



International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

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

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

ARTICLE TITLE	FROM PIXELS TO PROMPTS: A STRUCTURED NARRATIVE REVIEW OF SECURITY, PRIVACY, AND ADVERSARIAL ROBUSTNESS IN HEALTHCARE AI
----------------------	---

DOI	https://doi.org/10.31435/ijitss.4(52).2026.6340
------------	---

RECEIVED	15 July 2026
-----------------	--------------

ACCEPTED	09 October 2026
-----------------	-----------------

PUBLISHED	15 October 2026
------------------	-----------------

LICENSE



The article is licensed under a **Creative Commons Attribution 4.0 International License**.

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

FROM PIXELS TO PROMPTS: A STRUCTURED NARRATIVE REVIEW OF SECURITY, PRIVACY, AND ADVERSARIAL ROBUSTNESS IN HEALTHCARE AI

Patryk Sielaff

*Medical University of Warsaw, ul. Żwirki i Wigury 61, 02-091 Warsaw, Poland
ORCID ID: 0009-0005-8566-4715*

Artur Brudek

*Collegium Medicum, The Jan Kochanowski University, Kielce, 25-317, Poland
ORCID ID: 0009-0001-3984-5196*

Adrian Nowakowski

*Medical University of Warsaw, ul. Żwirki i Wigury 61, 02-091 Warsaw, Poland
ORCID ID: 0009-0004-7247-4464*

Hubert Piwar (Corresponding Author, Email: hubertpiwar@gmail.com)

*Medical University of Warsaw, ul. Żwirki i Wigury 61, 02-091 Warsaw, Poland
ORCID ID: 0009-0009-4333-9518*

Jan Pawłasek

*Medical University of Warsaw, ul. Żwirki i Wigury 61, 02-091 Warsaw, Poland
ORCID ID: 0009-0005-9244-8586*

Monika Lis

*Mazovian "Bródnowski" Hospital, Ludwika Kondratowicza 8, 03-242 Warsaw, Poland
ORCID ID: 0000-0003-4021-919X*

Karolina Hernik

*Medical University of Warsaw, ul. Żwirki i Wigury 61, 02-091 Warsaw, Poland
ORCID ID: 0009-0008-1097-8957*

Mateusz Sucharski

*Military Institute of Medicine, ul. Szaserów 128, 04-141 Warsaw, Poland
ORCID ID: 0009-0008-6460-2269*

Ewa Kacynel

*Medical University of Warsaw, Żwirki i Wigury 61, 02-091 Warsaw, Poland
ORCID ID: 0009-0002-8290-9818*

ABSTRACT

Artificial intelligence (AI) is increasingly used in healthcare, where evaluations usually ask whether a system is accurate on ordinary clinical data. This review asks instead what can happen when someone deliberately changes the data, instructions, or training to make a system fail, behave in hidden ways, or reveal confidential information, a field known as adversarial machine learning.

We reviewed 113 publications (2018–2026) and grouped the reported threats into six categories: altering an input after training; corrupting the training examples; teaching a hidden rule triggered later; using the model to learn private information; placing malicious instructions in language or image-and-language models; and attacking models trained jointly across hospitals or devices.

Laboratory studies showed that small deliberate changes to medical images could alter a diagnosis or a model's outline, corrupted training data could change what a model learned, and malicious prompts could produce unsafe medical responses. Effect sizes varied greatly across models, datasets, attacks, and test conditions. Several defenses helped in individual experiments, but no single defense was shown to work across all threats and settings; in one imaging study, attacks tailored to the tested defenses bypassed all five.

The studies do not show how often these attacks occur in practice or harm patients; their value is to demonstrate possible failure modes under controlled conditions. Medical AI should therefore be evaluated both on ordinary clinical data and in clearly described, clinically plausible attack scenarios.

KEYWORDS

Adversarial Machine Learning, Healthcare Artificial Intelligence, Medical Image Analysis, Data Poisoning, Adversarial Robustness

CITATION

Sielaff, P., Brudek, A., Nowakowski, A., Piwar, H., Pawłasek, J., Lis, M., Hernik, K., Sucharski, M., & Kacynel, E. (2026). From pixels to prompts: A structured narrative review of security, privacy, and adversarial robustness in healthcare AI. *International Journal of Innovative Technologies in Social Science*, 4(52). [https://doi.org/10.31435/ijitss.4\(52\).2026.6340](https://doi.org/10.31435/ijitss.4(52).2026.6340)

COPYRIGHT

© **The author(s) 2026.** This article is published as open access under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

AI systems are increasingly being tested or used in selected healthcare tasks. Examples include reading medical images, estimating risk from electronic health records (EHRs), analyzing physiological signals, supporting clinical documentation, and assisting with triage or decision-making. An EHR is the digital clinical record that may contain diagnoses, laboratory results, medicines, observations, and other information collected over time. Large language models (LLMs) process text, while vision-language models (VLMs) process both images and text. As these systems move closer to patient care, failures may affect patient safety, confidentiality, reimbursement, and trust (Qayyum et al., 2021; Clusmann et al., 2025; Akinci D'Antonoli et al., 2025; Abtahi et al., 2026).

Several basic terms are used throughout this review. A model is software that has learned patterns from examples. Training is the learning stage. Use, also called inference, is the later stage when the trained model receives a new image, record, signal, or instruction and produces an answer. The new data given to the model are called the input. A query is one request sent to the model. A workflow, sometimes called a pipeline, is the full chain of steps from collecting data to producing and using the model's output.

AI systems can make ordinary errors because the data are incomplete, the task is difficult, or the model performs poorly on patients who differ from the training population. Adversarial machine learning studies a different problem: a deliberate attempt to make the system give a wrong answer, follow an unintended instruction, or reveal information that should remain private. For example, an attacker might alter part of an image, place misleading cases in the training set, or repeatedly question a model to learn whether a particular patient's record was used during training (Ma et al., 2021; Jin & Li, 2023; Alber et al., 2025; Z. Zhang et al., 2022; Pedarla et al., 2023). These are controlled research scenarios, not evidence that the same events are common in hospitals.

Ma et al. (2021) suggested two possible reasons for the vulnerabilities they observed across three medical-image datasets: medical images contain complex biological textures with high-gradient regions that a model may rely on, and models that are more complex than the task requires can have a sharp loss landscape that small changes exploit. These explanations apply to the models tested in that study and should not be treated as universal causes.

The included studies examined several points in the life of an AI system. Some changed a new image or record only when the trained model was being used. Others corrupted training data, planted hidden trigger-based behavior, or tried to recover private information from a trained model (Allyn et al., 2020; Ozbulak et al., 2019; Zbrzezny & Grzybowski, 2023; Menon et al., 2021; Alber et al., 2025; Jin & Li, 2023; Z. Zhang et al., 2022; Subbanna et al., 2021).

Studies of language models examined three related problems. Prompt injection means placing an instruction in text, a document, or an image so that the model follows that instruction instead of its intended task. Jailbreaking means wording a request to bypass the model's safety rules. Malicious fine-tuning means adapting a model with harmful examples so that its later behavior changes (Clusmann et al., 2025; H. Zhang et al., 2024; Y. Yang et al., 2025; Huang et al., 2025).

Many defenses have been proposed, but the reviewed evidence does not identify one method that is broadly effective and ready for routine clinical use (Levy et al., 2024; Apostolidis & Papakostas, 2021).

Previous reviews often examine only one part of the problem. Some focus on medical imaging, some on radiology, and others on the privacy and security of healthcare LLMs (Apostolidis & Papakostas, 2021; Kaviani et al., 2022; Muoka et al., 2023; Dong et al., 2024; Surekha et al., 2024; Sorin et al., 2023; Dietrich et al., 2025; Akinci D'Antonoli et al., 2025; Alissa, 2026; Aroua et al., 2026). Broader surveys have also appeared: Abd El-Aziz et al. (2026) reviewed vulnerabilities, defenses, and recent research on adversarial threats across healthcare AI, Kalpana et al. (2025) surveyed the robustness of deep-learning models to adversarial attacks in wireless body-area networks, and Al-Dhabyani (2026) reviewed data privacy and security for large language models in medicine.

A combined view is useful because one clinical AI service may connect several components, such as an image model, a language model, retrieved documents, and shared pretrained software. A weakness in one component may therefore affect another. However, the included studies did not follow one compromised model across several real clinical products, and harmful behavior may change after later adaptation, filtering, or validation (Hanif et al., 2024; Alber et al., 2025). This review therefore compares possible routes of attack without assuming that every healthcare AI system contains all of them.

This review includes 113 publications published between 2018 and 2026. The year 2018 was the publication year of the oldest included study; it was not a pre-set eligibility threshold.

We first provide a plain-language map of six threat types. We then summarize findings from medical imaging, EHRs, physiological signals, genomic data, language models, and collaborative learning. Next, we compare proposed defenses and their limitations. Finally, we discuss what the evidence does and does not show for clinical practice, regulation, and future research. Technical names are retained so that readers can identify the original methods, but the main text focuses on what the researchers changed, what happened, and how the result should be interpreted.

2. Methods

2.1 Review design

This study was designed as a narrative review of cybersecurity risks affecting AI systems in healthcare. The review focused on adversarial threats, privacy risks, and mitigation strategies relevant to AI-based systems used in medical imaging, pathology, electronic health records, clinical decision support, Internet of Medical Things applications, wearable or sensor-based health monitoring, federated learning, large language models, vision-language models, and multimodal medical AI.

The review was intended to map the main categories of cybersecurity threats, describe recurring vulnerabilities, compare proposed defensive strategies, and identify research gaps relevant to the safe deployment of AI in healthcare. Because the literature included heterogeneous AI architectures, datasets, clinical contexts, threat models, and evaluation metrics, a narrative synthesis was considered more appropriate than a quantitative meta-analysis.

2.2 Databases and search strategy

A structured literature search was conducted in PubMed, Scopus, and Web of Science. The search strategy combined terms related to artificial intelligence and machine learning with terms referring to cybersecurity, adversarial attacks, adversarial robustness, data poisoning, backdoor attacks, prompt injection, jailbreaking, model inversion, membership inference, data leakage, gradient leakage, federated learning security, privacy-preserving machine learning, and healthcare or medical applications. The searches were run in July 2026, and records were exported on 5 July 2026.

After the three databases were combined and duplicate records were removed, the records covered 2014 to 2026. After full screening, the oldest included publication was from 2018 and the newest from 2026. The initial database export contained 339 PubMed records, 201 Scopus records, and 783 Web of Science records, giving 1,323 records before duplicate removal.

2.3 Which publications were included

We included English-language publications that met three conditions. First, the study had to examine a machine-learning or other artificial-intelligence system, such as deep learning, an LLM, a foundation model, or a system trained jointly by several sites. A foundation model is a large model trained on broad data and later adapted to more specific tasks. A vision-language model is one type of multimodal model that processes images and text together.

Second, the AI system had to be situated in a healthcare, clinical, biomedical, medical imaging, pathology, electronic health record, IoMT, wearable health monitoring, or health data context.

Third, the study had to examine a security, privacy, or robustness problem involving the model, its data, the route by which it produced an answer, the environment in which it was used, or its interaction with users. Eligible topics included deliberately altered inputs, corrupted training data, hidden triggers, malicious prompts, attempts to infer or reconstruct private data, leakage during collaborative training, attack detection, training with attacked examples, and privacy-preserving methods.

We excluded publications that discussed healthcare AI but did not examine security, privacy, robustness, or deliberate manipulation. We also excluded work on general hospital cybersecurity, such as ransomware, phishing, or network intrusion, when the AI model itself was not the subject of the study.

Eligible publication types included full journal articles, conference proceedings papers, reviews, detailed technical analyses, and preprints when no more complete peer-reviewed version was available. To enter the full-text synthesis, a report had to describe the healthcare AI system or data, the attack or defense, and either a measured result or a sufficiently detailed technical analysis. Short editorials, comments, posters, and conference abstracts without this information were excluded.

When the same work existed as both a preprint and a peer-reviewed article, we kept the peer-reviewed version. When a later journal article substantially expanded an earlier conference paper, we used the more complete report.

2.4 How publications were selected

We first removed duplicate database records. We mainly matched digital object identifiers (DOIs). When a DOI was missing, we matched cleaned titles and publication years. This removed 391 duplicates from 1,323 records and left 932 unique records.

We then screened titles and abstracts in two passes. The first pass removed 557 clearly irrelevant records, leaving 375. The second pass applied the eligibility criteria more closely and removed 259, leaving 116 reports for full-text assessment.

At full-text review, we excluded one poster abstract that did not report enough methodological detail, such as the dataset, the threat model, or the evaluation measure, to assess or reproduce its claims, and we set aside two earlier versions that had been replaced by more complete publications. This left 113 included publications.

We did not prospectively divide the included publications into separate numerical groups of primary and contextual evidence. Instead, the text states whether a publication provides an experiment, a technical method, or background and review context.

2.5 What information was extracted and how it was summarized

For each included publication, we recorded the citation, the type of AI system and medical data, the clinical setting, what the attacker was assumed to know or control, the attack or defense method, and the main reported result. In the main text, each study is described in terms of what was changed, what access the attacker had, what happened, and what the result does and does not mean. Exact figures were kept selectively when they made the magnitude of an effect or a study limitation easier to understand.

Experimental studies were used as evidence for measured attack or defense results. Reviews and conceptual papers were used for definitions, background, and research gaps; they were not treated as proof that a method was effective.

We grouped the evidence by threat type and compared findings across images, health records, physiological signals, genomics, and language. The categories sometimes overlap. For example, a hidden backdoor can be introduced through corrupted training data. Defenses were grouped by what they do, but their benefits and limitations were described separately. Publications that mainly broaden the scope or support the classification, rather than contributing a standalone numerical result, are described alongside the most relevant studies in the Results section.

Table 1. Summary of the review methodology.

Parameter	Details
Review type	Structured narrative review with documented stages of searching, screening, and selection. Results were summarized descriptively; no meta-analysis was performed.
Databases	PubMed, Scopus, and Web of Science
Search period	Searches were run in July 2026 and records were exported on 5 July 2026. PubMed was limited to 2020 onward; Scopus and Web of Science had no equivalent lower-year limit. Included publications were published from 2018 to 2026.
Search structure	Three groups of terms were combined: AI or machine learning; security or privacy; and healthcare.
Inclusion	English full text; a healthcare or biomedical AI system or dataset; a security, privacy, or robustness question; and enough detail to identify the attack or defense and its result or technical reasoning.
Exclusion	Healthcare AI without a security, privacy, or robustness focus; general hospital or network security when AI was not the object of analysis; short reports without enough detail; and earlier versions replaced by a more complete report.
Records identified	1,323 (PubMed 339; Scopus 201; Web of Science 783)
Records after duplicate removal	932
Full-text reports assessed	116
Evidence base	113 included publications. The text identifies whether a source provides an experiment, a technical method, or background and review context.
Synthesis	Studies were grouped by threat type and type of clinical data. Because their designs differed, no pooled effect or ranking of defense families was calculated.

Table 2. Study-selection flow for the narrative review.

Stage	Records	Decision or note
Records identified in PubMed	339	Database search
Records identified in Scopus	201	Database search
Records identified in Web of Science	783	Database search
Total before duplicate removal	1,323	Combined database records
Duplicates removed	391	Matched mainly by DOI; PubMed identifier and cleaned title/year were used when needed
Records after duplicate removal	932	Screened by title and abstract
Excluded in two screening passes	816	557 in the first pass and 259 in the second
Full-text reports assessed	116	Assessed against full-text eligibility criteria
Excluded: insufficient publication type	1	Poster abstract without enough methodological detail
Excluded: superseded versions	2	Earlier conference paper or preprint replaced by a later version
Review or background evidence	Not separately counted	Used for definitions, framing, and research gaps; included within the total of 113
Experimental or technical evidence	Not separately counted	Used for study-specific attacks, defenses, methods, and outcomes; included within the total of 113
Total included publications	113	All publications retained after full-text assessment and version reconciliation

3. Results

3.1 A plain-language map of deliberate attacks

Readers do not need to memorise algorithm names to understand an attack study. It is usually enough to know what the attacker is trying to achieve, whether the attack happens while the model is being trained or later when it is used, what information the attacker can see, and what the attacker can change. Together these define the threat model, which in plain language is simply the scenario the researchers chose to test.

An attacker may be able to change a new image or patient record, a training example or diagnosis label, an instruction given to a language model, or a model update. Put simply, a model update is a set of numbers describing how a model should change after learning from new data. In collaborative training, a hospital or device may send these numbers to a central server instead of sending the underlying patient records.

The six threat groups used in this review are explained below and summarized in Table 3.

Changing a new input after training (evasion): The attacker alters only the new image, record, or signal presented to an already trained model. For example, selected pixels in a chest radiograph may be changed so that the model reports a different diagnosis. The model itself is not retrained. The goal may be any wrong answer or one specific answer chosen by the attacker. The change may be difficult to see, but invisibility is not required (Ma et al., 2021; Tajak et al., 2025).

Corrupting the material used for training (data poisoning): The attacker changes examples, diagnosis labels, or text from which the model learns. This is similar to teaching from deliberately misleading cases. The aim may be to reduce performance in general or to change behavior only in selected situations (Abtahi et al., 2026).

Teaching a hidden trigger rule (backdoor or trojan): The model learns a concealed rule during training or later modification. The rule is activated only when a particular trigger is present. A trigger might

be a small image pattern or a particular pattern of missing values in an electronic health record. On ordinary cases without the trigger, the model may still appear highly accurate (Z. Y. Yang et al., 2024; Joe et al., 2022; Hanif et al., 2024).

Extracting private information (privacy or inference attack): The attacker's goal is not necessarily to change a diagnosis. Instead, the attacker tries to learn something about confidential data. This may include deciding whether a person's record was used for training, inferring a sensitive characteristic, or reconstructing an image or text that resembles the original data (Z. Zhang et al., 2022; Subbanna et al., 2021; Shi et al., 2025).

Manipulating language or image-and-language models: A malicious instruction may be placed in a user's message, a retrieved document, or even an image. Prompt injection makes the model follow an unintended instruction. A jailbreak tries to bypass safety rules. Malicious fine-tuning changes later behavior by adapting the model with harmful examples (Clusmann et al., 2025; H. Zhang et al., 2024; Y. Yang et al., 2025; Huang et al., 2025).

Attacking collaborative training: Several hospitals or devices may train one model together while keeping raw records at each site. A malicious participant or central server can send misleading numerical updates, plant a backdoor, or try to reconstruct private information from the numbers exchanged during training (Liyanage et al., 2025; Jin & Li, 2023; Shi et al., 2025).

The attacker's level of access determines both how difficult the attack is and how closely the experiment resembles a real clinical system. In a full-access or white-box study, the attacker knows the model's design, settings, and internal calculations. This is similar to testing a medical device while holding its complete technical blueprint. White-box studies are useful for finding worst-case weaknesses, but an outside attacker may not have this level of access.

In a partial-access or gray-box study, the attacker knows only some internal details. In a black-box study, the attacker cannot inspect the model and can only submit inputs and observe the answer or score. A closed-box study is even more restricted: the attacker may see only a final category or a ranked list. Black-box and closed-box studies may be closer to what an external user could attempt, but feasibility still depends on how many requests the system allows and how much information it returns (Sorin et al., 2023; Akhtom et al., 2024; Joseph, Jeffrey, et al., 2026).

A targeted attack tries to force one specific wrong answer, such as making a malignant lesion appear benign. An untargeted attack succeeds if it causes any wrong answer. These distinctions matter because a high success rate in a full-access, targeted laboratory test does not describe the same practical risk as a restricted attack through an ordinary user interface (Sorin et al., 2023).

Table 3. Six types of deliberate attack, explained for clinical readers.

Threat	When it happens	What happens in plain language	What access is usually needed	Healthcare example
Changing a new input (evasion)	When the trained model is used	A new image, record, or signal is altered so that the existing model gives a wrong or chosen answer.	From full technical access to ordinary user access, depending on the method	Altered chest radiographs, skin images, retinal images, or pathology slides (Ma et al., 2021; Ghaffari Laleh et al., 2022)
Corrupting training data (data poisoning)	During training or retraining	Misleading images, labels, or text are placed in the material from which the model learns.	Access to training data or the training process	Corrupted COVID-19 chest-radiograph data and vaccine misinformation in LLM training (Menon et al., 2021; Alber et al., 2025)
Hidden trigger rule (backdoor or trojan)	Inserted during training; activated during later use	The model behaves normally until a particular trigger appears, then gives the attacker's chosen answer.	Access to training data, the model, or collaborative updates	Trigger-based errors introduced through raw CT measurements and EHR missing-value masks (Z. Y. Yang et al., 2024; Joe et al., 2022)

Threat	When it happens	What happens in plain language	What access is usually needed	Healthcare example
Extracting private information	During or after training	The attacker tries to learn whether a record was used, infer a private attribute, or reconstruct data.	Model queries, shared updates, or internal model information	Membership inference on health data and reconstruction of brain MRI (Z. Zhang et al., 2022; Subbanna et al., 2021)
Malicious instructions to language or multimodal models	During use and sometimes during fine-tuning	An instruction in text, a document, or an image makes the model give unsafe, false, or policy-violating output.	Control of a prompt, retrieved content, an image, or fine-tuning data	Prompt injection in oncology models and jailbreaking of clinical LLMs (Clusmann et al., 2025; H. Zhang et al., 2024)
Attacking collaborative training	While several sites train one model together	A malicious participant or server sends misleading updates, plants a backdoor, or reconstructs private data from shared numbers.	Control of a participating site or central server	Poisoning, backdoors, and data reconstruction in federated healthcare models (Jin & Li, 2023; Faraoun et al., 2026; Shi et al., 2025)

3.2 Changing a medical image after the model has been trained

An evasion attack changes a new image after the model has already been trained. The attacker does not retrain the model. Instead, the attacker alters the copy of the image that is sent to the model and tries to change the answer. Researchers call the altered image an adversarial example.

Many studies used full-access stress tests with names such as FGSM (fast gradient sign method), PGD (projected gradient descent), and Carlini-Wagner. Each is a method for computing a deliberately misleading change to an image, and the names matter less than the shared idea: the researchers inspected the model's internal calculations (its gradients) and used them to find small image changes especially likely to cause an error. Because an outside user would not usually have this access, these experiments are best read as worst-case technical tests rather than as demonstrations of an easy hospital attack.

Full-access imaging stress tests

Ma et al. (2021) made extremely small, calculated changes to retinal images, chest radiographs, and dermoscopic photographs. Before the attack, the models were correct in roughly 9 of 10 cases. Under a white-box PGD attack (with a per-pixel limit of $1/255$), accuracy fell from 91.03% to 0% for fundus images, from 93.99% to 0% for chest radiographs, and from 87.62% to 0.43% for dermoscopic images. A separate detector in the same study told attacked from clean images with an AUC above 98%. The practical lesson is therefore two-sided: the models were highly vulnerable when the attacker had full internal access, but these particular attacks also left detectable traces.

Allyn et al. (2020) trained a classifier on the HAM-10000 dermoscopy dataset, which sorts skin lesions into seven types, including melanoma and basal cell carcinoma (both skin cancers). They then added small, humanly invisible changes to the pixels of each image, created by a separate neural network trained to fool the classifier, so that the model returned the same incorrect answer for every lesion: melanocytic nevus, which is a benign mole. In effect, cancers and other lesions were all relabeled as harmless. Overall accuracy fell from 84% to 67%, the attack sent all perturbed images to the melanocytic-nevus class, with a 100% fooling rate in each of the six non-target lesion categories. In a small observer experiment using 50 images, people were no better than chance at telling which images had been altered.

Ghaffari Laleh et al. (2022) studied models that read digitized pathology slides and predict the cancer subtype. They compared an older network design (a convolutional network) with a newer one (a vision transformer), and attacked both in the same way: adding tiny, invisible changes to the slide images to force a wrong prediction. On a kidney-cancer subtyping task, the older model's ability to tell the subtypes apart (measured by AUROC) collapsed to near chance, while the vision transformer held up much better; on a separate gastric-cancer task, the older model's

AUROC fell to zero. The study shows that the choice of model design affected how easily the model could be fooled; it does not show that vision transformers are always safe.

Kwon and Jeong (2022) studied segmentation, in which a model draws the boundary of a structure rather than choosing only one diagnosis. As in the classification studies, the attack added small, invisible changes to the image. After these changes, the model's outlines differed greatly from the reference outlines. The study therefore shows that an attack can distort not only a diagnostic label but also the area marked for measurement or review.

The attacks described so far relied on changes too small for a person to see, but an attack does not have to be invisible to succeed. Tajak et al. (2025) attacked brain-MRI classifiers by setting a single pixel to bright white. In some images, that one altered pixel was enough to change the predicted diagnosis, and the models were most vulnerable at low image resolution. Because a single white pixel is plainly visible, the study shows an oversensitivity to one pixel rather than a hidden, undetectable manipulation.

Attacks with less access to the model

Joseph, Jeffrey, et al. (2026) tested an attacker who could see only the model's final label for an image, meaning the single category it settled on (for example, benign or malignant), with no view of the confidence behind that label or of the model's internal calculations. The attacker repeatedly submitted slightly different versions of an image, on the order of a hundred attempts per image, until one produced a wrong answer, and this succeeded in more than 9 of 10 tested cases. At attack time this method used only final labels and no separate stand-in model; however, its offline setup trained a small helper network on clean images and attacked versions, so it was not entirely knowledge-free. Its main practical limitation is that it needs many repeated queries, which a service that rate-limits or monitors queries could disrupt.

W. Y. Li et al. (2025) attacked a medical image-search system, which takes an image and returns the most similar images from a database. The attacker could see only the ranked list of returned results, not the system's inner workings. Using those returned rankings, they trained a stand-in copy of the system, and then used that copy to craft small changes to a query image that pushed clinically relevant matches far down the results. They tested this on skin and chest-radiograph datasets, and the attack required repeated query access to the system's returned rankings.

Many medical AI tools come with a built-in explanation, usually a colored heat map laid over the image that highlights the areas the AI focused on. The idea is that a clinician can glance at the heat map and sanity-check the result, confirming the AI looked at the genuinely suspicious region rather than something irrelevant. W. Li and Zainon (2026) showed that this safety check can itself be defeated. Using the model's gradients, they computed a small perturbation to the image pixels, kept below a fixed limit so that it remained imperceptible, and tuned it so that this single change would fool both outputs at once: the model returned the wrong diagnosis, and the heat map shifted so that it no longer flagged anything unusual. Across all nine combinations of three image types and three AI models, classification accuracy fell to zero; the joint success against both the diagnosis and the explanation was high but varied with the explanation method and the settings. The concern is that a clinician could be shown a wrong answer alongside an explanation that still looks unremarkable, and so might not realize the AI had been fooled.

Other imaging studies filled in specific attack routes. Paschali et al. (2018) attacked brain-MRI segmentation and found that three networks with similar accuracy differed markedly in how easily they were fooled. Pal et al. (2024) measured how robustness fell when simultaneous patch attacks were applied to chest-radiograph and OCT images, while Sipola and Kokkonen (2021) explored whether changing a single pixel could mislead a medical-imaging model. Imam (2024) examined potential attacks on an explainable-AI-enabled digital twin in healthcare. Further studies reported adversarial vulnerabilities in COVID-19 diagnostic models on medical Internet-of-Things devices (A. Rahman et al., 2021), in a COVID-19 face-mask-detection system (Sheikh & Zafar, 2023), and in medical-image classifiers more generally (Krzton et al., 2025; Tsai et al., 2023), while Chen et al. (2025) developed a method to boost adversarial transferability in vision-language models.

Across these imaging studies, medical-image classifiers, segmentation systems, image-search tools, and explanation maps could all be manipulated under controlled conditions. This body of work reveals possible failure modes; it does not show how often such attacks occur in clinical practice or whether they have harmed patients.

3.3 Corrupting the data used to train a model

Data poisoning happens while a model is learning or being updated. The attacker changes the examples from which it learns, for example by altering an image, attaching the wrong diagnosis, or adding misleading text. A useful analogy is a trainee revising from notes into which a few incorrect answers have been slipped, so the mistakes are absorbed during learning rather than added afterwards.

Poisoning that broadly reduces performance

Menon et al. (2021) created contradictory chest-radiograph training examples by blending normal images into thousands of images that were still labeled as COVID-19. After retraining, accuracy fell from about 96% to 87%. This was a large, whole-image overlay manipulation applied to a chest-radiograph corpus combined from three public sources; the study did not measure how visible the change would be to a person. It shows that incorrect image-label pairs can damage learning, but it was not a subtle real-world poisoning campaign.

In this data-poisoning attack, Paracha et al. (2024) corrupted 10% of a skin-lesion image dataset. They added a small, engineered pattern of noise to one in ten of the training images, while leaving the diagnosis labels correct, so that these images would mislead the model during training. They then compared three types of model: two traditional machine-learning models and one neural network. After learning from this partly corrupted data, one traditional model's accuracy fell to about half and the other's to roughly a quarter, while the neural network changed very little. This shows that different models can react very differently to the same corrupted data. Because the study used only one dataset and fairly basic versions of each model, it does not prove which type of model is generally safest.

Singkorapoom and Phoomvuthisarn (2023) used generative AI to create artificial chest X-rays and deliberately attached the wrong diagnosis to some of them. All five diagnostic models they tested became less accurate as more of these mislabeled images were mixed into the training data. The lesson is that an image can look medically plausible yet still be harmful training material if the diagnosis attached to it is wrong.

Poisoning that teaches a hidden shortcut

S. B. Lee (2025) gave a chest-radiograph model a false lesson: that a white square means pneumonia. The researchers added a small white square to normal images and relabeled those images as pneumonia. When only a very small number of such fakes were added, the square had little effect. But when the square-marked fakes made up about one in six of the abnormal-class training images (100 fakes alongside 500 true pneumonia images, with 500 normal images as controls), the model increasingly treated the square itself as evidence of pneumonia, and it became worse on ordinary images. Heat maps (which show where the model was "looking") confirmed this: its attention shifted toward the square. This is a clear example of a model learning an artificial shortcut instead of a genuine clinical feature.

Poisoning language models

Alber et al. (2025) replaced just one word (technically, one "token") in every 100,000 of an AI language model's training text with vaccine misinformation. Afterwards, the model gave harmful answers more often, a relative increase of roughly 5 to 7% rather than a rise of that many percentage points, yet its scores on the standard performance tests (benchmarks) barely moved, so those routine tests did not reveal the problem. The study shows that a very small amount of targeted misinformation can distort the model's later answers without causing an obvious overall decline.

How much poisoned data matters depends on what was changed, how much was changed, which model was used, and how it was tested. Standard validation can miss a problem that appears only in the specific case the attacker is aiming for, such as when a hidden trigger is present, but the reviewed evidence does not show that every poisoning attack will stay hidden.

3.4 Teaching a model a hidden trigger rule: backdoors and trojans

In this context, a backdoor is a hidden rule learned by the model. The model behaves normally when the trigger is absent, but gives the attacker's chosen answer when the trigger appears. The attack therefore has two stages: the rule is planted during training or later adaptation, and the trigger is presented when the trained model is used.

CT analysis illustrates how a trigger can be placed before the final image even exists. A CT scanner first collects raw measurements, reconstruction software turns those measurements into an image, and a separate AI model may then read (classify) that image to make a diagnosis. Z. Y. Yang et al. (2024) contaminated raw

measurements used during training so that the final classifier learned a hidden trigger. Later, adding the same pattern to new scanner measurements caused the attacker's chosen diagnosis after reconstruction. The reconstructed scans remained visually very similar to the originals, and the model stayed accurate on ordinary cases, but almost every triggered case was pushed to the attacker's chosen diagnosis. This was a simulated compromise of the scanner-to-AI pathway, not an observed hospital incident.

A trigger does not have to be a mark on an image. Joe et al. (2022) used a particular pattern of missing laboratory values in an electronic health record. The pattern was made to look plausible, as though certain tests had simply not been ordered or recorded. After poisoning a small part of the training data, the model changed little on ordinary records but was strongly disrupted when that missing-value pattern appeared. In the simplest of the models they tested, poisoning less than 1% of the training set produced the attacker's chosen output in almost every triggered case.

Modern medical AI often starts from a large, general-purpose model (already trained on huge numbers of images and text) that is then customized for a specific job using only a small set of local examples. Hanif et al. (2024) showed that this customization step is itself an opening for attack. They adapted such a model to sort pathology images into tissue types, using the Kather dataset (a standard set of labeled colorectal-tissue images), but first they tampered with just a handful of the customization examples: 8 of the 288 used. Each poisoned example carried a hidden trigger, namely a small mark on the image plus a faint, deliberately designed noise pattern that is invisible to the eye. The finished model then behaved normally on untouched images, matching a clean model's accuracy. But whenever the trigger was present, over 90% of images were misclassified as the tissue type the attacker had chosen. In this study the attacker was assumed to be the party providing the customization (a prompt-learning service), with full access to the base model and control over part of the small local dataset.

Backdoors introduced through collaborative training

Faraoun et al. (2026) simulated 100 sites training one shared image model, and had malicious sites plant a hidden image trigger, a secret visual mark that makes the model give the attacker's chosen answer whenever it appears. Each malicious site placed the trigger in some of its local images before sending its training updates. This backdoor continued to work even after the malicious sites stopped taking part, and most images containing the trigger still produced the chosen result, showing that a backdoor can persist in a shared model after its original source has gone. The images were low resolution, and the attack required the attacker to model, or anticipate, how future training rounds would unfold.

Xi et al. (2021) poisoned a small part of one participant's local text data in a shared Alzheimer-disease classification task. The shared model behaved normally on ordinary text but gave the attacker's planted response whenever the hidden trigger was present, in almost every tested case. Methods intended to detect the malicious participant are defense strategies and are discussed in Section 3.9.

Jin and Li (2023) studied a related attack on a shared program that generates synthetic medical images. One malicious participant added mislabeled triggered images to its local data. Here the main harm was not one trigger-controlled diagnosis; instead the shared image generator (the model that produces the synthetic images) produced poor-quality images. The same basic manipulation can therefore cause different failures depending on what the system is designed to do.

A model can therefore appear accurate on ordinary test cases and still contain a hidden rule. Clean validation alone cannot rule out a backdoor, because the trigger may never appear in the ordinary test set.

3.5 Using a model to learn private information

Privacy attacks primarily seek to uncover confidential information rather than change a diagnosis. Instead, the attacker uses the model to learn about the patients whose data trained it. The reviewed studies fall into four types, described below.

Membership inference. This attack works out whether a particular patient's record was used to train a model or to build a synthetic dataset, without copying the record, by exploiting the statistical trace a model carries of what it learned from (using shadow models, differences in confidence on seen versus unseen records, or probing how the output changes when one field is altered). Z. Zhang et al. (2022) found it worked on partly synthetic data, while fully synthetic data was only marginally susceptible; Pedarla et al. (2023) found it fairly reliable on a diabetes dataset (correct about three times in four, catching about nine in ten true members). C. H. Zhang and Bonomi (2022) cut the attack to chance with a privacy safeguard, but cancer-subtype accuracy fell from 89% to 58%, and Nikolaidis and Efraimidis (2025) found federated learning reduced the leak, especially with client selection, though stronger attackers recovered part of the signal. Simply belonging to a dataset can be sensitive, and defenses that hide membership often cost accuracy.

Model inversion. Here the attacker rebuilds an input, or infers a private attribute, from information stored inside the model, for example with a decoder that reverses the model's compressed internal representation. Subbanna et al. (2021) reversed two brain-MRI models into blurred but recognizable scans and matched them to the correct patient in most cases (30 of 40 and 26 of 40). Krall et al. (2020) added controlled random noise during training and reduced a separate attack that infers a patient's sensitive attribute (scored by R-squared) to zero while keeping predictions accurate, in one model. Malik et al. (2026) built a detector that flagged most such attacks on chest-X-ray and ECG models. A model's compressed internal data can still hold recognizable patient information.

Collaborative-learning leakage. Instead of raw records, hospitals share numerical training updates, and these can be read back. A malicious server can tamper with the shared model so the returned updates encode the data (update or gradient inversion), even bypassing a protection called secure aggregation; in split learning, the shared feature maps can be inverted into images. El Azzouzi et al. (2026) recovered more than nine in ten words from clinical reports when reports were processed one at a time (recovery fell as more were processed together); Shi et al. (2025), with a method called MedLeak, rebuilt most or nearly all of the images and much of the text even under secure aggregation (the exact recovery ranged with batch size and setting); and Roth et al. (2022) inverted split-learning feature maps back into recognizable MRIs, finding that a later split plus moderate privacy noise reduced this. Keeping raw data local does not guarantee privacy; a dishonest server can engineer leakage.

Synthetic-note leakage. This checks whether AI-written clinical notes reproduce identifiers present in their (de-identified, dummy-PHI) source notes, comparing one-shot generation (a full example note) with keyword-only prompting. With Mistral 7B, one-shot generation reproduced the dummy place and date identifiers in about 85 to 87% of notes, while keyword-only prompting cut this sharply (to about 30% and near zero); separately, GPT-3.5 one-shot notes trained an ICD-9 diagnosis-coding model about as well as the source notes (Chuang et al., 2025). How the model is prompted matters: keyword generation lowers identifier exposure but does not guarantee anonymity.

Keeping raw data inside the hospital, sharing only the model's numerical updates instead of the records themselves, or training on synthetic (computer-generated) records can each reduce exposure, but none of these steps guarantees privacy. How much data leaks still depends on what the system reveals, who controls the central server, and what safeguards are in place.

3.6 Attacks on electronic records, clinical text, and physiological signals

An electronic health record is a clinical timeline. A prediction may depend not only on a laboratory value, diagnosis code, or observation, but also on when it appears. An attacker may therefore try to change a small number of values, timestamps, or coded events so that the record tells the model a different clinical story.

Changing an electronic health record

M. Y. Sun et al. (2018) tested a model that predicts a patient's risk of dying from the first 48 hours of intensive-care data. Such data can be pictured as a table with one row per measurement (each vital sign or laboratory value) and one column per time point, and the researchers altered only a small fraction of the individual entries in it, that is, a few specific measurements at specific times. Even so, more than half of the tested records could be pushed from predicted survival to predicted death; changing predictions in the opposite direction was harder. The study did not fully verify whether every altered combination would stay clinically plausible, meaning whether a real patient could actually show those values.

An et al. (2019) attacked patient histories built from diagnosis and procedure codes recorded across many visits over time (a longitudinal record), in a heart-failure prediction task. Their method first worked out which few codes the model leaned on most, then altered those codes (adding or removing a code) to change the prediction. Only a handful of changes caused a sharp drop in the model's ranking performance, its area under the precision-recall curve (AUPR) falling from about 0.50 to 0.08 for attention-based models rather than a fall in plain accuracy. The attack assumed the attacker knew how the model worked, fully or in part, and it did not check whether the edited histories would survive a clinician's review, meaning whether a doctor reading the record would notice it had been altered.

Papangelou et al. (2019) studied models that predict how a patient would fare under different treatment options. Changing only two or three of the 25 patient variables (for example a single lab value or characteristic) could flip both the predicted outcome and the treatment the model's prediction would point to. Because the patient variables came from 985 real participants in the IHDP randomised study while only the outcomes were simulated, this shows how sensitive the model is rather than proof that a real treatment decision was changed.

Changing clinical text

Mondal (2021) tested a black-box attack (one where the attacker cannot see inside the model and only works from its outputs) on systems that read written text to identify adverse drug events. The attack replaced drug or disease terms with medically related synonyms, changed spellings, or rewrote numbers, while trying to keep each sentence understandable. On two datasets, the model's accuracy was roughly halved. Human reviewers generally found the text readable, although some of the spelling changes looked visibly unnatural.

Changing signals from medical devices and wearables

Newaz et al. (2020) attacked a simulated system that combined data from eight medical devices (vital signs such as heart rate, blood pressure, oxygen, and glucose). The readings were not changed at random: knowing each device's normal range, the attacker nudged the numeric values by small, deliberately calculated amounts (some methods use the model's own math to find the most effective tiny change) while keeping values inside each device's preset normal range. In the general evaluations the largest accuracy drop was about 32 percentage points and the highest targeted attack-success rate was about 20%; a separate test that constrained changes to those preset ranges had smaller effects. The study did not verify that the altered data would pass a clinician's inspection, and it did not observe a real change of treatment.

Sahay et al. (2026) changed the timing of signals from wearable devices (for example the heart-rate trace from a smartwatch, recorded over time) without necessarily changing the measured values themselves. In other words, the same pattern was stretched, compressed, or shifted in time before it reached the model. Accuracy fell by as much as 30 percentage points, and with a moderate amount of this time-warping more than 4 in 10 attacks succeeded. The study did not evaluate a deployed wearable service (one actually in use by patients).

Several studies also tested detectors (tools that try to spot tampered inputs) or retraining methods (re-training the model so it resists such attacks) for records and signals. Because these are defenses rather than attacks, they are discussed in Section 3.9.

These studies show that a model can be misled by changing a small part of a clinical timeline, a few medically important words, or the timing of a physiological signal. The main unresolved question is whether the altered data would stay plausible and go undetected in a real clinical workflow.

3.7 Attacks on medical language and image-and-language models

Language models are AI systems that read and write text and accept free-form instructions, and image-and-language models accept both text and images. This flexibility creates a new attack route. Prompt injection hides an unwanted instruction inside content that the model reads (for example a line of text buried in a document or a note), so the model follows it as if it came from the user. A jailbreak is a prompt worded to get around the model's safety rules, the built-in limits meant to stop it producing harmful output. Malicious fine-tuning changes the model more permanently: fine-tuning means giving a model extra training examples to adjust how it behaves, and here those examples are deliberately harmful.

Instructions hidden in images or accompanying text

Clusmann et al. (2025) used 18 confirmed malignant lesions. The unwanted instruction was placed either in the text sent with the image or directly inside the image itself, including versions shown in high contrast, low contrast, or a small font. The instruction told the model to describe the organ but to say it looked healthy, or not to mention the lesion. Across four image-and-language models, the number of missed cancers rose substantially, reaching nearly 9 in 10 cases in some models. The test was small and controlled, but it shows that text hidden in a clinical image can override the actual medical findings in that image.

Güler et al. (2026) tested 52 dermoscopic images. For each image, the researchers took the opposite of the true diagnosis and inserted it as a visible word written over the image, as text in the accompanying metadata (the written information attached to the image file), or in both places. Accuracy across three models fell from around 60% without this false label to almost zero with it. This was a deliberately severe test, because the attacker knew the correct diagnosis and supplied its exact opposite.

Hidden instructions in clinical conversations

R. W. Lee et al. (2025) created simulated multi-turn medical conversations, meaning scripted back-and-forth exchanges over several messages rather than a single question and answer. In the main experiment, the researchers programmatically inserted hidden instructions into the user's message before it reached the language model; a separate, smaller proof-of-concept simulated interception on the user's own device for one scenario. Some instructions subtly pushed a risky recommendation; others supplied fabricated guideline excerpts or false meta-analyses to make an extremely harmful recommendation appear evidence-based. The injection changed the model's decision in 102 of 108 attack cases and often continued to influence later answers in the same conversation. The harmful recommendations included prescribing meperidine (a strong opioid painkiller) to a patient taking a monoamine oxidase inhibitor, a combination that can cause a dangerous drug

reaction, and prescribing thalidomide or methotrexate during pregnancy, when both are known to cause severe harm to the developing fetus. The dialogues were simulated, and no patient received this advice.

Zahra and Chin (2026) asked eight language models about cancer treatment and tried six different jailbreaks (prompts designed to get around the models' safety rules). The prompts asked the model to adopt a role, imagine a fictional setting, disregard its rules, accept a conspiracy article, answer in Malay, or knowingly violate guidance. Each version was also tested with and without emotional pressure, such as desperation or fear. Severe misinformation without a warning appeared in about 6% of ordinary prompts, 25% of non-emotional attack prompts, and 38% of emotionally framed attacks. The models differed greatly: Claude 3.5 refused all six tested variants, whereas Mistral produced severe content in five of six emotional variants.

Jailbreak prompts and malicious model adaptation

H. Zhang et al. (2024) began with 450 unsafe medical questions and rewrote them using several jailbreak approaches, including repeated automatic rewording, persuasive framing, and text transformations meant to disguise the request. The resulting 2,700 prompts were evaluated against seven language models: the commercial GPT models through their APIs and the open-source models run from their weights. Some models answered almost every disguised harmful request, even though they often refused the plain, direct version. The same study then trained two smaller models on safe examples of these attacks and greatly reduced how often they complied with the harmful requests, but the evaluation rested on a single test set and used another AI model, rather than human experts, to judge whether an answer was harmful.

Y. Yang et al. (2025) used 1,200 de-identified patient notes and tested three malicious goals: discourage COVID-19 vaccination, recommend unsafe drug combinations, and encourage unnecessary imaging tests. First, the researchers added a hidden malicious instruction to the prompt sent to GPT-4 (a prompt injection). They then used the resulting harmful answers to fine-tune (further train) other models, building the change into the models themselves so it was more lasting. In the prompt attack, vaccine recommendations almost disappeared, unsafe drug combinations became common, and recommendations for MRI rose sharply. These were controlled tests of three predefined goals, not an evaluation of routine clinical use.

Other ways of confusing medical language and multimodal systems

Table 4. Attacks on medical language and multimodal models: selected studies.

Study	What it did	Main finding	Scope and caveats
Harada (2026)	Hid an instruction either in the learner's own message (direct) or in reference material the model reads (indirect), targeting a virtual-patient training chatbot.	The direct version sharply reduced how faithfully the chatbot followed the assigned case (no role change was observed); the indirect version had little effect. A simple text filter restored the measured fidelity close to normal but did not remove a baseline leak of forbidden information.	One model, five computer-generated cases.
Joshi et al. (2026)	Disguised harmful health questions by replacing letters with emojis or reordering characters, in English, Spanish, and transliterated Hindi.	Plain harmful questions were usually refused, but the disguised versions often slipped past the safety filter, especially in transliterated Hindi.	One version of GPT-4; does not by itself measure clinical inequity.
Huang et al. (2025)	Paired a malicious request with an image-text mismatch (for example the wrong imaging method or body region); a stronger version also optimized an adversarial text and image.	The attack raised unsafe-compliance rates and lowered refusal rates, and the version that also altered the image was stronger. The study measured unsafe compliance and refusal, not diagnostic accuracy.	Medical image-and-language models (benchmark).
Lu et al. (2026)	Swapped medically important words for similar-looking or similar-meaning ones, and made subtle changes to selected regions of the image.	The combined text-and-image attack caused more errors than changing either part alone.	Laboratory benchmark; attack construction needed a surrogate model and tools, but the black-box transfer targets did not require internal access.
Thota et al. (2024)	A full-access (white-box) pixel attack on a pathology image-and-language model.	Forced the chosen tissue label across all nine tested tissue types.	Laboratory benchmark; required full internal access.

Two further language-model studies add context. Das et al. (2024) poisoned the training data of a clinical language model (BioGPT) in a breast-cancer case study, showing that biomedical language models are also vulnerable to data poisoning. Mondillo et al. (2025), in a brief report, examined the ethical and clinical risks of jailbreaking, using ChatGPT 3.5 and 4 in two medical scenarios as illustrations.

In language and multimodal systems, the attack may be ordinary-looking text rather than altered software. Untrusted instructions can be placed in a chat message, in metadata, in a retrieved document, or in an image, so testing needs to follow every route by which text or images can enter the model.

3.8 Attacks on federated and collaborative learning

Federated learning lets several hospitals or devices train one shared model together without sending all their raw patient records to a single central database. A typical round works like this: each site receives the current model, learns from its own records, and sends the server a set of numerical adjustments (its updates). The server averages those updates to create the next version of the shared model. This reduces central copying of records, but it also requires the server to trust the participating sites and the sites to trust the server.

Disrupting the shared model

Soulb et al. (2025) simulated three institutions training a brain-MRI classifier together. One participant deliberately changed 15% of its diagnosis labels (mislabelling its own training data). A second participant sent an update that pushed the shared model in the wrong direction and made it three times stronger, so it would outweigh the honest sites when the server averaged the updates. Together, these manipulations reduced final accuracy from 61% to 23%. The experiment used only four training rounds and two repeated runs, so it shows a possible failure rather than a stable estimate of real-world risk.

Recovering private data from shared updates

Keeping records at each hospital does not mean that the shared numerical updates are harmless. An update reflects how the local records changed the model, so it carries an imprint of those records. Under certain conditions, that imprint is detailed enough that parts of the original records can be reconstructed from it.

El Azzouzi et al. (2026) simulated three sites training a language model on French clinical reports (genetic, anaesthesia, and birth records). Normally the update a site sends back is a tangled mixture of everything it learned, which is hard to turn back into the original text. In this attack the server that coordinates the training is itself the attacker. Before sending the model out, it tampers with the model's internal settings: it switches off the part of the model that blends words together for context (called the attention mechanism) and adjusts the model's weights, so that each word leaves a clean, separate trace in the update rather than a blurred one. When a site then trains on its reports, the update it returns carries those separable traces, and the server works backwards from them to identify the words, their positions, and their order, reassembling the original text. The effect was strongest when a site trained on just one report at a time (rather than many mixed together): the server then recovered most of that report's words, including names, dates, and genetic markers. The attack only works if the server is able to tamper with the model in this way before sending it out.

Shi et al. (2025) extended this idea in a method they call MedLeak and showed that it can defeat a common privacy safeguard. As in the previous attack, the server distributes a manipulated model so that a hospital's own data, whether images or clinical text, can be reconstructed from the update it returns. Here the manipulation is a hidden recording component inside the model that copies whatever data passes through it into the update. The safeguard it defeats is called secure aggregation, a scheme in which the server is allowed to see only the sum of all hospitals' updates added together, never any single hospital's update on its own, so that no individual contributor can be isolated.

MedLeak circumvents this by ensuring that every hospital except the target contributes nothing: the targeted hospital receives the recording model, while every other hospital receives a model configured to return an update of zero. Because the others contribute zero, the sum equals the target hospital's own update, from which the server reconstructs the original data. Across the tests the server recovered most or nearly all of the images and text, and the reconstructed images were almost indistinguishable from the originals.

Split learning and membership leakage

Split learning divides one neural network into two parts, one held by the hospital and one by another organisation. The hospital passes an image through its part and sends onward only the intermediate result, a set of numbers produced at the layer where the network is divided, rather than the image itself. The difficulty is that when the network is divided at an early layer, this intermediate result still closely resembles the original image.

Roth et al. (2022) showed that in a brain-MRI model these early representations could be reconstructed into recognisable images. Moving the division point deeper into the network, or adding a moderate amount of noise, reduced the leakage, although stronger protection began to harm the model's segmentation task (its outlining of structures in the scan).

Membership inference can also be aimed at a whole person rather than a single record: instead of asking whether one entry was used in training, the attacker asks whether all of a given individual's data went into the shared model. Nikolaidis and Efrimidis (2025) trained practice models (their own copies of the model, built on data they controlled) and used them to judge whether a particular person's activity data had contributed to the shared model. Carefully choosing which sites took part, and adding privacy noise, weakened the tell-tale signal but did not remove it. The study used a single dataset and assumed the attacker knew part of the model's design.

Several studies also proposed ways to filter or limit suspicious site updates. Because these are defense strategies rather than attacks, their detailed results are discussed in Section 3.9.

Additional work looked at collaborative and privacy-preserving setups. Darzi et al. (2024) examined the transferability of adversarial examples in federated learning for medical imaging, reporting high attack-success rates for small perturbations on brain-tumour MRI (meningioma and glioma) and different transferability patterns on pathology images. Ayensu et al. (2025) proposed a defense framework that checks each site's model update and, instead of blindly averaging them all, discards the ones that look anomalous before combining them into the shared model, countering data poisoning, model poisoning, backdoor, and man-in-the-middle attacks; it flagged about 84% of attacks while holding about 86% accuracy under attack, close to its roughly 90% clean baseline.

Paul et al. (2021) defended retinal-image diagnostics against privacy attacks by training on computer-generated stand-in images (synthetic look-alikes produced by a generative model) instead of real patient scans, so the resulting model resists attempts to infer whether a real patient image was used to train it (a membership-inference attack); the authors did not target image reconstruction; they also introduced a combined score that balances diagnostic accuracy against an attacker's success. Ullah et al. (2025) built a security system for everyday consumer health devices such as wearables, which connect to the internet only from time to time; it trained across the devices themselves rather than pooling their data centrally, and it reported about 97% detection accuracy (97.45% and 96.7%) on two public network-security datasets.

In a dual-use example, Yu et al. (2026) put a backdoor to protective use: because a trained model that outlines structures in medical scans is valuable and can be copied, they hid a secret ownership marker inside it (StealthMark) to let developers prove the model is theirs; crucially, the marker did not change the outlines the model produced (Dice and AUC dropped by less than 1 percentage point in the highlighted tests), whereas earlier backdoor watermarking could degrade segmentation performance by as much as 39%.

Collaborative training reduces the need to centralize raw records, but it creates new trust questions. A malicious site can corrupt the shared model, and a malicious server may be able to recover data from information that hospitals believed was safe to share.

Table 5 gives one plain-language example for each threat class. It is not a ranking. The studies used different data, models, attacker access, and outcome measures, so their percentages should not be compared directly.

Table 5. Representative attack studies, explained for clinical readers.

Threat	What the attacker changes	Clinical data or task	Representative finding and how to read it
Evasion	A new image, record, or signal after the model has been trained.	Medical-image diagnosis, segmentation, image search, and explanation maps	In a stress test where the attacker knew the model's inner workings, tiny calculated changes to the image reduced three models from roughly 9 in 10 correct answers to almost none. Attacks that only probed the model from the outside, by trying many queries, also succeeded on most tested images (Ma et al., 2021; Joseph, Jeffrey, et al., 2026). This shows vulnerability under the stated laboratory access, not how often such attacks happen in the clinic.
Data poisoning	Examples, labels, or text used during training.	Chest radiographs, skin images, wearable data, and medical language models	Corrupting 10% of one skin-lesion dataset caused a large accuracy loss in one model type, a moderate loss in another, and little change in a third. Replacing about one token in 100,000 of the training text increased harmful language-model answers by roughly 5 to 7% (a relative increase, not percentage points) (Paracha et al., 2024; Alber et al., 2025). The effect depends strongly on the data and model.
Backdoor or trojan	A hidden rule is planted during training and activated later by a trigger.	CT measurements, missing EHR values, pathology images, and collaborative learning	Triggered cases were sent to the attacker's chosen answer in nearly all tests while ordinary cases remained mostly correct (Z. Y. Yang et al., 2024; Joe et al., 2022; Hanif et al., 2024). This is why clean validation alone can miss a backdoor.
Privacy or inference	Model answers, the model's internal number patterns, or training updates are used to work out or rebuild private data.	Tabular records, genomics, MRI, federated learning, split learning, and synthetic notes	Membership attacks sometimes identified training records better than chance, and reconstructed brain MRIs could be matched to the correct original in most of a 40-scan test. Malicious servers also recovered much of the tested clinical text and imagery (Subbanna et al., 2021; El Azzouzi et al., 2026; Shi et al., 2025). Local storage alone does not guarantee privacy.
LLM or multimodal attack	An unwanted instruction is placed in a message, metadata, document, or image, or harmful examples are used to fine-tune (further train) the model.	Medical advice, lesion interpretation, virtual patients, and visual question answering	Hidden instructions changed the decision in 102 of 108 simulated medical-advice attacks. In a separate dermoscopy test, adding the opposite diagnosis to the image or metadata reduced model accuracy from around 60% to almost zero (R. W. Lee et al., 2025; Güler et al., 2026). Text and metadata must therefore be treated as possible attack inputs.
Federated or split-learning attack	A malicious site sends harmful updates, a server alters the shared model, or the shared intermediate data is decoded back into records.	Multi-institution imaging, shared EHR models, split learning, and edge systems	False labels combined with a deliberately reversed update reduced a simulated brain-MRI model from 61% to 23% accuracy. Other malicious-server studies reconstructed most of the tested images or text from shared updates (Soulb et al., 2025; Shi et al., 2025). Collaborative learning reduces central data collection but adds new trust risks.

3.9 Defense strategies

A defense should be judged on three things: how much it improves performance under the tested attack, whether it reduces performance on ordinary clinical data, and whether it still works when the attacker knows the defense is present and designs the attack around it.

One defense is adversarial training, which is loosely analogous to building up immunity in advance. A model normally learns only from clean images, which leaves blind spots that tiny manipulations can exploit. In adversarial training, the team keeps generating attacked versions of the training images and shows them to the model together with the correct answer, so the model is repeatedly forced to give the right diagnosis even when the image has been tampered with; over many rounds it learns to see past those manipulations.

Neethunath et al. (2025) tried this on five model designs (architectures), using a dataset of 3,297 dermoscopic images, of which 660 were held out for testing. Without the defense, all five were accurate on ordinary images (85.15% to 89.55%) but collapsed under a strong test attack, called Auto-CE, scoring only 0% to 12.73%. After adversarial training (the attacked training images were generated with a standard method called PGD), the same attack was run again. For one model, ConvNeXt, the defense worked: it stayed accurate on ordinary images (82.27%) and this time held 73.79% under the attack instead of collapsing. For another, MobileNetV2, the same defense barely helped: under attack it still scored only 0.15%. So adversarial training can genuinely protect a model, but how much depends heavily on the model design.

UI Haque et al. (2025) combined three defenses on lung-nodule CT scans: adversarial training, plus two image-cleaning steps applied to each scan before the model reads it. The first cleaning step smooths away the fine noise the attacker adds while keeping the real anatomy (a technique called total variation minimisation), and the second (called bit-plane slicing) separates the image into levels of brightness from coarse to very fine, like separating a picture's bold shapes from its faintest shading; because the attacker's changes are so slight that they sit almost entirely in the very finest level, this step keeps the higher levels that carry the coarse anatomy and drops the lower levels that are most prone to small adversarial changes, removing the hidden tampering along with them.

Without any defense, accuracy fell from 93.17% on ordinary images to 13.73% and 1.32% under two standard attacks (FGSM and PGD). With all three defenses in place, accuracy under attack was restored to about 93%, and accuracy on ordinary images stayed high at 92.58%. As in the previous study, not every model design was tested, and the combined defense was not tested against an attacker who had redesigned the attack specifically to get past it.

Detection and rejection methods try to spot a suspicious input and set it aside before its prediction is used in care. Ma et al. (2021) built such a detector and, across four attacks on fundus, chest-radiograph, and dermoscopic models, it told attacked images from clean ones very reliably (an AUC above 98%, where AUC measures how well the detector separates the two and 100% would be perfect).

Joseph, George, et al. (2026) tested a tool called Sentinel on four medical-image datasets. Sentinel sits between the images and the diagnostic model like a gatekeeper, and it needs no retraining of the model. To spot an attack it makes several simplified versions of the image and checks whether the model's answer stays steady across them: a clean image gives a stable answer, whereas a tampered one makes the answer wobble, which flags the attack and even reveals its type. For attacks hidden in the image's fine detail, a built-in denoising step then removes the perturbation while keeping the real anatomy, so the image can still be read. It detected more than 90% of successful attacks. For that fine-detail family it recovered the correct label about 95% of the time; for ordinary attacks that use the model's own gradients it recovered about 70%, and the cleaned-up images were less true to the originals. Processing took about 80 to 95 milliseconds per image on an ordinary computer processor (a CPU).

W. J. Wang et al. (2020) tested a detector called RADAR on a mortality model (one that predicts a patient's risk of dying) built on MIMIC-III, a large public intensive-care dataset. Set so that 95% of ordinary records passed through, it caught more than 90% of the attacked records. When it flags an attack, RADAR simply flips the model's answer to the opposite one, which works here because the model has only two options (survive or die) but would not carry over to a diagnosis with many possible categories. The authors also tested a distribution-matched attack designed to be harder for the detector to spot, though it was not a fully adaptive attack optimized against every part of RADAR.

Detection can also target the data itself, to catch poisoning before it is baked into a model. Lourdu Mahimai Doss et al. (2026) built an automated checker that scanned training images and correctly flagged most of the poisoned ones (up to about 96%), even though the tampering was invisible to the eye rather than an obvious wrong label. Mithsara et al. (2024) took general-purpose chatbots (such as ChatGPT and Gemini)

and, without any special training, asked them to look at wearable-sensor data and pick out the readings that had been tampered with; ChatGPT-4 did this notably well, but only on a very small amount of data, so the result cannot yet support routine clinical use.

On its own, input purification tries to remove a suspected attack from an image before it reaches the diagnostic model. He et al. (2024) tested an AI cleaning method called ADDM, which uses a diffusion model (the kind of generative AI used to create images) to rebuild a clean version of the picture, on brain MRI and pneumonia chest radiographs. Two standard attacks (BIM and PGD) had reduced the undefended model's accuracy to 0%. After cleaning, accuracy recovered to 71.07% to 71.32% for brain MRI and 83.33% to 83.81% for pneumonia. The cleaning came at a small cost, though: on ordinary images it lowered accuracy by 5.45 to 7.36 percentage points compared with the untouched originals, and the method did not carry over well to a different dataset.

Model hardening changes the model itself to make it harder to fool. Chen et al. (2021) made two such changes to a standard image model (a ResNet34): they stripped out the model's least useful internal parts, and they added a component (called attention) that helps it focus on the meaningful parts of an image. On chest radiographs this barely affected ordinary accuracy, which slipped from 88.71% to 86.19%, but it made the model much harder to attack. A PGD attack (a gradient-based method), which had fooled the original model 100% of the time, now succeeded only 22.82% of the time, and a DeepFool attack only 30.51%. One caveat is that these attacks were built to fool the original model and then reused on the hardened one, rather than being specially redesigned to beat it.

Privacy defenses tackle a different problem from defenses against malicious inputs: their job is to stop the model leaking the data it was trained on. Krall et al. (2020) added carefully measured random noise during training, a standard privacy technique. The amount of noise is set by a privacy budget, where a smaller number means stronger privacy. At a very strict budget of 0.0001, a model-inversion attack (which here tried to infer a patient's sensitive attribute from the released regression model, scored by R-squared) was reduced to zero, while the model's accuracy stayed close to what it reached with no privacy noise at all, though the paper did not give one exact accuracy figure for that setting.

J. Wang et al. (2026) tested a privacy tool called HealthGuard on wearable activity and stress data. It works in two parts. On the wearable itself it adds carefully calibrated random noise to the readings, but not evenly: it puts the most noise on the rapid, most telling changes in the body (which carry the most sensitive information) and less on routine readings, so an attacker cannot infer private facts while the data stays medically meaningful. On the server, a smart reconstruction step then cleans the noised stream back into a clinically usable signal while still blocking anyone trying to recover the true underlying pattern. This cut the success of a membership-inference attack (working out whether a person's data was used to train the model) to 8.7%, while keeping 94.2% of the model's usefulness. Its extra cost on the wearable device was tiny in simulation: about 2 milliseconds of time and 38 millijoules of energy per 1,000 readings. Privacy methods like this protect confidentiality, but on their own they do nothing to stop poisoning, backdoors, or attacks that simply disrupt the service.

Certified robustness offers something stronger than testing: a mathematical proof. Instead of just trying some attacks and seeing whether the model holds, it proves that the model's answer cannot change as long as the input is altered by less than a fixed amount. That proven safe margin is the certificate. The catch is that the proof is narrow: it covers only one specific model and one specific kind of attack, and it says nothing about changes beyond the safe margin or about attacks of a different kind. Dietrich et al. (2025) add that in practice this safe margin is usually very small, which can limit its clinical relevance, so a certificate can look reassuring while actually guaranteeing very little. They also note that such formal methods have rarely been studied in radiological settings. A certificate is therefore only worth trusting if it states exactly what it assumes and how big its safe margin really is, and it should never be presented as complete protection.

Some work focuses on how to test systems before deployment rather than on a specific safeguard. Trabilsy et al. (2025) proposed a four-stage clinical red-teaming process (deliberately trying to make a system fail before it is used): plan the test, prepare realistic malicious prompts, run them, then evaluate safety and clinical appropriateness. It was offered as a structured framework and was still undergoing validation, so it should be seen as a proposed testing process rather than a proven safeguard.

The most important warning comes from adaptive testing. An adaptive attacker knows how the defense works and designs the attack specifically to get around it. Levy et al. (2024) did exactly this. They took five published medical-imaging defenses (mostly detectors that flag adversarial images, plus a denoiser and a few other add-on methods) and designed attacks specifically to defeat each one, across six datasets and eight

defense-and-task combinations. Against these adaptive attacks, classification accuracy still fell by about 50% on average, and segmentation quality (measured by the Dice score, which rates how closely the model's outline matches the true one) fell by about 30%. In three settings, the defended model's performance collapsed to between 0% and 4%. The authors could not rebuild two of the defenses exactly and had to use similar model designs instead, so the study does not prove that every version of these defenses would fail. But it does show why testing a defense only against the attacks its own developers chose can give false reassurance.

A further set of studies proposed or tested defenses beyond those above. Anand et al. (2020) compared self-supervision with transfer learning as routes to more robust biomedical image analysis, while Ahmed et al. (2022) used deep active learning to reduce the impact of evasion attacks. Singh et al. (2025) presented a lightweight two-phase adversarial-training model for radiology images, S. Rahman et al. (2025) proposed a transformer-based detector to catch adversarial attacks on Internet-of-Medical-Things image analysis, and Mohammed et al. (2025) compared nine classifiers within a trusted framework for orthopedic disease detection under adversarial conditions. Xu et al. (2025) found that uncertainty-aware fine-tuning gave only limited protection for diabetic-retinopathy screening, with attacks still succeeding, and others presented further defenses for medical images (Jeribi et al., 2025; Tsai et al., 2025).

Other proposed defenses targeted specific settings: SAFe-Health used a federated defense-averaging scheme against data poisoning in smart-healthcare federated learning (Mali et al., 2024), FedKC combined personalised federated learning with robustness to model poisoning (L. Sun et al., 2024), ECG-ATK-GAN used a conditional generative adversarial network to make ECG classifiers more robust to adversarial attacks (Hossain et al., 2022), and a graph-attention WideResNet detector was proposed for adversarial attacks in wireless body-area networks (Pipal & Kannan, 2026). Their differing datasets and outcome measures cannot be combined into one estimate of effectiveness.

Several studies focused specifically on detecting attacks rather than resisting them. X. Li and Zhu (2020) and Kishan and Kumar (2026) built detectors that flag adversarial medical images before diagnosis, and Albattah and Rassam (2023) did the same for a combined CNN-and-LSTM model used in healthcare monitoring. de Aguiar et al. (2024) instead used model explanations to expose manipulated inputs, while Vali et al. (2023) tested both attacks and their detection on a digital-pathology model. The main text focuses on detection studies that report the attack, the measured detection result, and its limitations more fully.

It is fair to claim that a defense improved a specific, measured result against a specific named attack; it is not fair to conclude from that alone that the whole system is secure. To let readers judge a defense properly, an evaluation should report separately: accuracy on ordinary inputs, accuracy under attack, how often it raises false alarms on clean inputs, its computational cost, what access the attacker was assumed to have, and how it held up against attacks specifically adapted to beat it.

Table 6. Defense strategies, representative results, and practical limitations.

Defense	What it does	Representative result	Main limitation
Adversarial training and architecture choice	Trains the model on deliberately altered examples and compares which model designs retain performance.	After adversarial training (PGD), ConvNeXt reached 73.79% accuracy under the Auto-CE attack, while MobileNetV2 reached 0.15% (Neethunath et al., 2025).	Expensive training; possible loss on ordinary data; result depended strongly on architecture and one dataset.
Detection and rejection	Flags a suspicious input before its prediction is used.	The image detector reliably separated attacked from clean images (AUC above 98%); Sentinel and RADAR detected more than 90% of the tested attacks.	False alarms require local calibration. Tests that are not adapted to the detector may overestimate protection.

Defense	What it does	Representative result	Main limitation
Input purification	Attempts to remove the attack from an image before classification.	ADDM restored accuracy from 0% to about 71% for brain MRI and about 83% for pneumonia radiographs (He et al., 2024).	Accuracy on ordinary processed images fell by 5.45 to 7.36 percentage points, and transfer to another dataset was weak.
Model hardening	Changes the model, for example by removing less important internal parts and adding an attention component.	Chest-radiograph PGD attack success fell from 100% to 22.82% (Chen et al., 2021).	Ordinary accuracy also fell, and the attacks were transferred from the original model rather than redesigned for the defense.
Combined safeguards	Uses more than one protection at training and input time.	A CT study raised accuracy under two attacks (FGSM and PGD) from 13.73% and 1.32% to about 93% (Ul-Haque et al., 2025).	One dataset, two full-access (white-box) attacks, and no adaptive test of the final combined defense.
Federated integrity defenses	Monitors or limits client updates before they affect the shared model.	HyperShield still separated malicious updates from honest ones well (AUROC at least 0.82) even with up to 50% malicious clients (Dao et al., 2026).	Public-data simulation rather than real multi-hospital deployment; protection still depends on server trust.
Privacy defenses	Adds controlled noise or limits the information released by the system.	HealthGuard reduced membership-inference success to 8.7% while retaining 94.2% of usefulness (J. Wang et al., 2026).	Privacy and performance must be balanced; these methods do not prevent poisoning, backdoors, or service disruption.
Certified robustness	Provides a mathematical guarantee only for one stated type of attack and a limited amount of input change (a proven safe margin).	No single pooled percentage is meaningful; each certificate applies only within its own reported safe margin.	The guaranteed safe margin may be small, and the certificate does not cover attacks outside its assumptions.

4. Discussion

4.1 Main lessons across threat types

The first recurring lesson is that protection can reduce performance on the intended clinical task. In the genomic study by C. H. Zhang and Bonomi (2022), a privacy safeguard reduced a membership attack to no better than a coin toss (its success measure, the AUC, fell from 0.71 to 0.50), but the model's own accuracy also dropped sharply, from 89% to 58%. Roth et al. (2022) found a smaller trade-off in a split-learning model (Split-U-Net): a moderate amount of privacy noise cost little accuracy, but a large amount of noise almost halved the model's ability to outline structures in the scan, with the Dice score falling from about 0.70 to 0.39 (where 1.0 is a perfect outline). He et al. (2024) recovered much of the accuracy lost under attack, but the cleaning step lowered accuracy on ordinary images.

These examples do not mean that every defense causes the same loss. They mean that security, privacy, and clinical performance must be measured together for the specific model and task.

The second lesson is that experimental assumptions strongly affect what a result means. White-box imaging studies give the attacker extensive inside access to the model. They are useful for worst-case stress testing but may be less realistic for an outside attacker (Dietrich et al., 2025; Levy et al., 2024). Black-box and closed-box studies allow the attacker far less access, but even then the attack's feasibility depends on how

many queries are allowed and how much the system reveals in its answers (Joseph, Jeffrey, et al., 2026; W. Y. Li et al., 2025).

Each threat also requires a different route into the system. Poisoning requires access to training data or model updates. Prompt injection requires control over content that the model reads. The federated reconstruction studies by El Azzouzi et al. (2026) and Shi et al. (2025) require a malicious server able to alter the model. Practical risk therefore depends on the real data flow, user interface, access controls, and monitoring of the deployed system.

The third lesson is that connected components create connected risks. A foundation model (a large general-purpose model built once and then reused) may later be adapted to several tasks, though harmful behavior does not necessarily survive every adaptation, filtering, or validation step (Hanif et al., 2024; Alber et al., 2025). A multimodal model may take in both images and text. A language model may also read documents fetched for it or call on external tools (Huang et al., 2025; Lu et al., 2026; Akinci D'Antonoli et al., 2025).

Every additional input or component creates another place where untrusted content or data could enter. This does not mean that every AI system has every vulnerability. It means that evaluation should follow the actual path from data collection and training, through user input and retrieval, to model output and any downstream clinical action.

4.2 Clinical and regulatory implications

Clinical use is what makes these weaknesses matter. If a manipulated output reaches care without an effective check, it could contribute to a missed finding, an incorrect classification, or unsafe advice. However, this review did not establish how often attacks occur in clinical practice or how often they cause patient harm. The included experiments demonstrate technical failure modes under defined conditions.

Reviews describe possible motives such as reimbursement fraud, sabotage, or manipulation of automated decisions (Kaviani et al., 2022; Sorin et al., 2023). Abtahi et al. (2026) also argue that testing whether a tool resists deliberate attacks (adversarial-robustness testing) is not yet a routine part of clinical validation or regulatory pathways. Pre-market and post-market evaluation should therefore reflect the intended use and the attack routes that are realistic for that particular system.

Ordinary accuracy testing and security testing answer different questions. Accuracy on an ordinary test set asks, "How often is the model correct on data similar to the evaluation data?" Security testing asks, "What happens when someone deliberately changes the input, training data, shared update, or instruction?"

A practical assessment should specify who might attack the system, what that person could access, what they could change, and what outcome they might seek. Safeguards can then be chosen for that route. Examples include recording where training data came from, checking for unauthorised changes, limiting and monitoring model queries, screening federated updates, testing for hidden triggers, keeping untrusted instructions separate from trusted clinical content, and requiring independent review before an AI output changes care.

The safeguards should also be tested against an attacker who knows they are present, as illustrated by Levy et al. (2024).

4.3 Limitations of the evidence and of this review

The evidence base has major limitations. Most studies used public datasets, a limited range of model families, and digitally created attacks in laboratory settings. They do not tell us how often attacks occur in live clinical workflows or whether they have harmed patients.

The studies also measured different outcomes. Some reported accuracy, some attack success, and others AUROC, Dice, retrieval quality, privacy leakage, or selected examples. These results cannot be combined or ranked without a common study design. Some defenses were tested only against attacks chosen by the defense developers, and the adaptive results of Levy et al. (2024) show why that can overestimate protection. We did not formally assess publication bias (the tendency for studies with striking results to get published more often), so we make no claim about its size or direction.

This review also has methodological limitations. It was not registered in advance, and it did not include a formal check of each study's risk of bias or a statistical combining of the results (pooling). We also did not document having two reviewers independently screen the studies and extract the data, or a preset way of resolving disagreements between reviewers.

The database searches were not identical. PubMed began at 2020, while the Scopus search required a review or survey term. Broad exclusion terms may have removed some publications that combined model security with wider infrastructure security. Only English-language publications were included.

Because this is a qualitative narrative review rather than a quantitative meta-analysis, it can describe attack mechanisms, study-specific results, and recurring limitations, but it cannot estimate how often attacks occur, rank all defense families, or identify a single universally effective safeguard. The field is also developing very quickly: specific models, tools, and results date rapidly, so the findings here should be read as a snapshot of a fast-moving area rather than a settled account.

4.4 Priorities for future research

First, the field needs shared benchmarks that reflect clinical use. Each benchmark should state the task, dataset, model, attacker access, permitted changes, number of allowed queries, ordinary baseline performance, and whether the attack was adapted to the defense. It should report both the harm caused by the attack and the cost of the defense. Levy et al. (2024) show why a fixed attack selected by the defense developer is not enough.

Second, testing should follow the components that the real system contains. Data pipelines need records of data origin and checks for unauthorised changes. Federated systems need tests with malicious clients and servers, as well as tests for leakage from shared updates. Public-facing model interfaces need realistic query limits and monitoring. Safeguards should be selected from the stated threat model (who might attack the system and how), not applied as one universal checklist (Abtahi et al., 2026; Dao et al., 2026).

Third, LLM and multimodal systems need separate tests for direct user instructions, untrusted retrieved documents, conflicting text and images, and multilingual prompts. Prompt filtering and the PREE red-teaming cycle (a proposed process for deliberately stress-testing a system before use) are proposed safeguards or testing frameworks, not clinically validated solutions (Akinci D'Antonoli et al., 2025; Trabilsy et al., 2025).

Comparative studies should measure not only attack success, but also how often a safeguard wrongly blocks genuine clinical content (its false-alarm rate), whether results change between model versions, and whether later fine-tuning removes or reintroduces the weakness.

Fourth, practical systems need several independent safeguards. Technical controls, access limits, monitoring, and qualified clinical review protect different parts of the workflow. Human oversight should not be treated as automatic protection; both what it adds and its own ways of failing should be tested.

A system should be reassessed whenever its data source, model version, retrieval component, external tool, or user interface changes in a clinically meaningful way.

5. Conclusions

This structured narrative review of 113 publications found experimentally demonstrated weaknesses at several points in healthcare AI systems. Deliberately altered inputs changed predictions. Poisoned data and backdoors changed what models learned. The model's answers, and the numerical updates shared during collaborative training, could expose private information. Malicious instructions changed the responses of language and multimodal models.

Several defenses improved results in individual experiments, but no defense was shown to work across all threats and settings. In one medical-imaging study, attacks designed specifically around five defenses bypassed all five (Levy et al., 2024).

The evidence does not show how often these attacks occur in clinical practice or how often they harm patients. Evaluation should therefore match the intended workflow and a clearly stated threat model (who might attack and how). When relevant, it should report performance on both ordinary and attacked inputs, together with privacy, computational, and clinical trade-offs. Testing should follow the attack routes created by the system's actual components and should be repeated after important changes to data, models, or interfaces.

Supplementary Materials: Appendix A. Database search strategies

Author Contributions: Conceptualization, methodology, literature search, data extraction and verification, writing of the original draft and review and editing were performed by all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This is a review of previously published literature and did not involve new studies on humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data discussed in this review are available in the cited published articles.

Acknowledgements: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Abd El-Aziz, A. A., El-Khoribi, R. A., & Khalifa, N. E. M. (2026). Adversarial threats in health care: A comprehensive analysis of vulnerabilities, defense mechanisms, and recent research. In *Next-gen healthcare: AI-powered medical innovations* (pp. 3–39). Springer Nature. https://doi.org/10.1007/978-3-032-07267-2_1
2. Abtahi, F., Seoane, F., Pau, I., & Vega-Barbas, M. (2026). Data poisoning vulnerabilities across health care artificial intelligence architectures: Analytical security framework and defense strategies. *Journal of Medical Internet Research*, 28, e87969. <https://doi.org/10.2196/87969>
3. Ahmed, U., Lin, J. C. W., & Srivastava, G. (2022). Mitigating adversarial evasion attacks by deep active learning for medical image classification. *Multimedia Tools and Applications*, 81(29), 41899–41910. <https://doi.org/10.1007/s11042-021-11473-z>
4. Akhtom, D., Singh, M. M., & Xinying, C. (2024). Unveiling hidden variables in adversarial attack transferability on pre-trained models for COVID-19 diagnosis. *International Journal of Advanced Computer Science and Applications*, 15(11), 1343–1350. <https://doi.org/10.14569/ijacsa.2024.01511131>
5. Akinci D'Antonoli, T., Tejani, A. S., Khosravi, B., Bluethgen, C., Busch, F., Bressemer, K. K., Adams, L. C., Moassefi, M., Faghani, S., & Gichoya, J. W. (2025). Cybersecurity threats and mitigation strategies for large language models in health care. *Radiology: Artificial Intelligence*, 7(4), Article e240739. <https://doi.org/10.1148/ryai.240739>
6. Al-Dhabyani, W. (2026). Data privacy and security in large language models for medical fields. In *Next-gen healthcare: AI-powered medical innovations* (pp. 175–197). Springer Nature. https://doi.org/10.1007/978-3-032-07267-2_7
7. Albattah, A., & Rassam, M. A. (2023). Detection of adversarial attacks against the hybrid convolutional long short-term memory deep learning technique for healthcare monitoring applications. *Applied Sciences*, 13(11), Article 6807. <https://doi.org/10.3390/app13116807>
8. Alber, D. A., Yang, Z., Alyakin, A., Yang, E., Rai, S., Valliani, A. A., Zhang, J., Rosenbaum, G. R., Amend-Thomas, A. K., Kurland, D. B., Kremer, C. M., Eremiev, A., Negash, B., Wiggan, D. D., Nakatsuka, M. A., Sangwon, K. L., Neifert, S. N., Khan, H. A., Save, A. V., ... Oermann, E. K. (2025). Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2), 618–626. <https://doi.org/10.1038/s41591-024-03445-1>
9. Alissa, K. A. (2026). A survey on privacy issues and mitigation strategies for LLMs in healthcare. *The Journal of Supercomputing*, 82(1), Article 26. <https://doi.org/10.1007/s11227-025-08146-1>
10. Allyn, J., Allou, N., Vidal, C., Renou, A., & Ferdynus, C. (2020). Adversarial attack on deep learning-based dermatoscopic image recognition systems: Risk of misdiagnosis due to undetectable image perturbations. *Medicine*, 99(50), Article e23568. <https://doi.org/10.1097/md.00000000000023568>
11. An, S., Xiao, C., Stewart, W. F., & Sun, J. M. (2019). LAVA: Longitudinal adversarial attack on electronic health records data. In *The World Wide Web Conference (WWW 2019)*. Association for Computing Machinery.
12. Anand, D., Tank, D., Tibrewal, H., & Sethi, A. (2020). Self-supervision vs. transfer learning: Robust biomedical image analysis against adversarial attacks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI 2020)*. IEEE.
13. Apostolidis, K. D., & Papakostas, G. A. (2021). A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17), Article 2132. <https://doi.org/10.3390/electronics10172132>
14. Aroua, R., Kammoun, I., Louati, M. A., Adjero, D. A., & Duan, T. (2026). Recent advances in defending the privacy attacks of large language models for healthcare applications: A concise review. *Frontiers in Artificial Intelligence*, 9, Article 1816692. <https://doi.org/10.3389/frai.2026.1816692>
15. Ayensu, F., Turner, C., & Osunmakinde, I. (2025). Enhancing federated learning security with a defense framework against adversarial attacks in privacy-sensitive healthcare applications. *International Journal of Advanced Computer Science and Applications*, 16(5), 1–13. <https://doi.org/10.14569/IJACSA.2025.0160501>

16. Chen, L., Chen, Y., Ouyang, Z., Dou, H., Zhang, Y., & Sang, H. (2025). Boosting adversarial transferability in vision-language models via multimodal feature heterogeneity. *Scientific Reports*, 15(1), Article 7366. <https://doi.org/10.1038/s41598-025-91802-6>
17. Chen, L., Zhao, L., & Chen, C. Y. (2021). Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism. *Medical Physics*, 48(10), 6198–6212. <https://doi.org/10.1002/mp.15208>
18. Chuang, Y. S., Sarkar, A. R., Hsu, Y. C., Mohammed, N., & Jiang, X. (2025). Robust privacy amidst innovation with large language models through a critical assessment of the risks. *Journal of the American Medical Informatics Association*, 32(5), 885–892. <https://doi.org/10.1093/jamia/ocaf037>
19. Clusmann, J., Ferber, D., Wiest, I. C., Schneider, C. V., Brinker, T. J., Foersch, S., Truhn, D., & Kather, J. N. (2025). Prompt injection attacks on vision language models in oncology. *Nature Communications*, 16(1), Article 1239. <https://doi.org/10.1038/s41467-024-55631-x>
20. Dao, B. T., Ma, V. D., Vu, T. A., Nguyen, H. T., Tran, K. P., & Truong, T. H. (2026). HyperShield: A personalized hypernetwork-based defense against poisoning attacks in federated learning for healthcare. *Expert Systems with Applications*, 299, Article 130237. <https://doi.org/10.1016/j.eswa.2025.130237>
21. Darzi, E., Dubost, F., Sijtsma, N. M., & van Ooijen, P. M. A. (2024). Exploring adversarial attacks in federated learning for medical imaging. *IEEE Transactions on Industrial Informatics*, 20(12), 13591–13599. <https://doi.org/10.1109/tii.2024.3423457>
22. Das, A., Tariq, A., Batalini, F., Dhara, B., & Banerjee, I. (2024). Exposing vulnerabilities in clinical LLMs through data poisoning attacks: Case study in breast cancer. *AMIA Annual Symposium Proceedings*, 2024, 339–348.
23. de Aguiar, E. J., Traina, C., Jr., & Traina, A. J. M. (2024). RADAR-MIX: How to uncover adversarial attacks in medical image analysis through explainability. In *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS 2024)*. IEEE.
24. Dietrich, N., Gong, B., & Patlas, M. N. (2025). Adversarial artificial intelligence in radiology: Attacks, defenses, and future considerations. *Diagnostic and Interventional Imaging*, 106(11), 375–384. <https://doi.org/10.1016/j.diii.2025.05.006>
25. Dong, J., Chen, J., Xie, X., Lai, J., & Chen, H. (2024). Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3), Article 79. <https://doi.org/10.1145/3702638>
26. El Azzouzi, M., Bellafqira, R., Coatrieux, G., Cuggia, M., & Bouzille, G. (2026). Recovering sensitive medical text in federated learning. *Studies in Health Technology and Informatics*, 336, 1466–1470. <https://doi.org/10.3233/shti260452>
27. Faraoun, H., Bellafqira, R., Coatrieux, G., & Kallas, K. (2026). A durable backdoor attack on medical imaging via federated learning. *Studies in Health Technology and Informatics*, 336, 600–604. <https://doi.org/10.3233/shti260241>
28. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G. P., Han, T., van Treeck, M., Buelow, R. D., Langer, R., Dislich, B., Boor, P., Schulz, V., & Kather, J. N. (2022). Adversarial attacks and adversarial robustness in computational pathology. *Nature Communications*, 13(1), Article 5711. <https://doi.org/10.1038/s41467-022-33266-0>
29. Güler, I., Kraus, A., Grieb, G., Satir, T., Eberz, P., & Stelling, H. (2026). AI in dermato-oncology: Diagnostic performance and prompt-injection vulnerability of vision-language models in dermoscopic skin cancer assessment. *Cancers*, 18(11), Article 1750. <https://doi.org/10.3390/cancers18111750>
30. Hanif, A., Shamshad, F., Awais, M., Naseer, M., Khan, F. S., Nandakumar, K., Khan, S., & Anwer, R. M. (2024). BAPLe: Backdoor attacks on medical foundational models using prompt learning. In *Medical image computing and computer assisted intervention – MICCAI 2024 (Part XII)*. Springer.
31. Harada, Y. (2026). Robustness of a large language model (LLM)-based virtual patient for Japanese history-taking training under direct and indirect instructional contamination. *Cureus*, 18(5), Article e109161. <https://doi.org/10.7759/cureus.109161>
32. He, Y. M., Pang, S. C., Du, A. N., Chen, H. C., Cong, L. L., & Orgun, M. M. (2024). ADDM: Adversarial defenses with diffusion model for medical imaging data mining. In *Advanced data mining and applications (ADMA 2024) (Part IV)*. Springer.
33. Hossain, K. F., Kamran, S. A., Tavakkoli, A., & Ma, X. J. (2022). ECG-ATK-GAN: Robustness against adversarial attacks on ECGs using conditional generative adversarial networks. In *Applications of medical artificial intelligence (AMAI 2022)*. Springer.
34. Huang, X. J., Wang, X. Y., Zhang, H. T., Zhu, Y. H., Xi, J. W., An, J. K., Wang, H., Liang, H., & Pan, C. W. (2025). Medical MLLM is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI-25) (Vol. 39, No. 4)*. AAAI Press.
35. Imam, N. H. (2024). Adversarial examples on XAI-enabled digital twin for smart healthcare systems. *Sensors*, 24(21), Article 6891. <https://doi.org/10.3390/s24216891>
36. Jeribi, F., Tahir, A., Alsabi, K. H., Ramakrishnan, J., & Rana, N. (2025). Defense against adversarial attack on medical imaging. In *AI revolution: Research, ethics and society (AIR-RES 2025) (Part II)*. Springer.
37. Jin, R., & Li, X. (2023). Backdoor attack and defense in federated generative adversarial network-based medical image synthesis. *Medical Image Analysis*, 90, Article 102965. <https://doi.org/10.1016/j.media.2023.102965>

38. Joe, B., Park, Y., Hamm, J., Shin, I., & Lee, J. (2022). Exploiting missing value patterns for a backdoor attack on machine learning models of electronic health records: Development and validation study. *JMIR Medical Informatics*, 10(8), Article e38440. <https://doi.org/10.2196/38440>
39. Joseph, N., George, S. N., Ameer, P. M., & Raja, K. (2026). Securing medical AI: Adversarial attack detection and mitigation in CAD systems. *IEEE Transactions on Emerging Topics in Computational Intelligence*. Advance online publication. <https://doi.org/10.1109/tetci.2026.3692606>
40. Joseph, N., Jeffrey, D., George, S. N., Ameer, P. M., & Raja, K. (2026). Stealth-black-box attack on Industry 4.0 medical AI: A low-rank perturbation approach. *IEEE Transactions on Industrial Informatics*, 22(3), 2458–2468. <https://doi.org/10.1109/tii.2025.3640725>
41. Joshi, S., Mehta, M. A., Mendoza, M., Rivera, Y. M., & Singh, V. K. (2026). Differences in safety risks across languages in health-relevant queries: Vulnerability analysis of large language model responses. *JMIR Formative Research*, 10, Article e87465. <https://doi.org/10.2196/87465>
42. Kalpana, R. N. L. S., Patro, A. K., & Rao, D. N. (2025). Survey on robustness of deep learning techniques on adversarial attacks in WBAN. In *Lecture notes in electrical engineering*. Springer.
43. Kaviani, S., Han, K. J., & Sohn, I. (2022). Adversarial attacks and defenses on AI in medical imaging informatics: A survey. *Expert Systems with Applications*, 198, Article 116815. <https://doi.org/10.1016/j.eswa.2022.116815>
44. Kishan, K., & Kumar, B. V. (2026). Adversarial attack detection in secure medical images using deep learning. *Information Security Journal: A Global Perspective*, 35(4), 601–615. <https://doi.org/10.1080/19393555.2025.2547206>
45. Krall, A., Finke, D., & Yang, H. (2020). Gradient mechanism to preserve differential privacy and deter against model inversion attacks in healthcare analytics. In *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'20)*. IEEE.
46. Krzton, K., Kolodziej, J., Widlak, A., Nawrocki, M., & Sigut, J. (2025). Vulnerabilities of machine learning algorithms to adversarial attacks in medical images. In *Proceedings of the 39th ECMS International Conference on Modeling and Simulation (ECMS 2025)*. European Council for Modeling and Simulation.
47. Kwon, H., & Jeong, J. (2022). AdvU-Net: Generating adversarial example based on medical image and targeting U-Net model. *Journal of Sensors*, 2022, Article 4390413. <https://doi.org/10.1155/2022/4390413>
48. Lee, R. W., Jun, T. J., Lee, J. M., Cho, S. I., Park, H. J., & Suh, J. (2025). Vulnerability of large language models to prompt injection when providing medical advice. *JAMA Network Open*, 8(12), Article e2549963. <https://doi.org/10.1001/jamanetworkopen.2025.49963>
49. Lee, S. B. (2025). Gradual poisoning of a chest x-ray convolutional neural network with an adversarial attack and AI explainability methods. *Scientific Reports*, 15(1), Article 21779. <https://doi.org/10.1038/s41598-025-02294-3>
50. Levy, M., Amit, G., Elovici, Y., & Mirsky, Y. (2024). The security of deep learning defenses in medical imaging. In *Proceedings of the 2024 Workshop on Cybersecurity in Healthcare (HealthSec 2024)*. Association for Computing Machinery.
51. Li, W., & Zainon, W. (2026). Dual adversarial attacks on explainable deep learning in medical image classification. *Computer Methods and Programs in Biomedicine*, 273, Article 109125. <https://doi.org/10.1016/j.cmpb.2025.109125>
52. Li, W. Y., Zhang, Z., Lan, X. Y., & Wang, Y. W. (2025). Untargeted closed-box attack against healthcare image retrieval via rank manipulation. *IEEE Transactions on Consumer Electronics*, 71(2), 4544–4555. <https://doi.org/10.1109/tce.2025.3559103>
53. Li, X., & Zhu, D. X. (2020). Robust detection of adversarial attacks on medical images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI 2020)*. IEEE.
54. Liyanage, S., Weerawardhane, V., Kheminda, J., Siriwardhana, Y., Weerasinghe, T., & Liyanage, M. (2025). SecFLH: Defending federated learning-based IoT health prediction systems against poisoning attacks. In *2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE.
55. Lourdu Mahimai Doss, P., Gunasekaran, M., Padmakala, S., Samikannu, R., Yahya, A., Badruddin, I. A., Kamangar, S., & Rohini, R. (2026). Invisible poison attacks and holistic defense strategy for pyramid vision transformers in medical imaging. *Scientific Reports*. Advance online publication. <https://doi.org/10.1038/s41598-026-58050-8>
56. Lu, Z. M., Xu, N., Tian, H. S., Wang, L. J., & Liu, A. A. (2026). Medical VLP model is vulnerable: Toward multimodal adversarial attack on large medical vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2), 2478–2491. <https://doi.org/10.1109/tcsvt.2025.3602970>
57. Ma, X. J., Niu, Y. H., Gu, L., Wang, Y., Zhao, Y. T., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110, Article 107332. <https://doi.org/10.1016/j.patcog.2020.107332>
58. Mali, B., Singh, P. K., & Mazumdar, N. (2024). SAFe-Health: Guarding federated learning-driven smart healthcare with federated defense averaging against data poisoning. *Security and Privacy*, 7(5), Article e403. <https://doi.org/10.1002/spy2.403>

59. Malik, A., Qaisar, S. M., Khan, M. Z., Khan, M. N., Aldhyani, T. H. H., & Alsisi, R. H. (2026). An efficient framework for detecting model inversion attacks in connected healthcare systems. *IEEE Transactions on Consumer Electronics*, 72(2), 4528–4539. <https://doi.org/10.1109/tce.2026.3658570>
60. Menon, K., Bohra, V. K., Murugan, L., Jaganathan, K., & Arumugam, C. (2021). COVID-19 diagnosis from chest X-ray images using convolutional neural networks and effects of data poisoning. In *Computational science and its applications (ICCSA 2021)* (Part IX). Springer.
61. Mithsara, W. K. M., Shahid, A. R., & Yang, N. (2024). Zero-shot detection and sanitization of data poisoning attacks in wearable AI using large language models. In *2024 International Conference on Machine Learning and Applications (ICMLA)*. IEEE.
62. Mohammed, T. J., Xinying, C., Albahri, A. S., Alnoor, A., & Wah, K. K. (2025). A novel trusted framework for orthopedic disease detection with reliability against adversarial attacks. *IEEE Access*, 13, 160193–160220. <https://doi.org/10.1109/access.2025.3609330>
63. Mondal, I. (2021). BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*. Association for Computational Linguistics.
64. Mondillo, G., Colosimo, S., Perrotta, A., Frattolillo, V., Indolfi, C., del Giudice, M. M., & Rossi, F. (2025). Jailbreaking large language models: Navigating the crossroads of innovation, ethics, and health risks. *Journal of Medical Artificial Intelligence*, 8, Article 6. <https://doi.org/10.21037/jmai-24-170>
65. Muoka, G. W., Yi, D., Ukwuoma, C. C., Mutale, A., Ejayi, C. J., Mzee, A. K., Gyarteng, E. S. A., Alqahtani, A., & Al-antari, M. A. (2023). A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. *Mathematics*, 11(20), Article 4272. <https://doi.org/10.3390/math11204272>
66. Neethunath, M. R., Gladston, R. S., & Pradeepan, P. (2025). Adversarial robustness of deep learning in medical imaging: A comprehensive survey and benchmark of state-of-the-art architectures. *International Journal of Advanced Computer Science and Applications*, 16(12), 831–841. <https://doi.org/10.14569/IJACSA.2025.0161278>
67. Newaz, A., Haque, N. I., Sikder, A. K., Rahman, M. A., & Uluagac, A. S. (2020). Adversarial attacks to machine learning-based smart healthcare systems. In *2020 IEEE Global Communications Conference (GLOBECOM)*. IEEE.
68. Nikolaidis, C. C., & Efraimidis, P. S. (2025). A study of membership inference attacks on a federated health care application. *Computing*, 107(7), Article 149. <https://doi.org/10.1007/s00607-025-01507-x>
69. Ozbulak, U., Van Messem, A., & De Neve, W. (2019). Impact of adversarial examples on deep learning models for biomedical image segmentation. In *Medical image computing and computer assisted intervention – MICCAI 2019* (Part II). Springer.
70. Pal, S., Rahman, S., Beheshti, M., Habib, A., Jadidi, Z., & Karmakar, C. (2024). The impact of simultaneous adversarial attacks on robustness of medical image analysis. *IEEE Access*, 12, 66478–66494. <https://doi.org/10.1109/access.2024.3396566>
71. Papangelou, K., Sechidis, K., Weatherall, J., & Brown, G. (2019). Toward an understanding of adversarial examples in clinical trials. In *Machine learning and knowledge discovery in databases (ECML PKDD 2018)* (Part I). Springer.
72. Paracha, A., Arshad, J., Ben Farah, M., & Ismail, K. (2024). Exploring data poisoning attacks against adversarially trained skin cancer diagnostics. In *2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC)*. IEEE.
73. Paschali, M., Conjeti, S., Navarro, F., & Navab, N. (2018). Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In *Medical image computing and computer assisted intervention – MICCAI 2018* (Part I). Springer.
74. Paul, W., Cao, Y. Z., Zhang, M. M., & Burlina, P. (2021). Defending medical image diagnostics against privacy attacks using generative methods: Application to retinal diagnostics. In *Clinical image-based procedures, distributed and collaborative learning, artificial intelligence for combating COVID-19 and secure and privacy-preserving machine learning (CLIP 2021, DCL 2021, LL-COVID19 2021, PPML 2021)*. Springer.
75. Pedarla, L. P., Zhang, X. Y., Zhao, L., & Khan, H. (2023). Evaluation of query-based membership inference attack on the medical data. In *Proceedings of the 2023 ACM Southeast Conference (ACMSE 2023)*. Association for Computing Machinery.
76. Pipal, A. K., & Kannan, R. J. (2026). Graph attention layer-based WideResNet for efficient adversarial attack detection in wireless body area networks. *Information Security Journal: A Global Perspective*, 35(1), 203–217. <https://doi.org/10.1080/19393555.2025.2493102>
77. Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156–180. <https://doi.org/10.1109/RBME.2020.3013489>
78. Rahman, A., Hossain, M. S., Alrajeh, N. A., & Alsolami, F. (2021). Adversarial examples: Security threats to COVID-19 deep learning systems in medical IoT devices. *IEEE Internet of Things Journal*, 8(12), 9603–9610. <https://doi.org/10.1109/jiot.2020.3013710>
79. Rahman, S., Pal, S., Fallah, A., Doss, R., & Karmakar, C. (2025). RAD-IoMT: Robust adversarial defense mechanisms for IoMT medical image analysis. *Ad Hoc Networks*, 178, Article 103935. <https://doi.org/10.1016/j.adhoc.2025.103935>

80. Roth, H. R., Hatamizadeh, A., Xu, Z. Y., Zhao, C., Li, W. Q., Myronenko, A., & Xu, D. G. (2022). Split-U-Net: Preventing data leakage in split learning for collaborative multi-modal brain tumor segmentation. In *Distributed, collaborative, and federated learning, and affordable AI and healthcare for resource diverse global health (DeCaF 2022, FAIR 2022)*. Springer.
81. Sahay, R., Lakshmi, L., Sai, P. V. J., Althobaiti, M. M., Salhi, A., Khan, M. I., & Ismoilov, M. (2026). Trust-centric mitigation of time-warping adversarial attacks in wearable sensor platforms. *IEEE Transactions on Consumer Electronics*, 72(2), 4932–4946. <https://doi.org/10.1109/tce.2026.3681806>
82. Sheikh, B. U. H., & Zafar, A. (2023). Untargeted white-box adversarial attack to break into deep learning based COVID-19 monitoring face mask detection system. *Multimedia Tools and Applications*. Advance online publication. <https://doi.org/10.1007/s11042-023-15405-x>
83. Shi, S. H., Haque, M. S., Parida, A., Zhang, C. Y., Linguraru, M. G., Hou, Y. T., Anwar, S. M., & Lou, W. J. (2025). MedLeak: Multimodal medical data leakage in secure federated learning with crafted models. In *2025 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE.
84. Singh, A., Singh, M. P., & Singh, A. K. (2025). Adversarially enhanced learning (AEL): Robust lightweight deep learning approach for radiology image classification against adversarial attacks. *Image and Vision Computing*, 154, Article 105405. <https://doi.org/10.1016/j.imavis.2024.105405>
85. Singkorapoom, P., & Phoomvuthisarn, S. (2023). Pre-trained model robustness against GAN-based poisoning attack in medical imaging analysis. In *Artificial intelligence applications and innovations (AIAI 2023) (Part I)*. Springer.
86. Sipola, T., & Kokkonen, T. (2021). One-pixel attacks against medical imaging: A conceptual framework. In *Trends and applications in information systems and technologies (WorldCIST 2021) (Vol. 1)*. Springer.
87. Sorin, V., Soffer, S., Glicksberg, B. S., Barash, Y., Konen, E., & Klang, E. (2023). Adversarial attacks in radiology: A systematic review. *European Journal of Radiology*, 167, Article 111085. <https://doi.org/10.1016/j.ejrad.2023.111085>
88. Soulb, H., Guelzim, I., Hirchoua, B., & Nait-Charif, H. (2025). Evaluating the robustness in federated learning for medical imaging. In *2025 6th International Conference on Data Analytics for Business and Industry (ICDABI 2025)*. IEEE.
89. Subbanna, N., Wilms, M., Tuladhar, A., & Forkert, N. D. (2021). An analysis of the vulnerability of two common deep learning-based medical image segmentation techniques to model inversion attacks. *Sensors*, 21(11), Article 3874. <https://doi.org/10.3390/s21113874>
90. Sun, L., Tian, J., & Muhammad, G. (2024). FedKC: Personalized federated learning with robustness against model poisoning attacks in the metaverse for consumer health. *IEEE Transactions on Consumer Electronics*, 70(3), 5644–5653. <https://doi.org/10.1109/tce.2024.3386932>
91. Sun, M. Y., Tang, F. Y., Yi, J. F., Wang, F., & Zhou, J. Y. (2018). Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18)*. Association for Computing Machinery.
92. Surekha, M., Sagar, A. K., & Khemchandani, V. (2024). Adversarial attack and defense mechanisms in medical imaging: A comprehensive review. In *Proceedings – International Conference on Computing, Power, and Communication Technologies (IC2PCT 2024)*. IEEE.
93. Surekha, M., Sagar, A. K., & Khemchandani, V. (2025). Robustness against adversarial attacks through a hybrid defense approach in medical imaging. *Electrica*, 25(1), Article 0127. <https://doi.org/10.5152/electrica.2025.25127>
94. Tajak, W., Nurzynska, K., & Piórkowski, A. (2025). Vulnerability to one-pixel attacks of neural network architectures in medical image classification. *Bio-Algorithms and Med-Systems*, 21(1), 58–70. <https://doi.org/10.5604/01.3001.0055.3261>
95. Thota, P., Veerla, J. P., Guttikonda, P. S., Nasr, M. S., Nilizadeh, S., & Lubner, J. M. (2024). Demonstration of an adversarial attack against a multimodal vision language model for pathology imaging. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI 2024)*. IEEE.
96. Trabilisy, M., Prabha, S., Gomez-Cabello, C. A., Haider, S. A., Genovese, A., Borna, S., Wood, N., Gopala, N., Tao, C., & Forte, A. J. (2025). The PIEE cycle: A structured framework for red teaming large language models in clinical decision-making. *Bioengineering*, 12(7), Article 706. <https://doi.org/10.3390/bioengineering12070706>
97. Tsai, M. J., Lee, Y. C., Lien, H. Y., & Liang, C. C. (2025). Adversarial defense for medical images. *Electronics*, 14(22), Article 4384. <https://doi.org/10.3390/electronics14224384>
98. Tsai, M. J., Lin, P. Y., & Lee, M. E. (2023). Adversarial attacks on medical image classification. *Cancers*, 15(17), Article 4228. <https://doi.org/10.3390/cancers15174228>
99. Ul Haque, S. B., Zafar, A., ul Haque, S. M., ul Haq, S. R., & Ahmad, M. (2025). Securing AI in healthcare: A three-layer defense to mitigate adversarial noise impact in radiology imaging. *Biomedical Signal Processing and Control*, 109, Article 107969. <https://doi.org/10.1016/j.bspc.2025.107969>
100. Ullah, F., Mostarda, L., Cacciagrano, D., Alenazi, M. J. F., Chen, C. M., & Kumari, S. (2025). Federated edge intelligence for enhanced security in consumer intermittent healthcare devices using adversarial examples. *IEEE Transactions on Consumer Electronics*, 71(2), 4574–4585. <https://doi.org/10.1109/tce.2024.3511615>

101. Vali, E., Alexandridis, G., & Stafylopatis, A. (2023). Adversarial attacks & detection on a deep learning-based digital pathology model. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE.
102. Wang, J., Kim, B. G., Rani, S., Li, K. Q., & Lv, J. H. (2026). HealthGuard: Privacy-preserving framework for consumer healthcare devices against adversarial inference attacks. *IEEE Transactions on Consumer Electronics*, 72(1), 1465–1476. <https://doi.org/10.1109/tce.2025.3649226>
103. Wang, W. J., Tang, P. F., Xiong, L., & Jiang, X. Q. (2020). RADAR: Recurrent autoencoder based detector for adversarial examples on temporal EHR. In *Machine learning and knowledge discovery in databases: Applied data science track (ECML PKDD 2020) (Part IV)*. Springer.
104. Xi, B. H., Li, S. F., Li, J. C., Liu, H., Liu, H., & Zhu, H. J. (2021). BatFL: Backdoor detection on federated learning in e-health. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*. IEEE.
105. Xu, T., Zhang, J., Wang, M., He, H. Y., & Zou, X. B. (2025). On the limits of uncertainty-aware fine-tuning for robust diabetic retinopathy screening. In *Ophthalmic medical image analysis (OMIA 2025)*. Springer.
106. Yang, Y., Jin, Q., Huang, F., & Lu, Z. (2025). Adversarial prompt and fine-tuning attacks threaten medical large language models. *Nature Communications*, 16(1), Article 9011. <https://doi.org/10.1038/s41467-025-64062-1>
107. Yang, Z. Y., Chen, Y. Y., Sun, M. Y., & Zhang, Y. (2024). Inject backdoor in measured data to jeopardize full-stack medical image analysis system. In *Medical image computing and computer assisted intervention – MICCAI 2024 (Part VII)*. Springer.
108. Yu, Q. K., Zhang, C., Jin, G. J., Huang, T. J., Zhou, W., Li, W. H., Jin, X. B., Huang, B., Zhao, Y. T., Yang, G., Lip, G. Y. H., Zheng, Y. L., Villavicencio, A., & Meng, Y. D. (2026). StealthMark: Harmless and stealthy ownership verification for medical segmentation via uncertainty-guided backdoors. *IEEE Transactions on Image Processing*, 35, 1290–1304. <https://doi.org/10.1109/tip.2026.3655563>
109. Zahra, N. A., & Chin, Z. (2026). Prompt injection is all you need: A systemic framework for evaluating healthcare misinformation in LLMs. In *Artificial intelligence in healthcare (AIIH 2025) (Part I)*. Springer.
110. Zbrzezny, A. M., & Grzybowski, A. E. (2023). Deceptive tricks in artificial intelligence: Adversarial attacks in ophthalmology. *Journal of Clinical Medicine*, 12(9), Article 3266. <https://doi.org/10.3390/jcm12093266>
111. Zhang, C. H., & Bonomi, L. (2022). Mitigating membership inference in deep learning applications with high dimensional genomic data. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI 2022)*. IEEE.
112. Zhang, H., Lou, Q., & Wang, Y. (2024). Towards safe AI clinicians: A comprehensive study on large language model jailbreaking in healthcare. *AMIA Annual Symposium Proceedings, 2024*, 1519–1528.
113. Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics*, 125, Article 103977. <https://doi.org/10.1016/j.jbi.2021.103977>

Appendix A. Database search strategies

The full search strings executed in each database are reproduced below exactly as run. Records were exported on 5 July 2026, yielding 339 records from PubMed, 201 from Scopus, and 783 from Web of Science (1,323 in total before deduplication).

A.1 Scopus

```
TITLE-ABS-KEY (
  ("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model*" OR
  "foundation model*" OR "generative AI" OR "federated learning")
  AND
  ("adversarial attack*" OR "data poisoning" OR "model poisoning" OR "backdoor attack*" OR
  "membership inference" OR "model inversion" OR "prompt injection" OR "data leakage" OR "privacy
  attack*" OR "cybersecurity threat*")
  AND
  (healthcare OR "health care" OR clinical OR hospital* OR "medical imaging" OR "electronic health
  record*" OR "medical device*" OR telemedicine OR "remote patient monitoring")
  AND
  (review OR survey OR "systematic review" OR "scoping review")
)
AND NOT TITLE-ABS-KEY (
  blockchain OR "smart grid" OR "autonomous vehicle*" OR manufacturing OR financial OR "fraud
  detection" OR "malware detection" OR "intrusion detection" OR phishing OR veterinary
)
```

A.2 PubMed

```
(
  ("artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "large
  language model"[tiab] OR "large language models"[tiab] OR "foundation model"[tiab] OR "foundation
  models"[tiab] OR "generative AI"[tiab] OR "federated learning"[tiab])
  AND
  ("adversarial attack"[tiab] OR "adversarial attacks"[tiab] OR "adversarial example"[tiab] OR
  "adversarial examples"[tiab] OR "adversarial machine learning"[tiab] OR "data poisoning"[tiab] OR
  "model poisoning"[tiab] OR "backdoor attack"[tiab] OR "backdoor attacks"[tiab] OR "membership
  inference"[tiab] OR "model inversion"[tiab] OR "prompt injection"[tiab] OR jailbreak[tiab] OR
  jailbreaks[tiab] OR "data leakage"[tiab] OR "privacy attack"[tiab] OR "privacy attacks"[tiab] OR
  "gradient leakage"[tiab] OR "cybersecurity threat"[tiab] OR "cybersecurity threats"[tiab])
  AND
  (healthcare[tiab] OR "health care"[tiab] OR clinical[tiab] OR hospital[tiab] OR hospitals[tiab] OR
  "medical imaging"[tiab] OR radiology[tiab] OR pathology[tiab] OR "electronic health record"[tiab] OR
  "electronic health records"[tiab] OR EHR[tiab] OR "medical device"[tiab] OR "medical devices"[tiab]
  OR "clinical decision support"[tiab] OR telemedicine[tiab] OR telehealth[tiab] OR "remote patient
  monitoring"[tiab])
)
NOT
(blockchain[tiab] OR "smart grid"[tiab] OR "autonomous vehicle"[tiab] OR "industrial control"[tiab]
OR manufacturing[tiab] OR financial[tiab] OR finance[tiab] OR "fraud detection"[tiab] OR "credit
card"[tiab] OR "malware detection"[tiab] OR "network intrusion"[tiab] OR "intrusion detection"[tiab]
OR phishing[tiab] OR veterinary[tiab] OR agriculture[tiab])
AND 2020/01/01:2026/12/31[dp]
AND english[la]
```

A.3 Web of Science

```
TS= (
  ("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model*" OR
  "foundation model*" OR "generative AI" OR "federated learning")
  AND
  ("adversarial attack*" OR "adversarial example*" OR "adversarial machine learning" OR "data
  poisoning" OR "model poisoning" OR "backdoor attack*" OR "membership inference" OR "model inversion"
  OR "prompt injection" OR jailbreak* OR "data leakage" OR "privacy attack*" OR "gradient leakage" OR
  "cybersecurity threat*")
  AND
  (healthcare OR "health care" OR clinical OR hospital* OR "medical imaging" OR radiology OR
  pathology OR "electronic health record*" OR EHR OR "medical device*" OR "clinical decision support"
  OR telemedicine OR telehealth OR "remote patient monitoring")
)
NOT TS= (
  blockchain OR "smart grid" OR "autonomous vehicle*" OR "industrial control" OR manufacturing OR
  financial OR finance OR "fraud detection" OR "credit card" OR "malware detection" OR "network
  intrusion" OR "intrusion detection" OR phishing OR veterinary OR agriculture
)
```