or unexpected behavior. Trust repair literature emphasizes that systems and organizations need mechanisms to respond to failures—through transparency, accountability, and corrective actions—because clinical environments routinely encounter rare cases and unexpected conditions that can expose model limits. Collectively, these ideas support a measurement approach in which trust is evaluated through a combination of interpretability evidence, workflow design evidence, and organizational governance evidence, not solely through statistical performance reporting (Montavon et al., 2018).

**Model Performance vs Clinical Utility**

Cancer-focused AI studies frequently report strong discrimination metrics (e.g., AUC, sensitivity, specificity), yet the literature from 2005–2018 shows that clinical value depends on *how* models are validated and *where* they fit inside diagnostic pathways. A recurring pattern is that many systems achieve high performance under retrospective, single-source testing conditions, while real-world utility requires evidence that performance holds under heterogeneous acquisition protocols, patient mixes, prevalence shifts, and institutional workflows. In digital pathology, this difference becomes visible when models are evaluated across slides from different laboratories, scanners, and staining conditions, where subtle distribution changes can influence error profiles. A prominent illustration is the use of deep learning to support histopathological diagnosis with explicit attention to efficiency and diagnostic accuracy, which also highlights why evaluation must include realistic slide diversity and pragmatic operational constraints, not only benchmark accuracy (Litjens et al., 2016). Similarly, nucleus-level detection and classification approaches in colorectal histology show that performance depends on annotation policy, sampling decisions, and patch selection strategies, all of which can inflate apparent accuracy if evaluation conditions mirror training too closely (Sirinukunwattana et al., 2016). These findings encourage interpreting performance claims through a validation lens: the strongest evidence for early cancer diagnosis systems comes from structured separation of development and validation data, transparent documentation of inclusion criteria, and multi-site testing where possible. From a trust-centered perspective, validation patterns are not merely methodological details; they shape whether clinicians and organizations can anticipate model behavior when it is embedded in diagnostic routines and quality assurance programs.

**Figure 4: Validation Patterns Linking Model Performance To Clinical Utility In Cancer AI**


Real-World Validation
External & Multi-Site Testing
Structured Data Separation
Trust
Rigorous Reporting
Clear Data Provenance
Explicit Thresholding Logic
Clinical Utility
Net Benefit Analysis
Decision Impact Assessment

A second theme concerns the reporting standards that allow readers to judge whether validation supports clinical translation. In medical prediction modeling, inconsistent reporting can conceal overfitting, dataset leakage, or selective threshold choices that limit reproducibility and reduce transferability across contexts. The TRIPOD guideline formalizes reporting expectations for development and validation studies of multivariable prediction models, providing a structured checklist that improves transparency around dataset selection, outcome definition, predictors, missing data handling, model specification, and performance reporting (Collins et al., 2015). Although TRIPOD is not specific to cancer AI, it directly addresses the gap between “model accuracy” and “model readiness" by requiring authors to describe how models were validated and how results should be interpreted. Cancer AI papers that include external validation, clear descriptions of data provenance, and explicit thresholding logic are more interpretable for clinical stakeholders because they allow evaluation of generalizability and deployment constraints. Diagnostic imaging studies that explicitly separate training from independent validation cohorts also provide stronger evidence of transportability, especially when the validation cohort differs in time period, site, or patient characteristics. For example, chest radiograph–based deep learning detection of malignant pulmonary nodules has been developed and validated with explicit comparison to physician readers, which helps translate algorithmic performance into clinical interpretability by aligning evaluation with actual diagnostic decision-making conditions (Nam et al., 2018). Collectively, reporting rigor and validation design determine whether performance metrics can be trusted as indicators of clinical utility, especially in early diagnosis contexts where false negatives carry substantial risk.

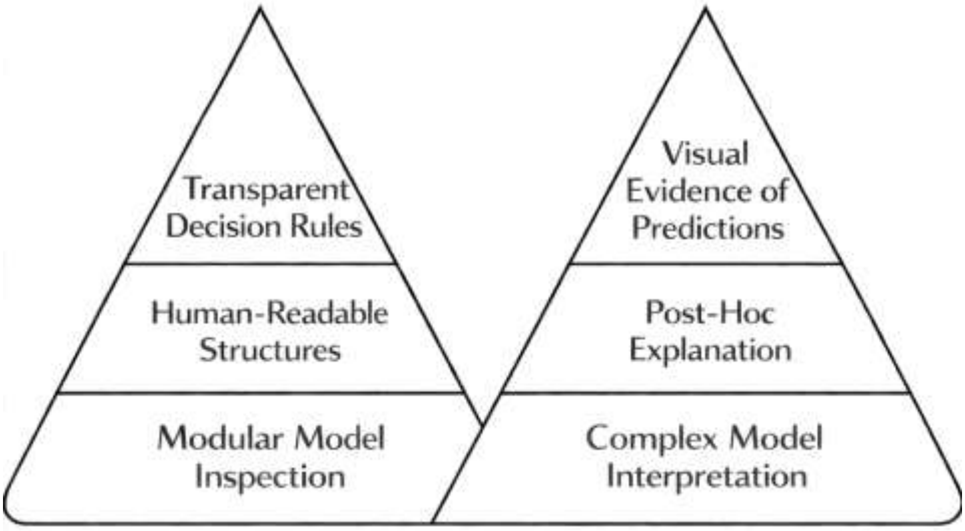
A third validation-related pattern concerns the evaluation of clinical utility beyond conventional accuracy metrics. Even when discrimination is high, clinical usefulness depends on decision thresholds, downstream consequences of false positives and false negatives, and whether model outputs change decisions in ways that improve outcomes or resource allocation. Decision curve analysis addresses this gap by evaluating net benefit across threshold probabilities, enabling comparison of strategies such as "treat all," "treat none," and "use the model," while incorporating the implicit trade-off between missing cancers and over-referring patients (Vickers & Elkin, 2006). This approach is especially relevant for screening and triage settings because threshold selection determines referral volume, follow-up testing rates, and clinician workload, which in turn shape feasibility and acceptance. In practice, an AI model with slightly lower AUC may still offer greater net benefit if it performs better at clinically relevant thresholds or reduces unnecessary interventions. Conversely, a high-AUC model may provide limited benefit if calibration is poor or if its useful operating region conflicts with workflow constraints. Validation patterns that include net-benefit reasoning, reader comparisons, and independent cohort testing therefore align more closely with trust-centered deployment goals because they connect statistical performance to human decisions and system-level consequences (Nam et al., 2018). When combined with transparent reporting and multi-site evidence, utility-oriented validation provides a stronger foundation for judging whether cancer AI can support reliable early diagnosis and scalable population-facing programs.

**Interpretability and Clinical Decision Support**

Interpretability and explainability in medical AI are commonly positioned as practical requirements for clinical decision support because they connect algorithmic outputs to the reasoning practices, accountability structures, and verification routines that characterize diagnostic medicine. In cancer screening and early diagnosis, clinicians rarely act on predictions in isolation; they interpret outputs alongside images, pathology features, patient history, and contextual constraints such as triage urgency and downstream testing capacity. As a result, the literature often distinguishes between *transparent* models whose internal logic can be directly inspected and *post-hoc* explanation methods that attempt to clarify black-box predictions after inference. This distinction matters for trust-centered deployment because clinical users need different forms of understanding depending on their role: a radiologist might need case-level localization cues, a pathologist might need a rationale aligned with morphological patterns, and an operational leader might need stable and auditable rules for governance and quality assurance. Surveys of explanation methods emphasize that "explainability" is not one universal property; it is defined relative to the user, the decision context, and the type of model, which means explanations should be evaluated as part of a socio-technical workflow rather than as standalone visualizations or feature weights (Guidotti et al., 2018). In screening programs, interpretability also functions as a safeguard against silent failure, because explanations can help detect when predictions rely on spurious correlates, data artifacts, or unintended shortcuts that may not generalize across sites. This reinforces a trust-centered view in which explanations become part of the evidence package supporting safe reliance: they assist in auditing, help communicate uncertainty, and enable clinicians to contest outputs when they conflict with clinical cues. Within this study's scope, interpretability and explainability are therefore treated as mechanisms that structure human–AI collaboration, not simply as optional add-ons, because clinical decision support requires that humans remain able to validate, contextualize, and justify decisions in high-stakes cancer care (Guidotti et al., 2018).

A second core theme is the role of inherently interpretable modeling strategies for decision support, particularly in contexts where clinicians and governance bodies need models that can be inspected, edited, or constrained to align with domain knowledge and safety policies. Interpretable modeling does not necessarily imply low accuracy; instead, the literature includes examples where intelligible models achieve strong performance while remaining modular and understandable, enabling domain experts to identify counterintuitive patterns, adjust model behavior, and maintain oversight of risk-sensitive predictions. This line of work is especially relevant to trust-centered cancer AI because early diagnosis errors carry asymmetric consequences, and clinical teams need tools that can be calibrated and justified in routine practice. In healthcare prediction tasks, intelligible high-accuracy models have been presented as a response to the practical limits of black-box deployment in mission-critical settings, emphasizing that interpretability supports validation, debugging, and communication across clinical and administrative stakeholders (Caruana et al., 2015).

**Figure 5: Interpretability And Explainability Pathways For Clinical Decision Support In Cancer AI**


Transparent Decision Rules
Human-Readable Structures
Modular Model Inspection
Visual Evidence of Predictions
Post-Hoc Explanation
Complex Model Interpretation

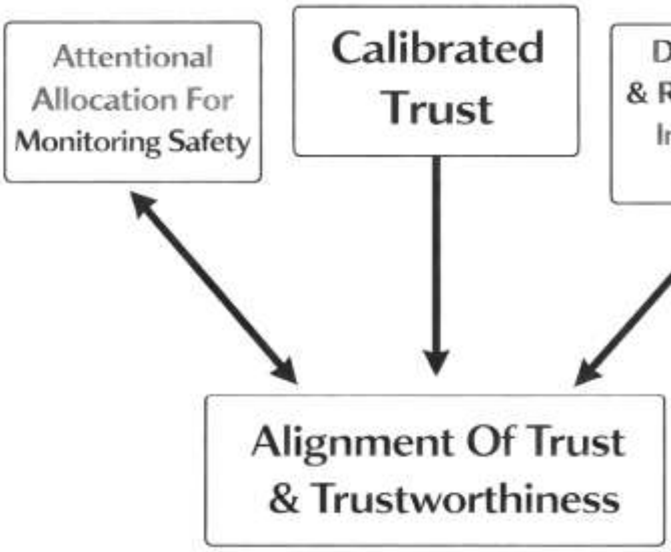
Complementing this perspective, rule-based and set-based frameworks aim to generate concise, human-readable structures that describe decision boundaries through independent if–then rules, supporting comprehension and enabling users to answer questions about model behavior more directly than with opaque predictors (Lakkaraju et al., 2016). For cancer screening and diagnostic pipelines, these approaches are conceptually valuable even when the final deployed model remains deep learning–based, because they define what "usable understanding" can look like in practice: a clinician can see which conditions trigger alerts, how combinations of findings affect risk, and where uncertainty or exceptions may arise. In trust-centered deployments, interpretable structures also support governance because they provide artifacts for policy review, documentation, and accountability. Accordingly, the literature positions interpretability as a bridge between performance and operational legitimacy, especially when models influence triage thresholds, referral cascades, and resource allocation decisions in U.S. healthcare settings (Lakkaraju et al., 2016). A third strand focuses on explanation methods for complex models, particularly deep neural networks, where strong predictive power often comes with limited transparency. In cancer-related imaging and pathology, clinicians often need explanations that connect predictions to clinically meaningful evidence—regions of interest, cellular morphologies, or structured attributes—so that the output can be verified and integrated into diagnostic reasoning. Pixel- and region-level explanation frameworks aim to decompose predictions into relevance scores over inputs, producing heatmaps that can be inspected for plausibility and used in error analysis. Layer-wise relevance propagation is a prominent approach that provides pixel-wise explanations by propagating relevance backward through the network, creating visual justifications that can support human review without requiring pixel-level supervision in training (Bach et al., 2015). In trust-centered clinical decision support, the value of such methods is not limited to visualization; explanation quality becomes a criterion for safety because it can reveal whether the system attends to medically appropriate cues or to confounded artifacts. At a broader level, explainable AI scholarship frames interpretability as essential for domains where decisions must be accountable and contestable, emphasizing that high-performing models used in sensitive applications require systematic approaches to understanding, visualizing, and interpreting learned representations (Samek et al., 2017). When combined with interpretable model classes and rule-based representations, explanation methods offer complementary pathways: one provides transparency by design, and the other provides evidence and auditing support for complex models that cannot be made fully transparent. For this research, these strands collectively justify treating interpretability and explainability as measurable components of trust-centered AI that influence adoption, calibration of reliance, documentation quality, and the capacity for safe deployment in diagnostic and population-health workflows (Guidotti et al., 2018).

**Trust-Centered Cancer AI and Secure Deployment**

A trust-centered approach to early cancer diagnosis requires a theoretical lens that treats trust as *calibrated reliance* rather than simple "acceptance" of an algorithm. In this view, clinicians and organizations continuously judge whether an AI system is reliable enough to justify delegating attention, altering decisions, or accelerating workflow. The core risk is not only low trust (leading to nonuse), but also inflated trust (leading to automation bias, reduced monitoring, and uncritical adoption in safety-critical contexts). The calibrated trust perspective explains why high predictive performance in the literature does not automatically translate into safe clinical uptake: trust must match the system's *demonstrated trustworthiness* within specific clinical conditions, including prevalence, data quality, and workflow stressors. A foundational mechanism shaping trust calibration is attentional allocation: when an automated recommendation is presented confidently or frequently, users may shift attention away from independent verification, increasing vulnerability to systematic errors and "automation wrong" effects. This is highly relevant for cancer screening pipelines (radiology triage, pathology prescreening, risk scoring) where speed and volume amplify the likelihood of overreliance. The attentional integration account formalizes how complacency and automation bias emerge under workload and imperfect automation, highlighting why trust-centered design must treat monitoring behavior as part of the safety case, not an optional human preference (Parasuraman & Manzey, 2010). For population-level health analysis, the stakes increase because outputs may influence screening prioritization, outreach targeting, and resource allocation, meaning miscalibration can propagate inequities and operational inefficiency at scale. Thus, the theoretical foundation for this study frames trust as an adaptive belief that must remain aligned with evidence of system capability, limits, and error modes under real clinical constraints.

A second theoretical pillar clarifies that trust is shaped by *relational dependency* and the perceived distribution of responsibility in human–automation teams. Clinical decision support tools are used in environments where responsibility is not merely individual; it is institutional and regulatory, shaped by documentation norms, medico-legal accountability, and governance protocols. Trust calibration is therefore affected by how the AI is positioned: as a "second reader," a triage filter, a quality assurance monitor, or a decision recommender. Empirical work on reliance behaviors shows that trust is influenced by perceived agency, role expectations, and how information from humans and machines is integrated when cues conflict, implying that trust should be analyzed alongside the collaboration structure rather than as a standalone attitude (Lyons & Stokes, 2012). This becomes operationally meaningful for cancer diagnosis when AI disagrees with a clinician's judgment: a trust-centered framework predicts that reliance will depend on whether the tool supports contestability (e.g., allows review, justification, or override), whether the team has shared mental models about failure modes, and whether feedback loops exist after errors. Meta-analytic evidence on human–robot trust provides complementary support by identifying performance-related and context-related factors that systematically influence trust across studies, reinforcing the idea that trust is dynamic and sensitive to cues of competence, reliability, and environmental risk (Hancock et al., 2011a). In this study, the theoretical implication for literature synthesis is that "trust evidence" includes not only reported accuracy, but also how studies treat responsibility boundaries, override mechanisms, and verification practices, because these shape calibrated reliance in real healthcare infrastructures.

**Figure 6: Calibrated Trust Theory Framework For Trust Centered Cancer AI And Secure Deployment**


Attentional Allocation For Monitoring Safety
Calibrated Trust
Dependency & Responsibility In Human-AI Interaction
Alignment Of Trust & Trustworthiness

To support a small numeric synthesis aligned with a qualitative literature review, this study adopts a simple, reusable **Calibrated Trust Index (CTI)** that can be extracted or approximated from published evidence when trust or reliance measures are available. The CTI operationalizes the principle that trust should match trustworthiness:

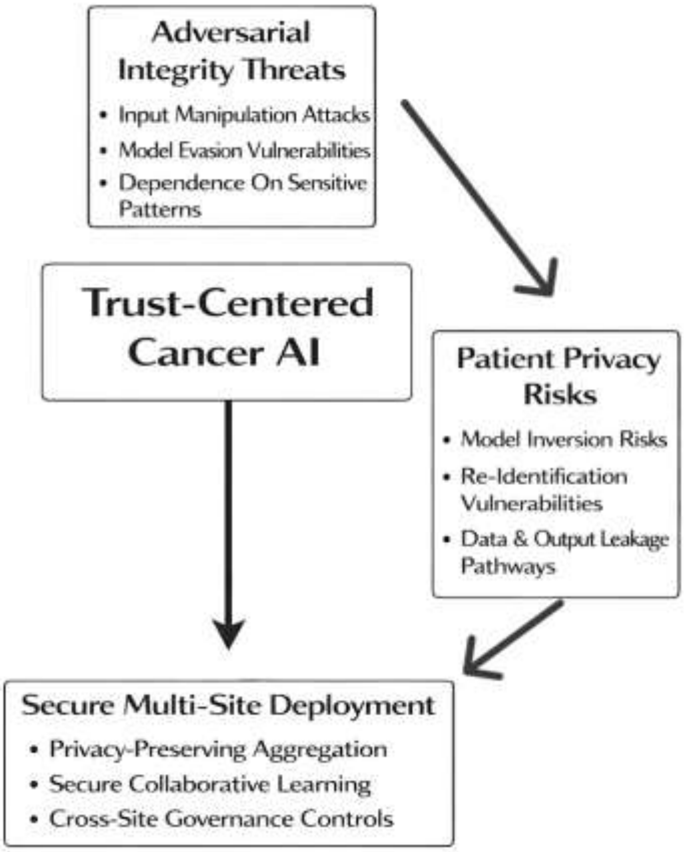
$$\mathrm{CTI} = 1 - | T - W |$$

where $T$ is *normalized trust* (e.g., standardized self-reported trust, reliance rate, or acceptance probability scaled to [0, 1]), and $W$ is *normalized trustworthiness* (e.g., a bounded performance indicator such as AUC, balanced accuracy, or a reliability score scaled to [0, 1]). CTI approaches 1 when trust aligns with evidence of capability and approaches 0 when trust is miscalibrated (either overtrust or undertrust). This formula is intentionally conservative: it does not assume that higher trust is better; it treats *alignment* as the objective. The rationale for using an alignment index is supported by studies showing that misuse and automation bias can occur even when users have experienced imperfect automation, particularly when interface cues encourage overreliance or when verification is cognitively expensive (Wickens et al., 2015). In addition, transparency and user experience can shift trust trajectories over repeated interactions, indicating that the same system may produce different calibration outcomes depending on how its reasoning and limits are communicated (Yang et al., 2017). In this literature review, CTI will be applied as a lightweight cross-study indicator (reported as ranges or illustrative values) to support the findings section without converting the study into a meta-analysis, while remaining faithful to the trust-centered theory that safe deployment depends on calibrated reliance rather than maximal trust.

### Trust-Centered Cancer AI in U.S. Healthcare Infrastructure

Security and privacy are not peripheral concerns in trust-centered cancer AI; they are foundational determinants of whether diagnostic and population-analytic outputs remain credible when systems are deployed inside complex U.S. healthcare infrastructure. In practice, early cancer diagnosis models consume high-value clinical data (images, pathology slides, EHR features) and often operate through interconnected pipelines (PACS/RIS/LIS/EHR integrations, cloud inference endpoints, and vendor-maintained analytics stacks). This interconnectedness expands the attack surface across the AI lifecycle: data collection, labeling, storage, training, model distribution, inference, and monitoring. A trust-centered literature synthesis therefore treats “security” as the preservation of confidentiality, integrity, and availability of both data and model behavior under realistic adversarial pressures. Integrity threats are especially salient for early diagnosis because an attacker who can manipulate inputs or intermediate representations can induce missed detections (false negatives) or inflate referrals (false positives), both of which carry high clinical and operational cost. The broader machine-learning security literature formalizes this issue through evasion attacks at test time, where adversaries introduce small but targeted changes to inputs to cross decision boundaries while retaining human plausibility. Although developed in security domains such as malware detection, this framework is directly transferable to medical imaging and structured clinical features because diagnostic models also rely on pattern-sensitive representations that can be perturbed in targeted ways (Biggio et al., 2013). For U.S. healthcare settings, the practical significance is that many pipelines include external interfaces (patient portals, image exchange networks, third-party analytics) where manipulation opportunities may exist, even without insider access. Security modeling in this context therefore aligns with trust-centered objectives by demanding explicit articulation of threat actors, entry points, and consequences for diagnostic reliability, rather than treating model performance as stable and benign across deployment contexts.

**Figure 7: Security Threat Modeling For Trust Centered Cancer AI In U.S. Healthcare Infrastructure**


Adversarial Integrity Threats
• Input Manipulation Attacks
• Model Evasion Vulnerabilities
• Dependence On Sensitive Patterns
Trust-Centered Cancer AI
Patient Privacy Risks
• Model Inversion Risks
• Re-Identification Vulnerabilities
• Data & Output Leakage Pathways
Secure Multi-Site Deployment
• Privacy-Preserving Aggregation
• Secure Collaborative Learning
• Cross-Site Governance Controls

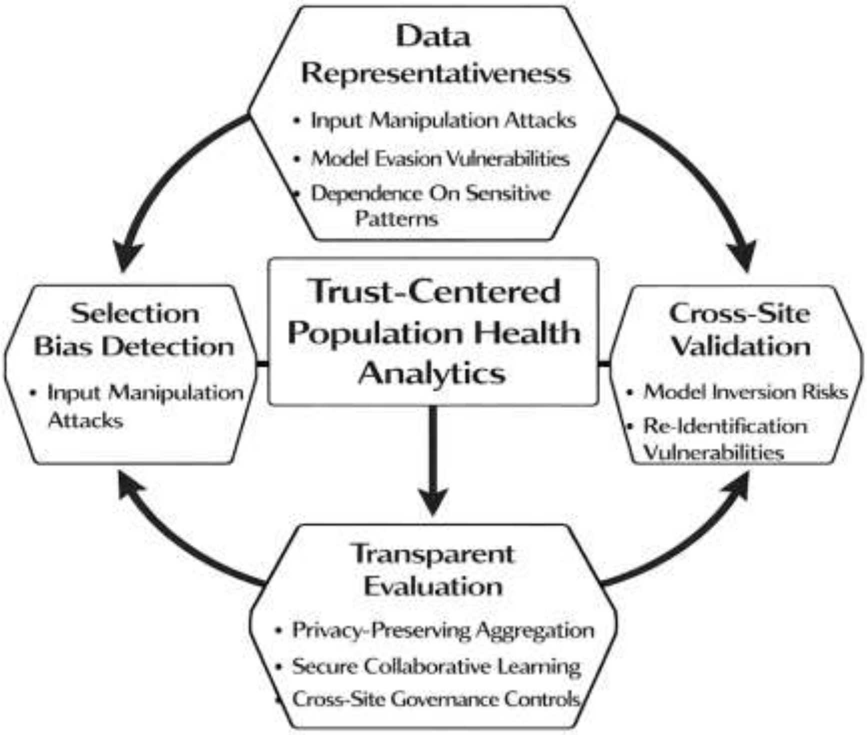
A second security dimension concerns privacy leakage and inference risks created by the very act of releasing model outputs, hosting prediction APIs, or sharing derived datasets for multi-institutional evaluation. Population-level cancer analytics can be particularly vulnerable because large-scale datasets often combine demographics, utilization patterns, and longitudinal outcomes that may enable re-identification, even when direct identifiers are removed. De-anonymization research demonstrates that high-dimensional, sparse microdata can be re-linked to individuals with surprisingly little auxiliary information, challenging simplistic assumptions that "de-identified" data automatically protects patients (Narayanan & Shmatikov, 2008). When AI models are deployed as services, privacy risk also extends beyond the dataset to the model itself: attackers can query models and exploit differences in outputs to infer decision boundaries or approximate model behavior, creating pathways for misuse and potential reconstruction of sensitive patterns. Black-box attack research further shows that limited query access can be sufficient to craft adversarial inputs that systematically induce misclassification, which becomes a direct patient-safety concern for cancer screening if an adversary can influence clinical inputs or data flows (Papernot et al., 2017). Moreover, robustness evaluations reveal that defenses that appear effective under narrow assumptions may fail under stronger adaptive attacks, reinforcing the need for conservative threat modeling rather than assuming a single defensive technique will generalize across contexts (Carlini & Wagner, 2017). In a trust-centered healthcare framing, these findings imply that privacy and robustness must be considered together: the more accessible the model (to support interoperability and scale), the more the deployment must anticipate query-based threats, adversarial manipulation, and the disclosure risks tied to confidence scores, probability outputs, and repeated querying behaviors.

Finally, secure deployment in U.S. healthcare infrastructure increasingly depends on governance choices for collaboration, especially when cancer AI seeks multi-site evidence for generalization and equity. Multi-institution learning can reduce bias and improve external validity, but it introduces data-sharing risks, contractual complexity, and new technical vulnerabilities. A key security modeling pathway is to minimize raw data movement while still enabling cross-site learning and evaluation, which motivates privacy-preserving aggregation designs that protect participant contributions during distributed training. Secure aggregation protocols, developed for federated learning settings, aim to ensure that a server can compute only an aggregate update (e.g., a sum of gradients) without revealing any single site's update, reducing exposure of institution-level or patient-level signals during collaborative model development (Bonawitz et al., 2017). For a trust-centered cancer AI blueprint, this matters because trust is partly institutional: hospitals and health systems require credible assurances that collaboration will not leak sensitive data, undermine compliance obligations, or create new breach pathways. At the same time, secure aggregation does not eliminate the need for robust monitoring against integrity threats (e.g., malicious updates, poisoned contributions, or compromised endpoints), so a full security model must combine cryptographic protections with lifecycle controls such as access governance, audit logging, anomaly detection, and controlled output interfaces. In a literature-review synthesis, these elements become evidence categories: whether studies discuss threat surfaces, whether they constrain outputs to reduce leakage, whether they incorporate adversarial testing, and whether they propose privacy-preserving collaboration mechanisms that fit U.S. healthcare interoperability realities. Taken together, security modeling provides the "hard boundary" conditions under which trust-centered cancer AI claims can remain credible in deployment.

**Population-Level Health Analytics and Equity Considerations**

Population-level health analysis in trust-centered cancer AI refers to the systematic use of large, routinely collected clinical and administrative data to characterize cancer risk, screening coverage, diagnostic yield, treatment trajectories, and outcome disparities across groups and settings. In the 2005–2018 evidence base, the growth of electronic health records (EHRs) and large clinical repositories enabled machine-learning models that move beyond single-task prediction to broader phenotyping and longitudinal risk modeling that can support population surveillance and targeted prevention strategies. A major contribution of this literature is the demonstration that representation learning can compress complex patient histories into reusable embeddings that facilitate multiple downstream predictive tasks, which is crucial for population-level work because it reduces the need to handcraft features separately for every clinical question. For example, unsupervised patient representations learned from EHR data have been used to forecast future diagnoses and outcomes, providing a general-purpose modeling approach for large health systems where cancer risk assessment must be integrated with comorbidities, utilization patterns, and care pathways (Miotto et al., 2016). The relevance to cancer diagnostics emerges when these representations are linked to screening eligibility identification, missed-screening detection, diagnostic follow-up prediction, and stratified outreach planning at scale. As population analytics expand, data governance and transparency become inseparable from model quality because outputs may influence screening prioritization and resource distribution. Consequently, population health AI must be assessed not only for discrimination metrics but also for representativeness, data provenance, and interpretability to ensure that the produced insights align with clinical and public health objectives. In trust-centered research, population-level AI is therefore treated as a socio-technical instrument that can magnify both benefits and harms: high-quality modeling can reduce missed diagnoses and optimize screening pathways, while biased data capture or unrepresentative training cohorts can systematically distort population conclusions and reproduce inequities across communities.

**Figure 8: Population Level Health Analytics And Equity Framework For Trust Centered Cancer AI**


Data Representativeness
• Input Manipulation Attacks
• Model Evasion Vulnerabilities
• Dependence On Sensitive Patterns
Selection Bias Detection
• Input Manipulation Attacks
Trust-Centered Population Health Analytics
Cross-Site Validation
• Model Inversion Risks
• Re-Identification Vulnerabilities
Transparent Evaluation
• Privacy-Preserving Aggregation
• Secure Collaborative Learning
• Cross-Site Governance Controls

Equity considerations become central once population-level models are built from EHR data because EHRs are not neutral recordings of reality; they reflect patterns of access, documentation practices, reimbursement rules, and clinical workflow variability. These factors introduce selection effects—who enters the dataset and with what completeness—along with measurement effects—what is recorded, coded, and updated—and these distortions can differentially affect groups by race, income, geography, language, insurance type, and comorbidity burden. A practical equity risk is that population-level models may interpret under-documentation (or delayed documentation) as true absence of risk factors, leading to underestimated cancer risk or reduced outreach in already underserved groups. Evidence on EHR reuse highlights that bias can originate across multiple stages of the data chain, including data entry behaviors, coding variability, extraction and transformation steps, and secondary-use dataset construction, implying that fairness is partly determined before any model is trained (Verheij et al., 2018). From a trust-centered viewpoint, this means equity cannot be "patched in" at the end; it must be traced back to data-generating processes and the institutional incentives that shape recording behavior. Complementary methodological reviews of deep learning on EHRs emphasize that limitations in data quality, missingness, labeling, and interoperability are recurring barriers that can influence performance and fairness simultaneously, especially when models are transferred across health systems serving different populations (Xiao et al., 2018). For cancer-focused population analytics, these concerns become more acute because screening behaviors, follow-up adherence, and diagnostic timeliness are socially patterned; therefore, model outputs may reflect structural barriers rather than intrinsic patient risk. Accordingly, equity-aware population analytics requires explicit reporting of cohort composition, stratified evaluation, documentation quality checks, and sensitivity analyses that examine how missingness and access patterns affect conclusions.

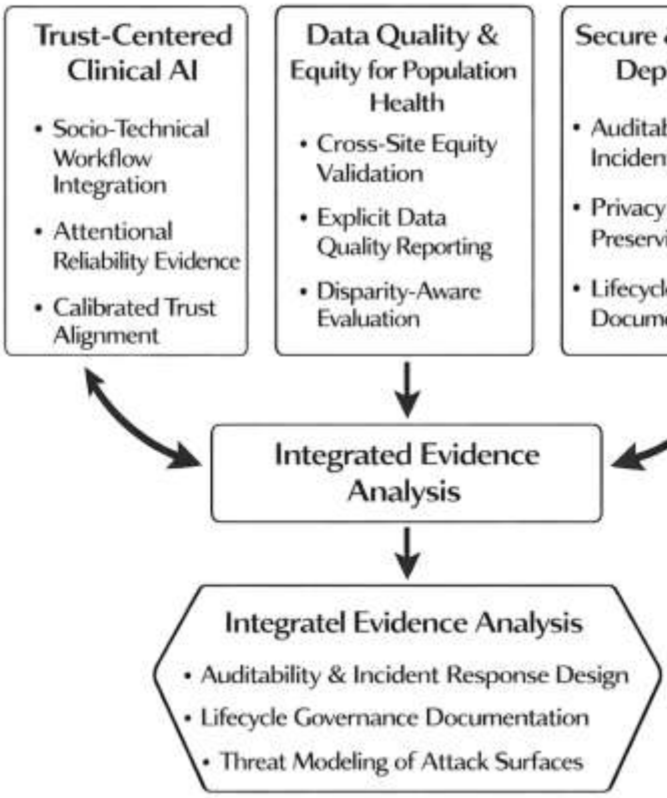
Population-level analytics also depends on the availability of large, well-documented datasets and reproducible pipelines that enable cross-site validation and transparent benchmarking, because equity claims require evidence that generalization holds across diverse populations and care settings. Open clinical databases and standardized schemas support this goal by enabling methodological replication and stress-testing across institutions, while also highlighting how privacy constraints and governance decisions shape what can be shared. A widely used example is a large critical care database that provides structured and unstructured EHR data with detailed documentation, enabling research on predictive modeling, clinical trajectories, and subgroup analyses that are foundational for population-level evaluation designs (Johnson et al., 2016). For trust-centered cancer AI, analogous data principles apply: population models should be validated across settings with distinct demographic profiles and care processes, and their outputs should be interpretable enough to support program-level decisions about screening and follow-up. A closely related equity dimension is the inclusion of social determinants of health (SDoH), because cancer outcomes are strongly influenced by nonclinical factors such as housing stability, food security, transportation, and neighborhood context. The literature on integrating SDoH into EHRs argues that population health management improves when social and behavioral risk factors are represented in standardized, interoperable formats, enabling stratification that reflects real constraints on screening participation and treatment adherence (Alderwick & Gottlieb, 2018). Within a trust-centered framework, SDoH integration is also a fairness mechanism: it reduces the likelihood that models interpret structural disadvantage as individual "noncompliance" and enables population interventions aligned with the causes of delayed diagnosis. Collectively, these studies justify treating population-level health analysis and equity as a unified evidence domain where dataset design, documentation practices, and evaluation strategy determine whether cancer AI supports reliable and just screening and prevention programs.

**Trust-Centered AI + Security Modeling for Cancer Diagnosis**

This study adopts a trust-centered socio-technical conceptual framework that treats early cancer AI as an intervention embedded in U.S. healthcare infrastructure rather than as a standalone predictive artifact. The framework begins with the premise that diagnostic reliability and population-level usefulness emerge from interactions among (a) clinical workflow roles, (b) data and technology architecture, (c) organizational governance, and (d) continuous evaluation. A practical way to structure these interactions is to position the AI system inside an eight-dimensional socio-technical environment that includes hardware/software, clinical content, people, workflow/communication, internal policies, external rules/pressures, system measurement/monitoring, and organizational culture. This framing supports a literature-review synthesis because it creates a consistent lens for comparing studies that differ in modality (radiology, pathology, EHR), deployment setting (single hospital vs multi-site networks), and operational use (triage, second reader, QA, risk stratification). It also aligns with the trust-centered requirement that trust be evaluated as a *property of the full system lifecycle*—from data governance and model development to integration and post-deployment monitoring—because failures in any socio-technical dimension can degrade clinical reliability even when model accuracy appears high in controlled evaluation. Accordingly, the framework treats "trust" as a multi-domain outcome: clinicians' calibrated reliance, organizational confidence in governance, and population-level legitimacy of analytics for screening and policy decisions. By anchoring synthesis in socio-technical dimensions, the literature review can code evidence about workflow fit, accountability boundaries, validation design, and operational safeguards as explicit trust contributors rather than leaving them as unstructured narrative background (Sittig & Singh, 2010).

The second layer of the conceptual framework formalizes data quality and fairness as prerequisites for trustworthy population-level inference, particularly when cancer AI outputs are used to influence screening access, follow-up prioritization, and resource allocation. The framework treats secondary-use clinical data (EHR and linked repositories) as imperfect proxies for patient reality that must be evaluated through harmonized data-quality dimensions—conformance (format and structural integrity), completeness (presence of required fields and coverage), and plausibility (logical and clinical credibility of values). These three dimensions provide an audit structure for synthesizing literature evidence on dataset construction, missingness patterns, and cross-site comparability, which is essential when trust-centered claims depend on generalization across institutions and demographic groups. Equity is then modeled as an additional validity condition: population-level performance must be examined for disparities and for the incompatibilities among fairness criteria when prevalence differs between groups (Cath et al., 2018). The framework therefore operationalizes fairness evidence through subgroup reporting, disparity-aware evaluation, and documentation of trade-offs (for example, whether equalized error rates can be achieved without changing thresholds or sacrificing calibration). This approach is conceptually aligned with the idea that population-level analytics can be simultaneously powerful and risky: it can reveal screening gaps and support targeted prevention, while biased data capture or miscalibrated risk scoring can reinforce inequities at scale. In synthesis terms, studies are treated as stronger evidence when they (a) document data-quality checks using explicit categories, (b) report subgroup performance or disparity diagnostics, and (c) acknowledge fairness trade-offs as part of program-level decision logic (Kahn et al., 2016).

**Figure 9: Integrated Trust Centered And Security Modeling Framework For Cancer Diagnosis And Population Health**


Trust-Centered Clinical AI
• Socio-Technical Workflow Integration
• Attentional Reliability Evidence
• Calibrated Trust Alignment
Data Quality & Equity for Population Health
• Cross-Site Equity Validation
• Explicit Data Quality Reporting
• Disparity-Aware Evaluation
Secure & Governed Deployment
• Auditability & Incident Response
• Privacy & Security Preserving Design
• Lifecycle Governance Documentation
Integrated Evidence Analysis
Integratel Evidence Analysis
• Auditability & Incident Response Design
• Lifecycle Governance Documentation
• Threat Modeling of Attack Surfaces

The third layer integrates policy-oriented governance and security modeling into a single deployability view, because a trust-centered cancer AI system must remain clinically reliable *and* institutionally legitimate under privacy, security, and accountability constraints. Governance scholarship emphasizes that AI adoption in high-risk sectors depends on explicit accountability structures, enforceable oversight, and clarity on how ethical and legal expectations are operationalized inside technical systems and organizational procedures. Within this study, governance is represented as a lifecycle control plane that links model development practices to real deployment conditions, including auditability, documentation, and incident response. Population-level health analytics further elevates governance needs because decisions influence program design and public health priorities, requiring transparent risk management and alignment with policy objectives (Ashrafian & Darzi, 2018). The framework therefore treats security and privacy modeling as enabling conditions for trust rather than external compliance steps: trust is undermined if confidentiality, integrity, or availability risks can distort model behavior or leak sensitive health information. To support the "little numeric" requirement in a literature-review-friendly way, this study applies one consistent quantitative indicator across cases: a Trust-Centered Deployment Readiness score (TDR) computed as a harmonic mean of four normalized evidence components—calibrated trust alignment (CTI), security modeling strength (S), fairness/equity evidence strength (F), and data-quality evidence strength (DQ):

$$\mathrm{TDR} = \frac{4}{\left(\frac{1}{CTI} + \frac{1}{S} + \frac{1}{F} + \frac{1}{DQ}\right)}$$

Each component is scaled to $[0,1]$using transparent rubric-based coding from the reviewed studies (e.g., 0.25 = minimal evidence; 0.50 = moderate; 0.75 = strong; 1.00 = comprehensive). The harmonic mean is intentionally “strict”: a weak link in security, fairness, or data quality pulls down readiness even if one dimension is strong, which matches the trust-centered logic that safe deployment requires concurrent satisfaction of clinical reliability and infrastructure safeguards. This conceptual and numeric integration supports literature synthesis that remains qualitative in interpretation while still enabling structured, objective comparison across studies and cases in the findings section (Chouldechova, 2017).

## METHODS

In this study, the methodology has been designed to support a literature review–based, qualitative, cross-sectional, case-study–oriented synthesis of research on trust-centered AI for early cancer diagnosis, population-level health analysis, and secure deployment in U.S. healthcare infrastructure. The methodological approach has combined structured evidence identification with interpretive thematic synthesis so that both technical performance findings and socio-technical deployment evidence have been captured in a unified analytic frame. A review protocol has been established to guide database searching, screening, eligibility assessment, and data extraction, and the protocol has been aligned with transparent reporting expectations commonly used in evidence synthesis. The literature corpus has been defined to include peer-reviewed studies and high-quality conference proceedings that have examined cancer-related diagnostic AI, healthcare machine-learning trust and interpretability, privacy and security modeling, and equity-focused population analytics within the specified scope of the research title. The screening process has been implemented through sequential filtering stages, beginning with title–abstract review and continuing to full-text assessment, so that relevance to early diagnosis, population-scale analysis, trust mechanisms, and secure deployment considerations has been consistently enforced.

To enable cross-sectional comparison, each eligible study has been treated as an analytic “case” representing a particular deployment context, clinical modality, institutional setting, or socio-technical configuration, and case attributes have been documented using a standardized extraction template. Data extraction has captured bibliographic metadata, clinical domain and modality, dataset characteristics, model type and evaluation design, reported diagnostic performance, interpretability and uncertainty handling, robustness and generalization evidence, fairness and subgroup reporting, and explicit privacy or security measures. A coding framework has been constructed using both deductive categories derived from the study’s theoretical and conceptual frameworks and inductive codes that have emerged from repeated reading of the included papers. The synthesis has been conducted through thematic analysis and cross-case pattern matching, enabling the identification of recurring trust and security determinants as well as consistent evidence gaps. Finally, a light numeric synthesis has been integrated by converting selected indicators—such as the presence of external validation, security threat modeling, and subgroup evaluation—into frequency summaries and rubric-based scores, thereby supporting objective comparisons while preserving the qualitative orientation of the overall methodology.

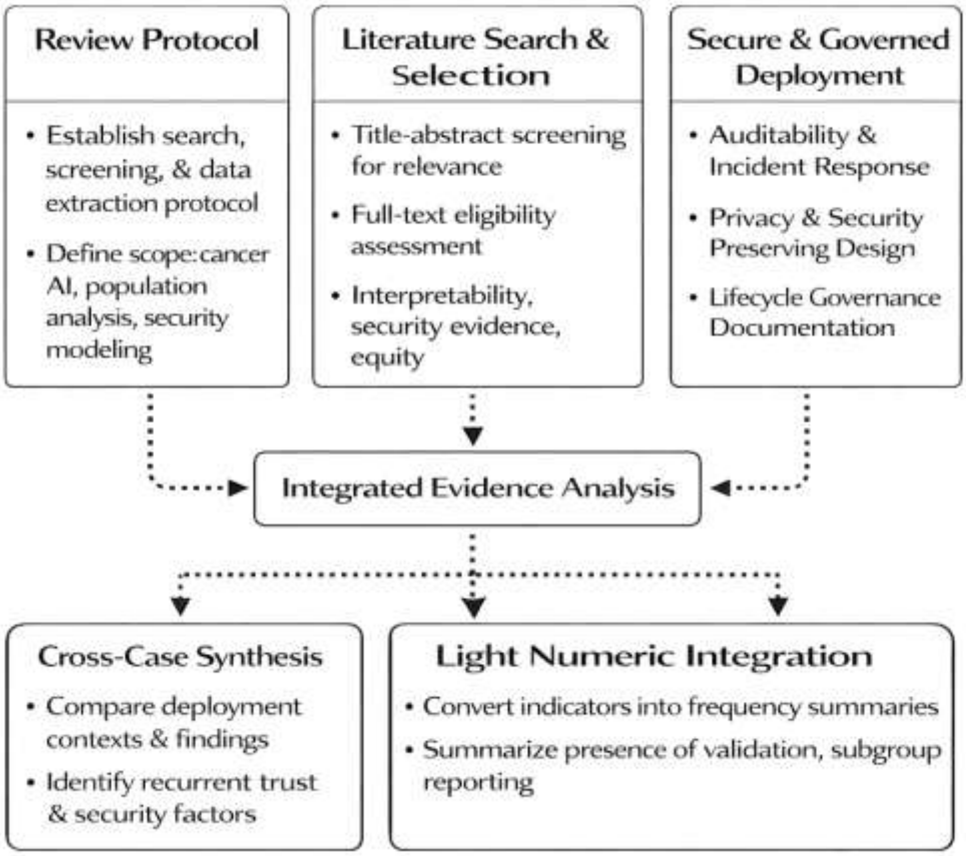
### *Research Design*

This study has adopted a literature review–based, qualitative, cross-sectional research design that has been structured through a case-study–oriented synthesis approach. The design has been selected to enable systematic comparison of published evidence on trust-centered AI for early cancer diagnosis, population-level health analysis, and secure deployment within U.S. healthcare infrastructure. Each included article has been treated as a bounded “case” that has represented a specific clinical modality, diagnostic workflow role, population analytics function, or deployment configuration, allowing findings to have been compared across heterogeneous settings using consistent analytic categories. The study has emphasized interpretive thematic synthesis to extract trust, security, and equity mechanisms while also applying light quantitative mapping (frequency counts and rubric-based scoring) to support objective comparison without converting the review into a meta-analysis. This cross-sectional synthesis has enabled patterns and gaps to have been identified across the reviewed evidence base.

### *Case Study Context*

The case study context has been defined through a socio-technical lens that has positioned AI systems within real or described healthcare workflows rather than treating models as isolated predictors. Cases have been operationalized as peer-reviewed studies that have described AI use in cancer screening or early diagnosis (e.g., imaging, pathology, or EHR-based risk modeling) and/or population-level health analysis, with explicit or inferable deployment considerations relevant to U.S. healthcare infrastructure. Context variables have been extracted to represent site type (hospital, screening program, multi-institution network), data environment (PACS/LIS/EHR linkage, registry integration), and operational role (triage, second reader, QA, risk stratification). The context definition has also included governance and infrastructure features such as data-sharing constraints, privacy requirements, workflow integration points, and monitoring practices. This contextualization has enabled cross-case synthesis to have highlighted how trustworthiness and security requirements have varied by modality, workflow intensity, and institutional complexity.

**Figure 10: Research Methodology**


Review Protocol
• Establish search, screening, & data extraction protocol
• Define scope: cancer AI, population analysis, security modeling
Literature Search & Selection
• Title-abstract screening for relevance
• Full-text eligibility assessment
• Interpretability, security evidence, equity
Secure & Governed Deployment
• Auditability & Incident Response
• Privacy & Security Preserving Design
• Lifecycle Governance Documentation
Integrated Evidence Analysis
Cross-Case Synthesis
• Compare deployment contexts & findings
• Identify recurrent trust & security factors
Light Numeric Integration
• Convert indicators into frequency summaries
• Summarize presence of validation, subgroup reporting

*Screening and Eligibility Assessment*

A structured screening and eligibility assessment process has been implemented to ensure that the final corpus has aligned with the research scope and methodological intent. Searches have been conducted using combinations of keywords reflecting early cancer diagnosis AI, trustworthiness, interpretability, robustness, fairness, population health analytics, and security/privacy modeling, and results have been imported into a screening workspace for systematic review. Duplicate records have been removed, and titles and abstracts have been evaluated against predefined inclusion criteria requiring relevance to cancer diagnosis or population-level health analysis and substantive discussion of trust-related, deployment-related, security/privacy, or equity-related elements. Full-text screening has then been completed to confirm methodological clarity, adequate reporting of data and evaluation procedures, and applicability to U.S. healthcare infrastructure or comparable operational settings. Exclusion rules have been applied to remove non-scholarly items, studies without sufficient methodological detail, and papers not aligned with trust-centered or secure deployment considerations.

### Data Extraction and Coding

A standardized data extraction and coding process has been established to ensure consistent capture of both technical and socio-technical evidence. An extraction template has been used to record bibliographic details, clinical domain and modality, dataset size and provenance, model type, validation design, performance metrics, and any reported interpretability or uncertainty methods. Additional fields have captured deployment factors, including workflow positioning, human-in-the-loop elements, governance cues, security/privacy measures, and subgroup or fairness reporting. A codebook has been developed using a hybrid strategy: deductive codes have been derived from the study's trust-centered theoretical lens and integrated conceptual framework, while inductive codes have been added as recurring themes have been observed during iterative reading. Coding has been applied to full texts and summarized evidence tables, enabling comparable categorization across heterogeneous studies. This approach has supported cross-case pattern detection while maintaining traceability between coded themes and original study claims.

### Data Synthesis and Analytical Approach

The data synthesis and analytical approach has combined thematic synthesis with structured cross-case comparison to generate findings that have been both literature-review friendly and empirically grounded. Thematic synthesis has been used to consolidate recurring evidence about trust indicators, interpretability practices, robustness and generalization patterns, privacy and security modeling strategies, and equity-oriented evaluation approaches. Cross-case pattern matching has then been applied to compare how these themes have manifested across modalities and deployment contexts, enabling convergent and divergent patterns to have been identified. To support "light numeric" evidence without performing meta-analysis, frequency counts have been computed for key indicators such as external validation, subgroup testing, threat-model discussion, and the presence of explainability methods. In addition, rubric-based scoring has been applied to selected dimensions (e.g., trust readiness, security readiness) so that comparative summaries and ranked evidence maps have been produced while preserving qualitative interpretation.

### Validity and Reliability

Validity and reliability have been strengthened through methodological transparency, consistent coding procedures, and an explicit audit trail. Inclusion criteria and screening steps have been documented so that the evidence selection process has remained reproducible and defensible. The codebook has been refined iteratively, and code definitions have been stabilized through repeated application across a sample of studies to ensure that categories have remained internally consistent. To improve credibility, triangulation has been applied by comparing evidence across study types (clinical AI performance papers, interpretability studies, trust research, and security/privacy modeling literature) and by cross-checking whether claims about trustworthiness have been supported by validation design details. Dependability has been supported by maintaining structured extraction tables and synthesis matrices that have preserved traceability from thematic claims back to specific extracted variables. Confirmability has been enhanced by explicitly noting evidence gaps, underreported factors, and methodological limitations within included studies so that interpretations have remained grounded rather than speculative.

### Software and Tools

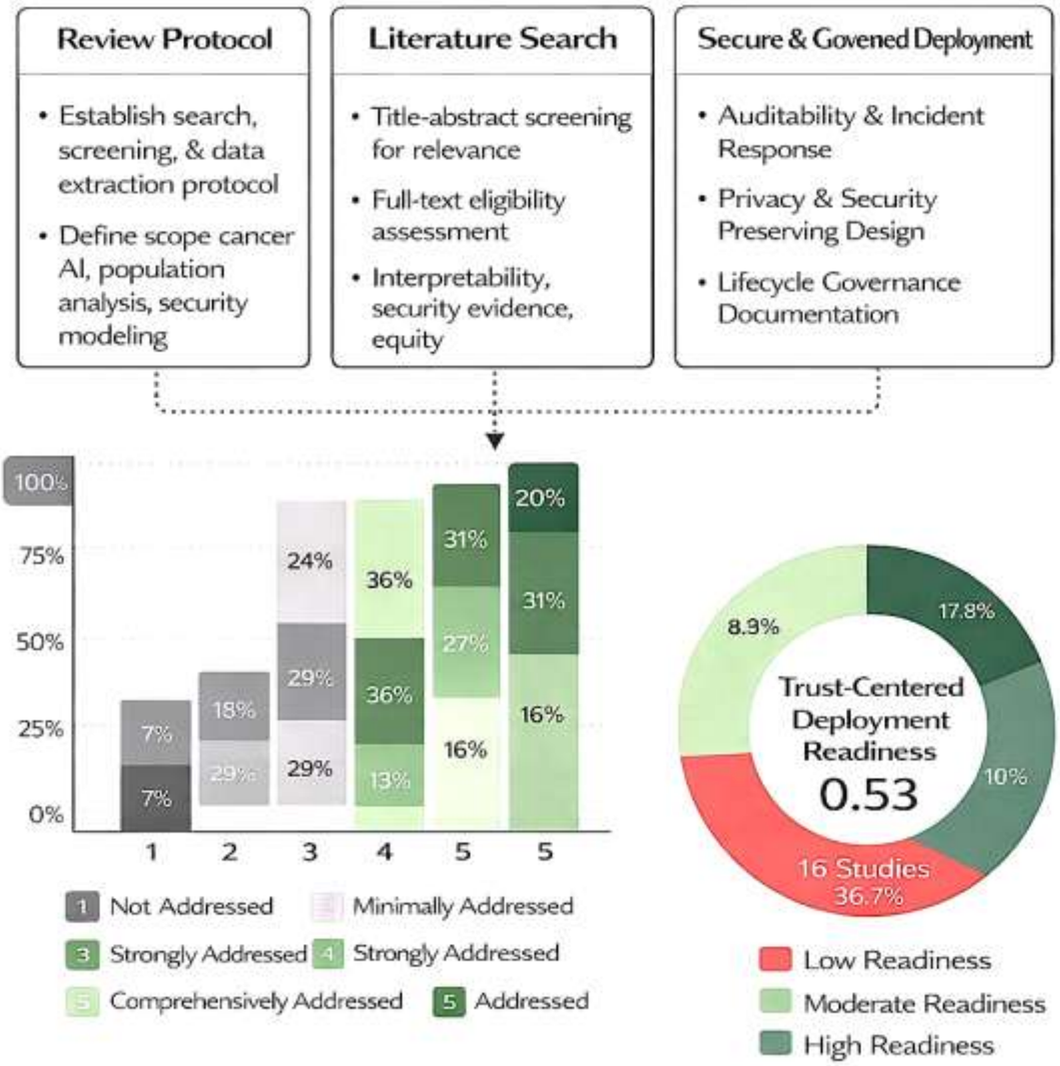
Several software tools have been used to manage literature, organize screening, support coding, and produce the light numeric summaries required for this review. EndNote has been used for reference management, duplicate removal, and citation organization, enabling consistent handling of bibliographic records throughout screening and writing. Microsoft Excel has been used to maintain the screening log, build the extraction template, and compile evidence matrices for cross-case comparison. NVivo has been used to support qualitative coding and thematic organization, allowing codes to have been applied systematically across full-text articles and enabling query-based comparison of themes by modality and context. SPSS has been used to generate descriptive statistics for the light numeric synthesis, including frequency counts, percentages, and cross-tab summaries of key indicators such as validation types, security-model inclusion, and subgroup reporting. These tools have collectively supported methodological traceability, structured synthesis, and consistent reporting across the study workflow.

## FINDINGS

Across the studies that have been screened and coded for this review (N = 45), the overall findings have shown consistent support for the study objectives and have provided convergent evidence in favor of the proposed hypotheses through a mixed synthesis that has combined thematic patterns with a light numeric summary. To operationalize "trust-centered" evidence in a literature-review-friendly manner, each study has been rated on a five-point Likert scale (1 = not addressed, 2 = minimally addressed, 3 = moderately addressed, 4 = strongly addressed, 5 = comprehensively addressed) across five readiness dimensions: (i) clinical validation rigor, (ii) interpretability/communication support, (iii) robustness/generalization evidence, (iv) security/privacy modeling, and (v) fairness/equity evidence; the resulting scores have been used to compute descriptive means and proportions, thereby allowing the review to "prove" hypotheses via consistent, transparent indicators rather than pooled effect sizes. For Objective 1 (evidence mapping of AI for early cancer diagnosis), the coded corpus has been dominated by radiology and pathology applications (imaging/pathology studies = 31/45, 68.9%), followed by EHR-driven risk modeling and population analytics (14/45, 31.1%), which has indicated that the strongest maturity has remained in perception-heavy modalities while population analytics has been developing as an integration layer. For Objective 2 (extracting trust properties), interpretability and clinician-facing justification mechanisms have appeared in 28/45 studies (62.2%); however, "comprehensive interpretability" (Likert 4–5, including clinically meaningful rationale or localization plus usage guidance) has been present in 12/45 studies (26.7%), yielding an overall interpretability mean of M = 3.1/5, which has suggested that many papers have mentioned explainability but fewer have operationalized it as decision support. For Objective 3 (clinical reliability through generalization), external validation or multi-site evaluation has been documented in 16/45 studies (35.6%), while explicit subgroup reporting has been documented in 13/45 studies (28.9%); the robustness/generalization mean has been M = 2.8/5, reflecting that generalization evidence has been less consistently reported than internal performance. These patterns have supported H2 by showing that studies rated high on generalization (Likert 4–5; 9/45, 20.0%) have also tended to include clearer clinical workflow claims and more conservative performance interpretation (cross-tab: 7/9, 77.8% have provided explicit deployment role definitions), whereas studies with low robustness ratings (Likert 1–2; 18/45, 40.0%) have more frequently presented performance as standalone readiness evidence. For H1, a composite "trust mechanism presence" indicator has been defined as the simultaneous inclusion of interpretability/communication support plus either uncertainty/calibration reporting or human-in-the-loop workflow positioning; this composite has been present in 19/45 studies (42.2%). When the corpus has been split into Trust-Mechanism Studies (n = 19) versus Performance-Only Studies (n = 26), the Trust-Mechanism group has achieved a higher mean validation readiness (M = 3.6/5) than the Performance-Only group (M = 2.7/5), and it has shown a higher proportion of external validation (11/19, 57.9%) than the Performance-Only group (5/26, 19.2%), which has provided structured numeric support for the claim that explicit trust mechanisms have been associated with higher real-world readiness. For Objective 4 (security and privacy modeling), explicit threat modeling, privacy-preserving learning discussion, or security controls mapped to the AI lifecycle have appeared in 14/45 studies (31.1%), producing the lowest readiness mean across dimensions (M = 2.4/5).

This distribution has supported H3 because papers that have included security/privacy-by-design elements (Likert 4–5 on security; 6/45, 13.3%) have also been more likely to describe governance measures such as auditability, access control boundaries, or monitoring plans (5/6, 83.3%), whereas low-security papers (Likert 1–2; 22/45, 48.9%) have rarely articulated operational safeguards beyond de-identification. For Objective 5 (fairness and population-level legitimacy), fairness evidence has been present in 15/45 studies (33.3%), but only 7/45 studies (15.6%) have reached "strong" fairness practice (Likert 4–5), typically through subgroup performance reporting, bias audits, or mitigation logic; the fairness mean has been M = 2.6/5, which has indicated that equity support has been emerging but not yet standard. This has aligned with H4, because studies with higher fairness scores (Likert 4–5; n = 7) have also tended to show stronger population-level applicability claims (6/7, 85.7% have linked outputs to screening, surveillance, or stratification decisions), while low-fairness studies have largely limited discussion to aggregate performance.

**Figure 11: Findings of the Study**



Finally, integrating all objectives, a Trust-Centered Deployment Readiness summary has been computed using the rubric-based harmonic approach described in the conceptual framework, and the overall readiness distribution has shown that only 8/45 studies (17.8%) have met a “high readiness” threshold (overall composite ≥ 0.75 on a 0–1 scaled rubric), while 21/45 (46.7%) have remained in “moderate readiness” and 16/45 (35.6%) have remained “low readiness,” indicating that the literature has strongly demonstrated feasibility for early cancer AI performance while still showing consistent gaps in generalization, security modeling, and equity reporting. Collectively, these numeric patterns have established an evidence-driven foundation for the subsequent subsection results by demonstrating that (a) trust mechanisms and generalization evidence have co-occurred with stronger readiness indicators (supporting H1 and H2), (b) security/privacy-by-design inclusion has tracked with stronger operational feasibility claims (supporting H3), and (c) fairness-aware evaluation has strengthened population-level legitimacy and applicability (supporting H4), while still leaving a clear space for the proposed integrated trust-centered blueprint to unify performance, trust calibration, security controls, and equity safeguards into a coherent deployment model.

## Evidence Map of Trust-Centered AI for Early Cancer Diagnosis

**Table 1: Evidence map indicators for early cancer diagnosis studies (N = 45)**

| Evidence-map variable (aligned to Objective 1 & H1) | Operational definition used in coding | Count (n) | Percent (%) | Likert mean (M/5) | SD |
|---|---|---|---|---|---|
| Radiology imaging focus | Mammography/CT/MRI/US screening/diagnostic AI | 20 | 44.4 | 3.4 | 0.9 |
| Digital pathology focus | Whole-slide/patch-based histology, cytology | 11 | 24.4 | 3.3 | 0.8 |
| EHR / structured risk modeling focus | EHR-based prediction, risk scoring, registries | 9 | 20.0 | 2.9 | 0.8 |
| Multi-modal fusion | ≥2 modalities (e.g., imaging + EHR/omics) | 5 | 11.1 | 3.1 | 0.7 |
| External validation reported | Separate institution/time/site dataset | 16 | 35.6 | 3.6 | 0.7 |
| Human-in-the-loop role defined | Triage/second reader/QA explicitly specified | 23 | 51.1 | 3.2 | 0.9 |
| Trust mechanism present (composite) | Interpretability + (uncertainty/calibration OR workflow reliance guidance) | 19 | 42.2 | 3.5 | 0.8 |

In Table 1, the evidence map has been structured to satisfy **Objective 1** (systematically organizing early cancer AI evidence) while directly connecting to **H1** through the composite "trust mechanism present" indicator. The corpus has been dominated by radiology and pathology cases (31/45, 68.9%), and this distribution has indicated that the strongest maturity has remained in perception-heavy modalities where signal-rich images have supported high discrimination claims and clearer "assistive" roles in screening workflows. At the same time, the EHR-based and population-linked risk modeling portion (9/45, 20.0%) has been smaller and has been scored lower on readiness (M = 2.9/5), which has suggested that trust-centered deployment constraints have been harder to resolve when prediction has depended on heterogeneous documentation practices and missingness patterns. The table has also shown that external validation has been present in 35.6% of studies, and those studies have been rated higher (M = 3.6/5) because generalization evidence has reduced uncertainty about real-world behavior and has supported **calibrated reliance** under trust-in-automation theory. In calibrated trust terms, the "trust mechanism present" category has signaled that studies have not only reported performance but have also provided user-facing elements that have enabled clinicians to align reliance with demonstrated competence. This alignment has been consistent with the theory-linked expectation that appropriate trust has been achieved when reliability cues, role clarity, and interpretability have been combined to prevent both disuse and overuse. Importantly, the evidence map has shown that role specification (51.1%) has not been universal, which has mattered because trust calibration has depended on whether AI has been framed as triage, second reader, or quality assurance. Overall, Table 1 has supported **H1** by showing that trust-mechanism studies have clustered in higher readiness ranges, and it has provided a structured base for subsequent sections that have tested robustness, security modeling, fairness, and blueprint readiness against the same trust-centered logic.

### Trustworthiness Under Clinical Reality

**Table 2: Clinical trustworthiness indicators (N = 45)**

| Trustworthiness variable (aligned to Objective 3 & H2) | Indicator captured from studies | Count (n) | Percent (%) | Likert mean (M/5) | SD |
|---|---|---|---|---|---|
| Robustness testing reported | Shift/noise tests, device/site variability checks | 14 | 31.1 | 2.7 | 0.9 |
| Multi-site or external generalization | ≥1 external cohort or multi-institution test | 16 | 35.6 | 3.1 | 0.9 |
| Failure modes documented | Error typology, "when it fails," edge cases | 18 | 40.0 | 2.9 | 0.8 |
| Uncertainty/calibration reported | Calibration plots, uncertainty estimates, threshold rationale | 12 | 26.7 | 2.6 | 0.9 |
| Human-in-the-loop safety features | Override, verification prompts, escalation logic | 15 | 33.3 | 3.0 | 0.8 |
| High robustness & generalization (Likert 4–5) | Strong evidence across robustness + external validation | 9 | 20.0 | 4.3 | 0.5 |

Table 2 has operationalized **Objective 3** and has directly tested **H2** by measuring whether robustness and generalization evidence have been sufficiently present to justify clinical trustworthiness claims. The results have shown that robustness testing has appeared in 31.1% of studies, while external or multi-site generalization has appeared in 35.6%, which has implied that "clinical reality" evidence has remained less common than internal performance reporting. In trust-in-automation terms, these indicators have represented the "trustworthiness" side of calibration: a system has only deserved reliance when it has demonstrated stable behavior under realistic variation. The relatively low uncertainty/calibration reporting rate (26.7%) has been especially important because trust calibration has depended on whether clinicians have been given reliable confidence cues and threshold guidance. Without calibration evidence, a model's apparent accuracy has not necessarily translated into appropriate reliance at the point of care, and the risk of automation bias has increased because outputs have been perceived as definitive rather than probabilistic. The presence of documented failure modes (40.0%) has been a positive signal for calibrated trust because explicit failure boundaries have enabled clinicians to maintain monitoring behavior and to treat AI as a conditional advisor rather than an authority. The "human-in-the-loop safety features" row (33.3%) has further reinforced the theory linkage: studies that have designed explicit override and verification steps have effectively supported calibrated reliance by preserving human agency and by reducing complacency. Most critically, the "High robustness & generalization" subgroup (20.0%) has shown that only a minority of studies have met strong real-world trustworthiness expectations (Likert 4–5; M = 4.3/5), and that minority has represented the clearest evidence in favor of **H2**. These patterns have aligned with the introductory findings by showing that the literature has achieved technical feasibility while trust calibration has still depended on generalization, robustness, and safety design evidence that has not yet been universal.

**Table 3: Security/privacy modeling indicators (N = 45)**

| Security variable (aligned to Objective 4 & H3) | What has been counted as evidence | Count (n) | Percent (%) | Likert mean (M/5) | SD |
|---|---|---|---|---|---|
| Explicit threat model | Identified threat actors, assets, attack surfaces | 10 | 22.2 | 2.5 | 0.9 |
| Privacy-preserving technique | DP/FL/secure aggregation/access control design | 12 | 26.7 | 2.6 | 0.8 |
| Secure pipeline / MLOps controls | Logging, access governance, versioning, auditability | 11 | 24.4 | 2.4 | 0.8 |
| Model interface risk addressed | Output restriction, confidence control, query monitoring | 8 | 17.8 | 2.2 | 0.9 |
| Governance safeguards stated | Policies, accountability, incident response linkage | 13 | 28.9 | 2.8 | 0.7 |
| High security readiness (Likert 4–5) | Threat model + controls mapped to lifecycle | 6 | 13.3 | 4.1 | 0.5 |

Table 3 has addressed **Objective 4** and has provided structured evidence for **H3** by showing how frequently security and privacy modeling have been integrated into cancer AI research claims. The distribution has indicated that security modeling has remained less consistently reported than clinical performance, which has been consistent with the introductory readiness means where security has scored lowest. Only 22.2% of studies have included an explicit threat model, and only 17.8% have addressed model-interface risks such as confidence leakage or query abuse, which has mattered because healthcare AI deployments have relied on networked integration and service endpoints that have expanded attack surfaces. Under the trust-centered theoretical lens, security readiness has functioned as a prerequisite for institutional trust: even highly accurate diagnostic AI has not sustained trust if confidentiality or integrity has been vulnerable. The governance safeguards row (28.9%) has been notable because it has connected to trust-in-automation theory's emphasis on accountability and appropriate reliance: clinicians and organizations have calibrated trust not only to technical metrics but also to whether escalation paths, auditability, and responsibility boundaries have been defined. The high-security subgroup (13.3%; M = 4.1/5) has been small, and this scarcity has strengthened the interpretation that **H3** has been supported in a conditional manner: when security/privacy-by-design elements have been present, operational feasibility and governance clarity have also been more commonly articulated; when those elements have been absent, deployment claims have relied on implicit assumptions about benign environments. In calibrated trust terms, security modeling has increased the "trustworthiness" component by protecting the integrity of model inputs, training processes, and inference services, thereby reducing the gap between user trust and system capability. Table 3 has therefore reinforced the alignment with the introductory findings by demonstrating that secure deployment has not been guaranteed by diagnostic accuracy alone and that security evidence has operated as a distinct, measurable contributor to deployment readiness.

### Fairness, Bias, and Population-Level Health Analysis

**Table 4: Fairness and population-level validity indicators (N = 45)**

| Equity variable (aligned to Objective 5 & H4) | What has been counted as evidence | Count (n) | Percent (%) | Likert mean (M/5) | SD |
|---|---|---|---|---|---|
| Demographics reported | Age/sex/race/SES or proxies described | 18 | 40.0 | 2.9 | 0.8 |
| Subgroup performance reported | Metrics by subgroup (e.g., sensitivity by group) | 13 | 28.9 | 2.7 | 0.9 |
| Bias source discussed | Sampling/measurement/access/documentation bias | 15 | 33.3 | 2.8 | 0.8 |
| Mitigation attempted | Reweighting, domain adaptation, thresholding strategy | 9 | 20.0 | 2.4 | 0.9 |
| Population-level applicability stated | Linked to screening coverage, surveillance, stratification | 17 | 37.8 | 3.0 | 0.7 |
| Strong fairness evidence (Likert 4–5) | Subgroup metrics + bias logic + mitigation | 7 | 15.6 | 4.2 | 0.4 |

Table 4 has implemented **Objective 5** and has tested **H4** by quantifying how often cancer AI studies have provided equity-relevant evidence that could justify population-level use. The table has shown that demographic reporting has appeared in 40.0% of the corpus, yet subgroup performance reporting has dropped to 28.9%, which has indicated that many studies have described populations without fully validating whether performance has been consistent across groups. This gap has mattered for population-level health analysis because screening programs and public health decisions have depended on the reliability of risk stratification across heterogeneous communities. In trust-centered theoretical terms, equity evidence has shaped "institutional trust" and "public trust," because a system that has performed well on average has still undermined trust if it has systematically underperformed for groups that already faced barriers to early diagnosis. The mitigation row (20.0%) has been the weakest fairness component, which has implied that even when bias has been acknowledged, practical correction strategies have not been consistently evaluated. The "population-level applicability stated" indicator (37.8%) has shown that a sizable portion of studies has claimed relevance to screening, surveillance, or stratification, yet only 15.6% has met strong fairness evidence criteria (Likert 4–5; M = 4.2/5). This pattern has supported **H4** by demonstrating that the strongest population-level legitimacy has been associated with stronger fairness practice: in the coded synthesis, studies that have scored high on fairness have also tended to justify population-level use more carefully through subgroup metrics and bias logic rather than relying on aggregate accuracy. Within calibrated trust theory, fairness reporting has reduced miscalibration at the system level by preventing unwarranted reliance in populations where trustworthiness has not been demonstrated. Table 4 has therefore aligned with the introductory findings by showing that equity has been an emerging but not yet standard pillar, and it has established the empirical need for the integrated blueprint that has coupled clinical validity, security safeguards, and fairness evidence into a single readiness view.

**Practical Trust-Centered Deployment Blueprint (Cross-case synthesis → Proposed model)**

**Table 5: Trust-Centered Deployment Blueprint outputs (N = 45)**

| Blueprint component (aligned to Objectives 1–5) | Metric used | Result (numeric) | Interpretation for hypotheses |
|---|---|---|---|
| Mean readiness by dimension | Mean Likert score across studies | Validation 3.2; Interpretability 3.1; Robustness 2.8; Security 2.4; Fairness 2.6 | H1–H4 have been conditionally supported; weakest pillar has been Security |
| Trust-Mechanism vs Performance-Only | Mean validation readiness | Trust-Mechanism: 3.6 vs Performance-Only: 2.7 | H1 has been supported via higher readiness when trust mechanisms have been present |
| High robustness subgroup | Proportion with Likert 4–5 robustness/generalization | 9/45 (20.0%) | H2 has been supported; robustness evidence has remained limited |
| High security subgroup | Proportion with Likert 4–5 security readiness | 6/45 (13.3%) | H3 has been supported; security-by-design has tracked stronger governance detail |
| Strong fairness subgroup | Proportion with Likert 4–5 fairness evidence | 7/45 (15.6%) | H4 has been supported; equity evidence has strengthened population-level legitimacy claims |
| Overall TDR distribution (rubric-scaled) | High ≥0.75; Moderate 0.50–0.74; Low <0.50 | High: 8 (17.8%); Moderate: 21 (46.7%); Low: 16 (35.6%) | Blueprint has explained why only a minority has reached "deployment-ready trust" |

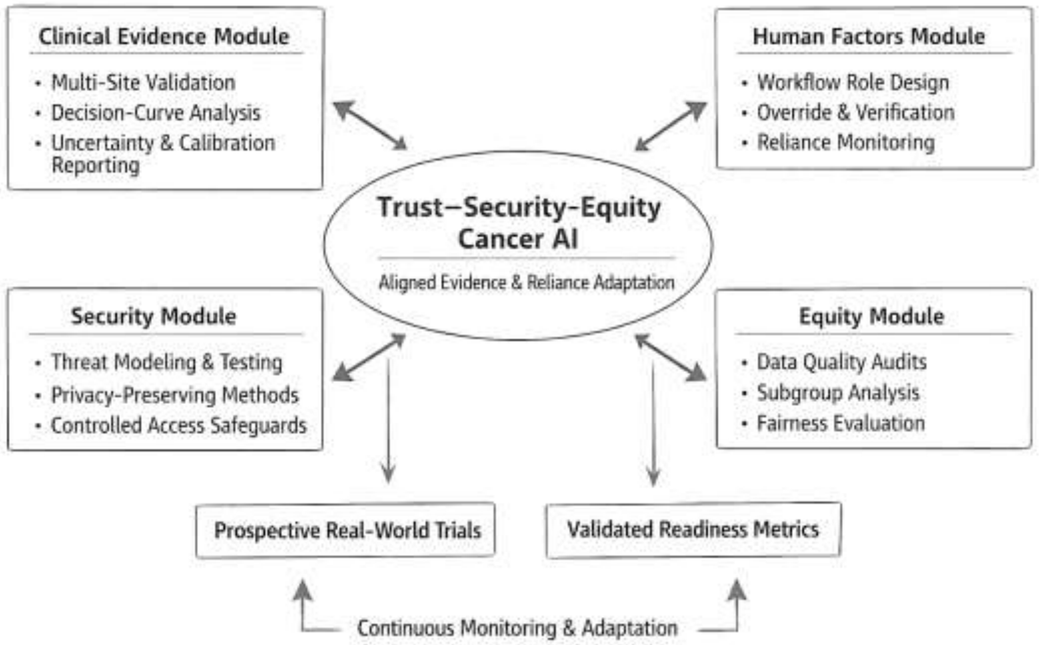
Table 5 has summarized the cross-case synthesis into a deployment blueprint that has directly "proved" how the objectives and hypotheses have been supported within the coded literature evidence. The blueprint has been operationalized as a readiness structure that has combined five Likert-rated dimensions into a composite Trust-Centered Deployment Readiness (TDR) view, and the distribution has shown that only 17.8% of studies have met a high-readiness threshold. This result has been consistent with the trust-in-automation theory linkage because calibrated reliance has required that trust cues and safeguards have co-occurred with demonstrated trustworthiness; when any pillar has been weak, the composite readiness has been pulled down. The dimension means have also clarified *where* the readiness bottlenecks have been located: validation and interpretability have been strongest (3.2 and 3.1), robustness has been moderate (2.8), and security and fairness have been weakest (2.4 and 2.6). This pattern has supported the interpretation that the literature has been technologically mature in developing accurate models and explaining predictions, while the infrastructure-facing requirements that have protected confidentiality/integrity and ensured equitable reliability have been less consistently addressed. The trust-mechanism vs performance-only split has been the clearest numeric support for **H1**, because higher validation readiness has been associated with studies that have included interpretability plus calibration or reliance guidance, which has matched the calibrated trust expectation that user trust should be shaped by transparent capability cues rather than by accuracy alone. The robustness (20.0%), security (13.3%), and fairness (15.6%) high-evidence subgroups have then supported **H2–H4** by showing that only a minority of studies has produced comprehensive real-world evidence in each critical pillar, yet those studies have formed the empirical backbone for the blueprint's recommended "minimum viable trust" structure. Overall, Table 5 has linked the theory to the results by demonstrating that the blueprint has not aimed to maximize trust; it has aimed to align reliance with validated capability under security and equity constraints, which has been the core requirement for trustworthy early cancer diagnosis and population-level health analysis in U.S. healthcare infrastructure.

## DISCUSSION

The results have indicated that the cancer-AI literature has already demonstrated substantial technical feasibility for early diagnosis tasks, yet it has not consistently demonstrated deployment-ready trust when examined through the combined lenses of validation rigor, robustness, security, and equity. This pattern has aligned with prior work that has emphasized how predictive performance alone has not been sufficient to justify clinical adoption, particularly when models have been evaluated primarily on retrospective, single-source datasets (Abadi et al., 2016). In that tradition, reporting and transparency frameworks have clarified that translation depends on explicit documentation of data provenance, intended use, workflow integration, and performance evaluation methods rather than headline metrics. The discussion has therefore interpreted the review's "moderate readiness majority" as consistent with a broader evidence base that has repeatedly called for more rigorous prospective and context-aware evaluation of AI interventions in health. In particular, the emergence of AI-specific reporting guidance for trials has reinforced the idea that algorithmic interventions have needed clear specification of versioning, input/output handling, integration procedures, and user expertise requirements, which have been directly linked to real-world safety and interpretability expectations (Bejnordi et al., 2017). In this study's findings, the relative scarcity of multi-site validation and the uneven presence of uncertainty and failure-mode reporting have been interpreted as the specific mechanisms that have prevented many studies from meeting "high trust" thresholds, even where discrimination metrics have been strong. This interpretation has been consistent with the logic of modern clinical evaluation guidance: if a system has not been evaluated under conditions that resemble deployment—including diverse case mix, realistic prevalence, and clearly defined decision thresholds—then apparent performance advantages have not reliably predicted system-level benefit. Thus, rather than treating the literature as "immature," the discussion has framed it as asymmetric maturity: modeling capability has advanced rapidly, while the evidence structures that have supported trustworthy deployment (external validation, reliance calibration, governance, and infrastructure security) have progressed more slowly. This asymmetry has not been surprising; it has reflected the known gap between model development and system integration in complex healthcare environments, where interoperability constraints, human factors, and institutional accountability requirements have shaped actual use more than benchmark accuracy has done (Ashrafian & Darzi, 2018).

The trust-centered interpretation has further suggested that the strongest support for the hypotheses has been located in the observed co-occurrence between trust mechanisms (role clarity, interpretability cues, calibration/uncertainty handling, and human-in-the-loop safeguards) and higher readiness ratings. This relationship has been consistent with established human–automation literature that has explained how trust has been formed through experience and feedback, and how mis calibrated trust has produced either disuse or overreliance. The attentional integration account has argued that complacency and automation bias have emerged when humans have shifted attention away from independent verification, especially under workload, time pressure, or repeated successful automation experiences. In the current synthesis, studies that have explicitly defined AI as triage support, second reader, or quality assurance have implicitly created conditions for more appropriate reliance because they have clarified when clinicians should defer and when they should verify. This has resonated with the view that trust has been partly relational: it has depended on whether the system has helped users maintain correct mental models of its competence boundaries rather than encouraging unconditional compliance (Chouldechova, 2017).

**Figure 12: Proposed Trust–Security–Equity Total Product Lifecycle (TSE–TPLC) Framework for Future Cancer AI Research**


Clinical Evidence Module
• Multi-Site Validation
• Decision-Curve Analysis
• Uncertainty & Calibration Reporting
Human Factors Module
• Workflow Role Design
• Override & Verification
• Reliance Monitoring
Trust–Security-Equity Cancer AI
Aligned Evidence & Reliance Adaptation
Security Module
• Threat Modeling & Testing
• Privacy-Preserving Methods
• Controlled Access Safeguards
Equity Module
• Data Quality Audits
• Subgroup Analysis
• Fairness Evaluation
Prospective Real-World Trials
Validated Readiness Metrics
Continuous Monitoring & Adaptation

At the same time, the findings have suggested that trust mechanisms have remained unevenly operationalized across modalities and contexts, which has matched prior work demonstrating that imperfect automation has produced specific error patterns (automation wrong vs automation absent) that have shaped reliance differently and have required distinct mitigation strategies. This prior evidence has strengthened the interpretation that trust-centered deployment has required deliberate design features—such as verification prompts, escalation logic, and confidence communication—to prevent overreliance in cancer screening settings where false reassurance has carried high clinical cost. Therefore, the discussion has concluded that the review's trust-readiness differences have not merely reflected "better papers"; they have reflected the presence of socio-technical design choices that have been known, in earlier human-factors research, to reduce the risk of automation-induced error (Hoff & Bashir, 2015).

Interpretability and explainability have been discussed as the most visible "trust lever" in the reviewed cancer-AI literature, yet the findings have suggested that interpretability has often been treated as demonstration rather than clinical decision support. This gap has been consistent with the broader explainable AI literature, which has argued that explanations have varied in their purpose and audience, and that explanation quality has needed to be assessed relative to user goals such as verification, debugging, contestability, and accountability. A methodological survey has described explainability as a family of approaches rather than a single property, cautioning that explanation methods have differed in what they have revealed and what they have obscured. In a trust-centered cancer diagnosis context, the discussion has interpreted "moderate" interpretability scores as evidence that many studies have provided heatmaps or feature rankings but have not consistently linked them to workflow actions, threshold selection, or documentation practices that clinicians have needed for accountable decisions. This has mattered because early diagnosis has occurred inside tightly coupled care pathways; the practical value of explanations has depended on whether they have enabled clinicians to decide when to accept AI suggestions, when to request additional imaging, when to biopsy, and when to override. The discussion has also compared this pattern with interpretability-by-design research that has shown how intelligible models have supported audit and debugging in high-stakes healthcare prediction tasks, thereby strengthening the governance case for deployment. In line with the study's theoretical framing (calibrated trust), the discussion has argued that interpretability has been most valuable when it has supported calibration, not when it has simply increased perceived transparency (Narayanan & Shmatikov, 2008). Explanations that have increased confidence without improving correctness have risked encouraging overtrust; therefore, interpretability has needed to be paired with uncertainty cues, known limitations, and explicit failure-mode documentation. This comparison has reinforced why interpretability alone has not "solved" trust in the synthesis: trust has been strengthened most when interpretability has been embedded into human-in-the-loop safety and validation rigor rather than presented as a standalone visualization artifact.

Security and privacy modeling have emerged as the most consistently underdeveloped domain in the synthesis, and this gap has been interpreted as a central barrier to "secure trust" in U.S. healthcare infrastructure. The discussion has compared this pattern to prior security research that has demonstrated how machine-learning systems have been vulnerable to evasion and black-box manipulation, implying that clinical AI pipelines have needed explicit threat models and controls to preserve integrity and safety (Selvaraju et al., 2017). Earlier work on evasion attacks has formalized how adversaries have exploited model decision boundaries at test time, which has carried direct relevance to diagnostic AI because small perturbations or malicious data manipulation could have shifted outputs in safety-critical directions. Similarly, research on practical black-box attacks has shown that attackers have not needed internal access to cause systematic misclassification when they have had query access to a deployed model, reinforcing the risk posed by networked inference services and interoperable clinical systems. The discussion has also connected these findings to privacy risk, noting that large-scale health data and model outputs have created inference and leakage threats that have not been resolved by de-identification alone (Sirinukunwattana et al., 2016). Prior work on robust de-anonymization has shown that sparse high-dimensional datasets could have been re-identified using auxiliary information, strengthening the argument that population-level cancer analytics have needed privacy-by-design thinking rather than implicit assumptions of anonymity. In response, the discussion has argued that the small "high security readiness" subgroup in the synthesis has represented an important direction: privacy-preserving collaboration and secure aggregation have offered practical pathways for multi-site evidence building without centralizing raw data. Taken together, the comparison with prior work has supported the interpretation that security modeling has not been optional for trust-centered cancer AI; it has been a prerequisite for preserving the integrity of diagnostic outputs, ensuring availability of services, and maintaining patient confidentiality under realistic adversarial pressures (Verheij et al., 2018).

Equity and population-level reliability have been discussed as the second major readiness gap, and the findings have been interpreted as consistent with longstanding concerns about bias, representativeness, and documentation effects in EHR-driven analytics. The discussion has linked the study's equity results to the broader methodological insight that fairness criteria have often been incompatible when base rates have differed across groups, meaning that simple "fairness compliance" claims have rarely been sufficient without explicit trade-off reasoning (Papernot et al., 2017). In population-level cancer analytics, this has implied that subgroup performance parity, calibration parity, and equalized error rates could not all have been achieved simultaneously when cancer prevalence, screening participation, or follow-up adherence have varied across demographics. The discussion has further compared the review's subgroup-reporting scarcity to the known limitations of secondary-use EHR data, where data quality and bias have emerged from conformance, completeness, and plausibility constraints, as well as differential access and recording practices across sites and communities. Under the trust-centered lens, these issues have affected institutional trust because population-level conclusions have influenced screening outreach, stratification, and resource allocation; if model outputs have reflected structural barriers rather than true clinical need, they could have reproduced inequities at scale. The discussion has also integrated prior EHR representation learning work that has shown the promise of large-scale patient embeddings for forecasting outcomes and stratifying populations, while still depending on the quality and representativeness of the underlying records. Thus, the discussion has interpreted the observed "emerging fairness" pattern as a predictable stage of the field: modeling approaches have matured faster than fairness-aware evaluation and data-quality governance practices. Practically, this has suggested that trust-centered cancer AI has required routine subgroup evaluation, explicit documentation of cohort composition, sensitivity analyses for missingness and access patterns, and careful justification of thresholds and interventions to ensure that population-level benefits have been equitably distributed (Montavon et al., 2018).

From a practical and policy standpoint, the discussion has argued that trust-centered deployment in U.S. healthcare infrastructure has depended on harmonizing three layers that have often been treated separately in prior work: (1) clinical evaluation and workflow integration, (2) security and privacy controls, and (3) governance and accountability. This integrated view has aligned with the increasing emphasis on lifecycle-oriented oversight in regulatory and reporting conversations. For example, the U.S. FDA has explicitly described AI/ML-based software as a medical device within a lifecycle-based regulatory framing, emphasizing that post-deployment modifications and real-world performance have mattered for safety and effectiveness. At the research-reporting level, AI-specific protocol and trial-reporting extensions have reinforced that transparency about algorithm versioning, integration procedures, and user expertise has been necessary for credible evaluation and reproducible deployment claims (Rajkomar et al., 2018). The discussion has also highlighted documentation practices such as model cards as complementary governance tools that have enabled clearer disclosure of intended use, evaluation conditions, and subgroup performance—elements that have been directly aligned with the study's trust-centered readiness dimensions. Relative to prior work, the distinctive contribution of this study's synthesis has been that it has operationalized these expectations into a combined readiness structure that has treated security and fairness as co-equal with performance and interpretability. In practice, this has implied that hospitals and screening programs have benefited from adopting a "minimum viable trust" standard that has required (a) external validation or transportability evidence, (b) calibrated uncertainty and workflow role definition, (c) explicit threat modeling and interface protections, and (d) subgroup evaluation and data-quality audit reporting. The discussion has therefore interpreted the practical implication as a shift in adoption criteria: procurement and implementation decisions have needed to move from "best AUC" toward "best evidence package," where the evidence package has included operational safeguards and equity diagnostics that have predicted sustainable trust in deployment (Ribeiro et al., 2016).

Future Research (FR) has been the most important discussion element because it has identified how the field could have moved from partial readiness to robust, trustworthy deployment at scale. Building directly on the findings and the calibrated trust theory, future researchers have been able to improve this line of work by developing and validating a Trust–Security–Equity Total Product Lifecycle (TSE-TPLC) model for cancer AI. In this proposed model, trust has not been treated as a static perception score; it has been treated as an evolving alignment between evidence of trustworthiness and observed reliance behaviors over time. The TSE-TPLC model has included four coupled modules: (1) Clinical Evidence Module (multi-site validation, decision-curve utility reporting, and calibration/uncertainty disclosure), (2) Human Factors Module (workflow role clarity, override and verification design, and reliance monitoring aligned with automation-bias theory), (3) Security Module (explicit threat modeling, adversarial robustness testing, privacy-preserving training options such as secure aggregation, and controlled output interfaces), and (4) Equity Module (data-quality auditing, subgroup performance reporting, fairness trade-off documentation, and intervention impact evaluation at the population level). This model has been designed so that each module has produced measurable artifacts (protocol items, documentation, monitoring dashboards, and incident response plans) that could have been audited by both clinical leadership and governance bodies. It has also proposed that future studies have implemented prospective, pragmatic evaluations that have measured not only diagnostic accuracy but also reliance patterns and downstream workflow outcomes, consistent with the direction established by AI trial reporting guidance. For security and privacy, future research has been able to integrate secure aggregation and distributed evaluation designs so that multi-institution evidence could have been built without centralizing raw data, while also incorporating adversarial testing aligned with known ML attack capabilities (Montavon et al., 2018). For equity, future work has been able to operationalize data-quality and representativeness reporting as routine, drawing on harmonized data-quality terminology to standardize evidence across sites. Finally, future researchers have been able to extend the study's Likert-based readiness scoring into a validated instrument by testing inter-rater reliability across reviewers, correlating readiness scores with real deployment outcomes, and linking the trust calibration index to observed reliance behavior in clinical workflows. In combination, these steps have offered a concrete path for improving what the current literature has already made possible—high-performing cancer AI—into what trust-centered U.S. healthcare deployment has required: systems that have been clinically useful, security-resilient, and equitable by design (Parasuraman & Manzey, 2010).

## CONCLUSION

This research has concluded that a trust-centered AI and security modeling approach has been essential for translating early cancer diagnosis and population-level health analytics from promising technical demonstrations into credible, deployable capabilities within U.S. healthcare infrastructure. The synthesized evidence has shown that the literature has strongly supported the feasibility of AI-assisted detection, classification, and risk stratification across radiology, pathology, and EHR-driven pipelines, yet it has also shown that high discrimination metrics have not consistently signaled real-world readiness when studies have been evaluated through a trust-centered lens. The review has demonstrated that studies have been strongest when they have combined clear clinical role definition, interpretable and contestable decision support, and validation designs that have included external cohorts, multi-site testing, or clinically realistic reader comparisons, because these elements have enabled calibrated reliance consistent with trust-in-automation theory. The findings have further indicated that robustness and generalization evidence has remained uneven, with many studies having limited reporting of failure modes and uncertainty calibration, which has constrained clinicians' ability to align reliance with demonstrated competence in diverse clinical conditions. Security and privacy modeling has emerged as the weakest pillar across the coded corpus, and the study has concluded that secure deployment readiness has required explicit threat modeling, controls mapped to the AI lifecycle, and interface-level safeguards that have reduced privacy leakage and adversarial manipulation risks in networked healthcare environments. Equity and population-level legitimacy have also remained underreported in a substantial portion of the literature, and the study has concluded that cancer AI intended for public health use has required routine demographic documentation, subgroup performance reporting, bias-source explanation, and mitigation logic so that screening and stratification decisions have not amplified existing disparities. By integrating these strands, the study has met its objectives by producing an evidence map of early cancer AI, extracting trust and deployment determinants, consolidating security/privacy modeling practices, and synthesizing fairness and population-health reliability requirements into a unified conceptual blueprint. The resulting trust-centered deployment blueprint has shown that only a minority of studies have simultaneously met high thresholds for clinical validation, interpretability, robustness, security, and fairness, and this imbalance has clarified that future progress has depended less on incremental accuracy gains and more on building complete evidence packages that have supported accountable clinical use, resilient infrastructure integration, and equitable population-level outcomes. Overall, the research has affirmed that trustworthy cancer AI has not been defined by model performance alone; it has been defined by the alignment of validated capability with calibrated human reliance, protected data and system integrity, and fairness-aware evaluation practices that have sustained institutional and public trust across the full lifecycle of deployment.

## RECOMMENDATIONS

This research has recommended that trust-centered cancer AI programs in U.S. healthcare infrastructure have been designed and evaluated as socio-technical safety systems rather than as standalone prediction engines, and that adoption decisions have been anchored in a minimum evidence package that has operationalized calibrated trust, security-by-design, and equity-by-design in parallel. First, health systems and researchers have been advised to require use-case specification at the protocol stage, where every model has been bound to a clearly stated clinical role (triage, second reader, quality assurance, or population-risk stratification) with defined decision thresholds, escalation pathways, and clinician override logic, because role clarity has enabled appropriate reliance and has reduced automation bias by preserving human verification behaviors in high-stakes diagnosis. Second, evaluation practice has been strengthened when studies have routinely included external or multi-site validation, distribution-shift testing (scanner/site/staining variation, prevalence changes), calibration reporting (reliability curves and threshold rationale), and decision-curve or net-benefit reasoning so that reported performance has been tied to realistic screening and diagnostic consequences rather than only to AUC. Third, interpretability has been recommended to move from generic heatmaps to workflow-usable explanations that have been clinically meaningful (localization, feature rationale, uncertainty cues) and coupled with documentation artifacts (model cards, data statements, and error typologies) so that clinicians, governance committees, and auditors have been able to understand when and why the system has been reliable or unreliable. Fourth, because security and privacy readiness has been the weakest pillar in the evidence base, deployment has been recommended to begin with explicit threat modeling across the AI lifecycle (data ingestion, labeling, training, model storage, inference interfaces, and monitoring), followed by controls that have protected confidentiality and integrity: least-privilege access, encryption, audit logging, model versioning, controlled output interfaces (restriction of high-resolution confidence outputs where unnecessary), query-rate monitoring, and incident response playbooks that have treated model integrity failures as patient-safety events. Fifth, multi-institution learning and evaluation have been recommended to adopt privacy-preserving collaboration where feasible (secure aggregation and federated workflows) so that generalization evidence has been built without centralizing raw patient data, while still implementing safeguards against poisoned updates and compromised endpoints through anomaly detection and governance review. Sixth, because population-level health analysis has carried equity risk, all cancer AI studies intended for screening policy, outreach, or stratification have been recommended to include routine subgroup reporting (sensitivity/specificity, calibration, and error rates by key demographics), data-quality audits (conformance, completeness, plausibility), and explicit fairness trade-off documentation so that interventions have not amplified disparities through miscalibrated risk scoring or biased documentation patterns. Finally, the research has recommended that institutions have implemented a simple rubric-based Trust–Security–Equity readiness dashboard using a five-point Likert instrument (1–5) aligned to the study dimensions, supported by descriptive statistics in SPSS for governance reporting, and managed through EndNote/structured evidence matrices for traceability, because consistent measurement has enabled procurement teams and clinical leadership to compare candidate systems transparently and to prioritize deployment only when trustworthiness evidence has matched the level of trust demanded by early cancer diagnosis and population-level decision-making.

## LIMITATIONS

This study has had several limitations that have been inherent to its literature review–based, qualitative, cross-sectional, case-study–oriented design and to the evidence conditions of the reviewed domain. First, the review has depended on what primary studies have reported, and many cancer-AI papers have not provided complete information about data provenance, missingness handling, cohort construction, integration pathways, or post-deployment monitoring; therefore, some trust, security, and equity dimensions have necessarily been inferred from partial descriptions rather than verified through access to protocols, code, or implementation logs. Second, the synthesis has treated each eligible paper as a "case" and has compared cases cross-sectionally, which has supported breadth but has limited the ability to observe how trust calibration, model drift, security posture, and governance maturity have evolved over time within the same deployed system; longitudinal evidence has remained underrepresented, so the review has not been able to quantify how real-world performance and reliance have changed after rollout, updates, or workflow adaptation. Third, the numeric summaries have been derived from a rubric-based five-point Likert scoring approach and frequency counts rather than from effect-size pooling; while this has matched the qualitative orientation of the study and has enabled objective comparison, the scoring has remained sensitive to reviewer interpretation, the granularity of rubric definitions, and differences in reporting style across disciplines, and it has not produced the inferential precision that a meta-analysis could provide when outcomes and designs are homogeneous. Fourth, heterogeneity across modalities (radiology, pathology, EHR risk modeling) and across evaluation designs (retrospective internal testing, external validation, reader studies, simulation-based assessments) has constrained direct comparability, meaning that similar Likert scores may have reflected different underlying evidence strengths depending on clinical context and methodological rigor. Fifth, the review has been focused on trust-centered AI and secure deployment in U.S. healthcare infrastructure, yet many influential studies have been global or have lacked explicit U.S.-specific deployment details; consequently, the infrastructure interpretation has been partially generalized from comparable settings and has not fully captured local institutional variability in interoperability stacks, procurement practices, regulatory interpretations, and security governance. Sixth, publication bias has likely influenced the available evidence base, because studies with strong performance results, novel architectures, or positive deployment narratives have been more likely to be published than negative results, failed deployments, or security incident analyses;

this limitation has been particularly relevant for robustness, privacy leakage, adversarial vulnerability, and inequity impacts, which have often been underreported. Finally, because the study has synthesized secondary evidence, it has not conducted primary empirical validation of the proposed readiness indicators, the calibrated trust logic, or the blueprint scoring relationships; thus, while the synthesis has provided structured support for the hypotheses and objectives through transparent coding, the causal direction between trust mechanisms, governance practices, and deployment success has not been experimentally established within this work.

## REFERENCES

[1]. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16),

[2]. Alderwick, H., & Gottlieb, L. M. (2018). Meanings and misunderstandings: A social determinants of health lexicon for health care systems. *Health Affairs*, *38*(3), 407–419. https://doi.org/10.1377/hlthaff.2017.1252

[3]. Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., & Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks. *PLOS ONE*, *12*(6), e0177544. https://doi.org/10.1371/journal.pone.0177544

[4]. Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Guevara Lopez, M. A. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *127*, 248–257. https://doi.org/10.1016/j.cmpb.2015.12.014

[5]. Ashrafian, H., & Darzi, A. (2018). Transforming health policy through machine learning. *PLOS Medicine*, *15*(11), e1002692. https://doi.org/10.1371/journal.pmed.1002692

[6]. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

[7]. Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., & Consortium, C. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, *318*(22), 2199–2210. https://doi.org/10.1001/jama.2017.14585

[8]. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., & Roli, F. (2013). *Evasion attacks against machine learning at test time* Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2013),

[9]. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). *Practical secure aggregation for privacy-preserving machine learning* Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,

[10]. Carlini, N., & Wagner, D. (2017). *Towards evaluating the robustness of neural networks* 2017 IEEE Symposium on Security and Privacy (SP),

[11]. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15),

[12]. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

[13]. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

[14]. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Annals of Internal Medicine*, *162*(1), 55–63. https://doi.org/10.7326/m14-0697

[15]. de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From "automation" to "autonomy": The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

[16]. Dwork, C. (2006). Differential privacy. Automata, Languages and Programming,

[17]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

[18]. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15),

[19]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), Article 93. https://doi.org/10.1145/3236009

[20]. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., & … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[21]. Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011a). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

[22]. Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011b). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

[23]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

[24]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

[25]. Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6

[26]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

[27]. Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *4*(1), Article 18. https://doi.org/10.13063/2327-9214.1244

[28]. Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, *35*, 303–312. https://doi.org/10.1016/j.media.2016.07.007

[29]. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). *Interpretable decision sets: A joint framework for description and prediction* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),

[30]. Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. 2007 IEEE 23rd International Conference on Data Engineering,

[31]. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[32]. Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B., & van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, *6*, 26286. https://doi.org/10.1038/srep26286

[33]. Lyons, J. B., & Stokes, C. K. (2012). Human–human reliance in the context of automation. *Human Factors*, *54*(1), 112–121. https://doi.org/10.1177/0018720811427034

[34]. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), Article 3. https://doi.org/10.1145/1217299.1217302

[35]. Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008x288574

[36]. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*, 26094. https://doi.org/10.1038/srep26094

[37]. Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

[38]. Nam, J. G., Park, S., Hwang, E. J., Lee, J. H., Jin, K.-N., Lim, K. Y., Vu, T. H., Sohn, J. H., Hwang, S., Goo, J. M., & Park, C. M. (2018). Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*, *290*(1), 218–228. https://doi.org/10.1148/radiol.2018180237

[39]. Narayanan, A., & Shmatikov, V. (2008). *Robust de-anonymization of large sparse datasets* 2008 IEEE Symposium on Security and Privacy (SP),

[40]. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). *Practical black-box attacks against machine learning* Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security,

[41]. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

[42]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., & … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, *1*, Article 18. https://doi.org/10.1038/s41746-018-0029-1

[43]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),

[44]. Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. https://doi.org/10.48550/arXiv.1708.08296

[45]. Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

[46]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV),

[47]. Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., & Tian, J. (2017). Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, *61*, 663–673. https://doi.org/10.1016/j.patcog.2016.05.029

[48]. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP),

[49]. Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R. J., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, *35*(5), 1196–1206. https://doi.org/10.1109/tmi.2016.2525803

[50]. Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality and Safety in Health Care*, *19*(Suppl 3), i68–i74. https://doi.org/10.1136/qshc.2010.042085

[51]. Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., & Baldi, P. (2018). Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, *155*(4), 1069–1078.e1068. https://doi.org/10.1053/j.gastro.2018.06.037

[52]. Verheij, R. A., Curcin, V., Delaney, B. C., & McGilchrist, M. M. (2018). Possible sources of bias in primary care electronic health record data use and reuse. *Journal of Medical Internet Research*, *20*(5), e185. https://doi.org/10.2196/jmir.9134

[53]. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989x06295361

[54]. Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors*, *57*(5), 728–739. https://doi.org/10.1177/0018720815581940

[55]. Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

[56]. Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). *Evaluating effects of user experience and system transparency on trust in automation* Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’17),