STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–28.
Published online in International Academic Press (www.IAPress.org)

Zero-Shot Anomaly Detection in Healthcare: A Systematic Review

Faisal A. Althobaiti*, Tarig M. Ahmed, Rayan A. Alsemmeari

Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Abstract Healthcare organizations need to detect anomalies that nobody has labeled in advance, such as new billing-fraud schemes, irregular activity in electronic health records (EHRs), and unexpected physiological or device behavior in Internet of Things (IoT) monitoring. Zero-shot anomaly detection (ZSAD) is designed for this setting, but it is unclear how much of the healthcare literature actually tests zero-shot generalization. Following PRISMA 2020, we searched PubMed/MEDLINE, IEEE Xplore, Scopus, Google Scholar and the ACM Digital Library (last search January 2026) for studies published from 2022 to 2025. Of 832 records, 295 full-text reports were assessed and 49 met the eligibility and risk-of-bias threshold: 40 primary studies and 9 reviews used only for context. Every primary study was coded on an operational zero-shot scale (Z0, supervised, to Z3, strict zero-shot), a validation scale (V0 to V4) and three machine-learning-adapted risk-of-bias domains, and the results were synthesized without meta-analysis following the SWiM guideline. One study evaluated strict zero-shot detection (leave-one-dataset-out medical imaging; mean AUROC 0.80, which rose to 0.89 with four labeled shots), and one more combined foundation-model priors with normal-only training. Nineteen studies were anomaly-label-free, 14 were supervised and 5 did not report the learning regime. Two studies used external validation and none was prospective. Reported AUROC ranged from 0.70 to 0.998 (median 0.92; 14 studies), and the highest values came from supervised classification of known attacks on IoT testbed data. Fraud detection contributed two studies. Seven studies applied explainability methods, and none measured their effect on false alarms. One study was at low overall risk of bias, and the certainty of evidence was low or very low for every main finding. ZSAD in healthcare is therefore at an early, imaging-led stage. The proposed zero-shot scale, evidence map, governance indicators and composite generalization index provide testable criteria for designing and reporting future studies.

Keywords Zero-Shot Anomaly Detection, Few-Shot Learning, Foundation Models, Machine Learning, Healthcare, Healthcare Fraud, Electronic Health Records, Internet of Medical Things, Systematic Review, Risk of Bias, Ethical Governance, Explainability

DOI: 10.19139/soic-2310-5070-5033

1. Introduction

The anomalies that matter most in healthcare are often the ones nobody has labeled yet: a new upcoding pattern in insurance claims, an access sequence in an EHR audit log that no rule anticipates, or a sensor trace that reflects either device failure or clinical deterioration. Supervised detectors need labeled examples of each anomaly type. Such examples are scarce, expensive to annotate and quickly outdated as fraud schemes, software and patient populations change [8, 48]. Severe class imbalance, distribution shift between development and deployment sites, and privacy limits on data sharing make the problem harder [11, 25, 28]. Zero-shot anomaly detection (ZSAD) is aimed at exactly this setting. A ZSAD model detects anomalies in a target task for which it has seen no labeled anomalies and, in the strict sense, no target training data at all. It does so by transferring knowledge from pretrained representations, auxiliary datasets or natural-language descriptions of normality and abnormality [19, 30, 62, 63, 64].

*Correspondence to: Faisal A. Althobaiti (Email: falthobaiti0018@stu.kau.edu.sa). Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

In the settings covered here, anomaly detectors usually serve as clinical or administrative decision support. They raise alerts for a human reviewer and leave the decision to that reviewer [44]. Several included studies describe this role for billing patterns, EHR activity and physiological signals from IoT devices [26, 35, 45]. Whether such alerts improve patient or financial outcomes, however, has not yet been tested prospectively in the evidence reviewed here (Section 3).

This review covers three application domains: fraud detection in insurance and billing, anomaly identification in EHRs and health-information systems, and real-time patient monitoring through IoT medical devices. We examine them together for three reasons. They share the problem structure that motivates ZSAD, because anomalies are rare, heterogeneous and evolve over time, sometimes adversarially (fraud and cyberattacks). They also form a chain of data within one episode of care: a device reading becomes an EHR entry, and the EHR entry becomes a claim, so an anomaly in one stream can explain or corroborate a signal in another. A sensor fault can write implausible vital signs into the record, and a phantom claim can lack any supporting EHR activity. Finally, the three domains fall under overlapping privacy, security and safety rules, so the same deployment questions recur across them. Medical imaging and diagnostic signals were not a planned domain, but the search returned them and they currently contain the only strict zero-shot evidence, so we report them as a fourth stratum. Fraud detection is kept as a core domain even though it contributed only two primary studies, because the scarcity of evidence there is itself a result that practitioners need to know.

1.1. Motivation and gap in prior reviews

Several reviews have examined anomaly detection in healthcare, but none of them answers the question posed here. Table 1 compares them on scope, period and sources, quality appraisal, coverage of zero-shot methods and coverage of governance. Earlier surveys organized deep anomaly-detection methods for medical data without a systematic search or quality appraisal [75, 23]. Domain reviews covered wearables [45], medical time series [38, 36], ensembles of autoencoders [29] and healthcare fraud [85], each within a single domain or method family. The BMAD benchmark standardized the evaluation of medical-image anomaly detection, but it is a benchmark rather than an evidence synthesis [28]. None of these works distinguishes zero-shot from few-shot, one-class and supervised detection with an operational rule, none covers fraud, EHR and IoT monitoring under one protocol, and none appraises risk of bias with machine-learning-specific criteria.

The present review makes six specific contributions relative to these works:

(1) An operational and testable definition of ZSAD with a four-level zero-shot degree scale (Z0 to Z3) and decision rules (Section 1.3). Applying the scale to every included study shows how little of the label-efficient literature in healthcare actually tests zero-shot generalization.
(2) A study-level evidence map of the 40 primary studies by domain, zero-shot degree, mechanism family and validation level, together with an evidence-anchored risk-of-bias re-appraisal and GRADE-style certainty ratings.
(3) A metric-specific synthesis without meta-analysis. It replaces pooled averages across incompatible metrics with study-level values, stratified medians and ranges, and six sensitivity analyses.
(4) A classification framework that states what it inherits from, revises in, and adds to existing anomaly-detection taxonomies and clinical-evaluation frameworks, verified by coding all 49 studies.
(5) A reformulated composite generalization index with explicit normalization, a weight-sensitivity analysis and a worked example on published results. We present it as a theoretical proposal that still requires validation.
(6) Governance recommendations translated into audit indicators, responsibility assignments and illustrative deployment thresholds, each linked to how often the included studies already provide the required evidence.

1.2. Background

Classical anomaly detection includes statistical rules such as z-score thresholds and Gaussian mixture models, clustering and density methods such as k-means and DBSCAN, and supervised classifiers such as support vector machines and random forests trained on labeled normal and anomalous samples [71, 22, 41]. Supervised classifiers can be accurate when labeled anomalies are plentiful, but they struggle with class imbalance and with anomaly

Table 1. Structured comparison of the present review with prior reviews and benchmarks on anomaly detection in healthcare

Work	Type	Scope	Period and sources (as reported)	Quality appraisal	Zero-shot coverage	Governance and ethics	Gap relative to this review
Fernando et al. [75]	Survey	Deep anomaly detection across medical data types	Not a systematic search	None	Not covered (pre-VLM)	Interpretability only	No operational zero-shot definition; no appraisal
Tschuchnig and Gadermayr [23]	Mini review	Medical imaging	Google Scholar, four search terms	None	Not covered	Not covered	Imaging only; no fraud, EHR or IoT
Sunny et al. [45]	Perspective review	Wearable data	Google Scholar (2021)	None	Not covered	Briefly	Single domain; no appraisal
Kumaraswamy et al. [85]	Narrative review	Healthcare fraud data mining	About 22 studies, 2006–2021; databases not named	None	Not covered	Investigative workflow only	Fraud only; pre-foundation-model era
Yang et al. [38]	Review	Deep learning for medical time series	Search strategy not reported	None	Not covered	Briefly	Single data type; no appraisal
Cekić [36]	Selective review	GANs for medical time series	Selective	None	Not covered	Briefly	Single method family
Nawaz et al. [29]	Narrative review	Ensembles of autoencoders, biomedical data	IEEE, Springer, ScienceDirect, Google Scholar (2023); 16 papers	None	Not covered	Not covered	Single method family; not systematic by the authors' statement
Bao et al. (BMAD) [28]	Benchmark	Six medical-imaging datasets, 15 methods	Public datasets	Not applicable	One-class methods; zero-shot not evaluated	Not covered	Imaging only; not an evidence synthesis
This review	Systematic review (PRISMA 2020, SWiM)	Fraud, EHR/health-information systems, IoT monitoring; imaging as a fourth stratum	Five sources, 2022–2025, searched January 2026; 49 studies	Three-domain risk of bias at eligibility and full-text re-appraisal; GRADE-style certainty	Operational Z0–Z3 scale applied to every study	Regulatory requirements mapped to audit indicators and to the evidence	First cross-domain, appraisal-based synthesis of ZSAD evidence

types absent from training, both of which are common in healthcare [43, 91]. Hybrid designs that combine a temporal model with a density-based detector, such as LSTM with the local outlier factor, have been proposed for IoMT and clinical data [88]. Deep one-class and reconstruction-based methods, reviewed in [69, 70], learn a model of normality and flag deviations from it. In medical imaging this line of work predates our search window and includes f-AnoGAN [76], MADGAN [77] and deep perceptual autoencoders [78], surveyed by Fernando et al. [75]. These methods need no anomaly labels, but they are trained on normal data from the target task, so they are one-class rather than zero-shot detectors.

Zero-shot learning (ZSL) enables a model to recognize classes for which it saw no training examples by relying on auxiliary information such as attributes, text descriptions or shared embeddings [74, 73]. ZSAD transfers this idea to anomaly detection. In the strict setting a pretrained model scores samples from a new task, dataset or modality without any training samples from that target, usually by comparing image or sequence features with text prompts that describe normal and abnormal states. Few-shot anomaly detection relaxes this by allowing a small number k of target samples, which may be normal only or include abnormal examples. Two related terms are often confused with ZSAD. Out-of-distribution (OOD) detection flags inputs that differ from the training distribution of a classifier [92, 72], and open-set recognition classifies known classes while rejecting unknown ones [93]. Both address unseen inputs, but neither requires transfer to a target task without target data, which is the defining property of ZSAD.

Foundation models made strict ZSAD practical. Contrastive vision-language pretraining (CLIP) aligned images with natural-language descriptions at scale [62]. WinCLIP showed that CLIP features compared with normal and abnormal text prompts could detect industrial defects without target training data [63], and AnomalyCLIP learned object-agnostic prompts that generalize across datasets [64]. Generalist few-shot detectors such as InCTRL use a handful of normal samples as in-context prompts [94]. In healthcare, MVFA adapted CLIP with multi-level residual adapters and was evaluated on unseen imaging modalities [30], and Ano-NAViLa combined a pathology vision-language model with term pools for normal and abnormal histology [19]. Beyond CLIP, self-supervised visual encoders (DINOv2) [65], sigmoid-loss language-image pretraining (SigLIP) [66], and biomedical or pathology-specific vision-language models such as BiomedCLIP [67] and CONCH [68] provide alternative representations for zero-shot scoring. For structured EHR data, transformer models trained on clinical event sequences, such as

EHR-BERT, capture contextual irregularities [20]. These representations enable semantic transfer, but their benefit in fraud, EHR and IoT data, which are not image-text pairs, has to be shown empirically. This review examines that evidence.

1.3. Operational definitions

Reviewer-independent classification requires definitions that can be checked against what a study reports. We therefore define the following.

Definition 1 (ZSAD). A detector performs zero-shot anomaly detection on a target task $\mathcal{T}$ if it assigns anomaly scores to samples from $\mathcal{T}$ without using any labeled or unlabeled training samples from $\mathcal{T}$, and if its knowledge of normality and abnormality comes from pretraining, auxiliary tasks or textual descriptions. A target task is defined by its dataset, site or modality.

Definition 2 (zero-shot degree). Each evaluation reported by a study is assigned one level of the scale in Table 2. A study is coded at the highest level at which it reports results, and all levels it evaluates are recorded. Use of a pretrained foundation or vision-language model (FM) is recorded as a separate flag, because FM priors can be combined with any level.

Table 2. Zero-shot degree scale used to classify each included study (decision rules are applied in order; the codebook with examples is in Supplementary S7)

Level	Name	Decision rule (what the study must report)	Typical mechanisms
Z3	Strict zero-shot	No samples of any kind from the target dataset, site or modality are used for training, tuning or prompt construction; performance is reported on that unseen target.	Prompt or text alignment with pretrained vision-language models; leave-one-dataset-out transfer
Z2	Few-shot	A stated number k of target samples (normal only, or normal and abnormal) is used for adaptation; k and the sample types are reported.	Memory banks, adapters or in-context prompts built from k shots
Z1	Anomaly-label-free	Training uses normal-only or unlabeled target data; no anomaly labels are used for training or model selection. Flag "+FM" if pretrained foundation-model priors are used.	One-class, reconstruction, isolation, density, distance and self-supervised sequence models
Z0	Supervised	Anomaly labels, attack labels, pseudo-labels produced by another detector, or active-learning queries are used for training.	Classifiers, ensembles, pseudo-label pipelines, active learning
NS	Not specified	The report does not state which data and labels were used for training.	Not determinable

Definition 3 (validation level). V0: no reproducible evaluation reported; V1: evaluation only on simulated, synthetic or non-health data; V2: retrospective evaluation on real health data with an internal split; V3: evaluation on a dataset, site or modality not used for training (external, cross-dataset or leave-one-domain-out); V4: prospective or deployed evaluation. These levels follow the staged logic of clinical-AI evaluation frameworks such as DECIDE-AI [60].

1.4. Research Questions

This paper aims to answer the following research questions:

RQ1: What are the current zero-shot anomaly detection (ZSAD) mechanisms and techniques applied in healthcare, and to what degree do they meet the operational definition of zero-shot detection?
RQ2: How can these ZSAD mechanisms be systematically classified and compared in terms of methodological approaches, validation level and performance across healthcare modalities?
RQ3: What are the primary open challenges, limitations, ethical/governance considerations, and future research directions for advancing ZSAD implementation and scalability in real-world healthcare systems?

1.5. Research Objectives

RO1: To categorize current ZSAD mechanisms in healthcare and grade their zero-shot degree.
RO2: To classify and compare ZSAD mechanisms through a structured framework and taxonomy, stratified by domain, zero-shot degree and validation level.
RO3: To identify open challenges, ethical considerations, and future directions for scalable adoption, and to translate governance requirements into auditable indicators.

2. Material and method

2.1. Review type, protocol and registration

This study is a systematic review with a narrative and descriptive quantitative synthesis. It follows the PRISMA 2020 statement [51] and its search extension PRISMA-S [52], and it reports the synthesis according to the Synthesis Without Meta-analysis (SWiM) guideline [53]. A meta-analysis was not planned, because the included studies use different tasks, datasets and metrics and rarely report variances. The review was not registered prospectively and no protocol was published before the search, which we acknowledge as a limitation (Section 5). To make the methods auditable, the protocol as applied is documented in Supplementary S6. S6 lists every deviation from the criteria reported in the original submission, with its reason: the retention of nine secondary studies for context only, the retention of three preprints, the corrected wording of the risk-of-bias threshold, the full-text re-appraisal and re-extraction carried out during revision, and the post hoc stratification by zero-shot degree and validation level. The completed PRISMA 2020 checklist is given in Supplementary S2.

2.2. Eligibility criteria

Table 3 lists the eligibility criteria as they were applied. Two differences from the criteria stated in the original submission are reported transparently. First, nine included records are secondary studies (reviews, surveys and a perspective paper) [2, 8, 15, 23, 26, 29, 36, 38, 45]. They passed screening because their topic was relevant, and they are retained to keep the included set reproducible, but they are used only for context and are excluded from the risk-of-bias re-appraisal and from every quantitative summary. Second, three included records are preprints [5, 8, 24], although preprints were listed as an exclusion criterion. They were identified through Google Scholar and kept because they report complete methods. A sensitivity analysis that removes them is reported in Section 3.5.1.

The publication window of 2022 to 2025 was chosen because large-scale vision-language pretraining became available in 2021 [62] and the first prompt-based ZSAD methods appeared from 2023 [63, 64]. Studies from earlier years apply one-class or supervised detection without semantic transfer. These foundational studies inform the background of this review [75, 76, 77, 78], but they cannot contribute evidence on zero-shot transfer, which is the subject of RQ1 and RQ2. The scope of the topic criterion was intentionally broader than strict ZSAD. It admitted label-efficient anomaly detection that claims to address unseen anomalies, so that strict zero-shot studies could be compared with the methods practitioners actually use. The zero-shot degree scale (Table 2) was applied after inclusion to make this breadth explicit.

2.3. Information sources and search strategy

The online databases that were searched are PubMed/MEDLINE for clinical topics, IEEE Xplore for technical, IoT and machine-learning venues, Scopus for broad coverage, Google Scholar for comprehensiveness and grey literature, and the ACM Digital Library for computing applications. No trial or study registers were searched, because ZSAD studies are not registered. IEEE Xplore indexes the CVPR, ICCV and WACV proceedings and their workshops, and Google Scholar indexes the NeurIPS, ICML and ICLR proceedings, OpenReview, arXiv, ResearchGate and Research Square; the included CVPR, CVPR-workshop and ICCV papers [30, 28, 19] and the three preprints were retrieved through these routes. Technical reports and industry white papers were not searched separately.

Table 3. Eligibility criteria as applied (deviations from the original submission are marked † and explained in Supplementary S6)

Criterion Type	Inclusion Criteria	Exclusion Criteria
Publication type	Peer-reviewed journal articles and full conference papers; preprints and book chapters reporting a complete method and evaluation† (flagged and tested in sensitivity analysis)	Editorials, letters, abstracts, theses and books
Language	English	Non-English
Publication date	2022–2025 (version of record)	Before 2022
Topic relevance	Anomaly detection in healthcare (fraud and billing, EHR and health-information systems, IoT and physiological monitoring) that addresses unseen or unlabeled anomalies through zero-shot, few-shot, label-free or hybrid mechanisms	Non-healthcare applications; general machine learning without an anomaly focus
Study design	Primary studies describing a method and its evaluation; secondary studies retained for context only†	Protocols without results
Risk of bias (full text)	At least two of three risk-of-bias domains judged low (Section 2.5)	Two or more domains judged unclear or high

The primary Boolean query was (*"Zero-Shot Learning" OR "Zero-Shot Detection" OR "Zero-Shot Anomaly Detection" OR "Generalized Zero-Shot"*) AND (*"Healthcare" OR "Medicine" OR "Clinical" OR "Medical Imaging" OR "EHR" OR "Biosignals"*) AND (*"anomaly detection" OR "outlier detection"*). It was expanded with enabling terms (*"few-shot", "prompt-based", "CLIP", "vision-language"*) and domain terms (*"IoT", "fraud"*). The search strings, filters, record counts and the date of each search are reported per database in Supplementary S2, and Table 4 summarizes them. All searches were run in January 2026 and limited to 2022–2025 and English. Google Scholar was screened up to the first 200 results of each query. Synonyms from neighboring fields, such as "out-of-distribution detection", "open-set recognition" and "novelty detection", were not included. These fields study classifiers that reject unknown inputs rather than transfer to a target without target data (Section 1.2), and adding them would have retrieved a large non-healthcare literature. We discuss the resulting risk of missed studies in Section 5.

2.4. Selection process

Duplicates were removed with EndNote and then verified manually. Titles, abstracts and keywords were screened for eligibility, and the full texts of potentially eligible reports were then assessed. Screening of titles, abstracts and full texts was carried out by one reviewer (F.A.A.). Uncertain records were discussed with the co-authors and decided by consensus. Dual independent screening was not performed, so inter-rater agreement statistics for screening cannot be reported. Reports whose full text could not be obtained were recorded as not retrieved and were not assessed. The process is shown in Figure 1 and in Supplementary S3.

2.5. Risk-of-bias threshold at the eligibility stage

Each of the 295 full-text reports was appraised in three domains, and the appraisal was used as the final eligibility step. The domains adapt the analysis, participants and applicability domains of PROBAST [54] and of its update for artificial-intelligence models, PROBAST+AI [55], together with the reporting items of TRIPOD+AI [56], to anomaly-detection studies. PROBAST-derived criteria are appropriate here for a specific reason. An anomaly detector that outputs a score and a threshold is a prediction model for a binary state (anomalous or not), and the main threats to its validity are those PROBAST+AI targets: unrepresentative data, information leakage between training and test data, optimistic evaluation and poor applicability to the intended setting. Two features of anomaly

Table 4. Search strategy summary (full strings per database in Supplementary S2)

Database	Timeframe	Key Search Terms (Representative Boolean Query)	Date Accessed	Initial Records
PubMed/MEDLINE	2022–2025	(“Zero-Shot Learning” OR “Zero-Shot Detection” OR “Zero-Shot Anomaly Detection” OR “Generalized Zero-Shot”) AND (“Healthcare” OR “Medicine” OR “Clinical” OR “Medical Imaging” OR “EHR” OR “Biosignals”) AND (“anomaly detection” OR “outlier detection”) [**adapted with:** “few-shot” OR “prompt-based” OR “CLIP” OR “vision-language” OR “IoT” OR “fraud”]	January, 2026	148
IEEE Xplore	2022–2025	Same as above, filtered to journals/conference	January, 2026	95
Scopus	2022–2025	Same as above, limited to English peer-reviewed	January, 2026	210
Google Scholar	2022–2025	Same as above, first 200 results per query	January, 2026	312
ACM Digital Library	2022–2025	Same as above	January, 2026	67
Total (before duplicates)	—	—	—	**832**

detection that PROBAST does not cover were added: whether the evaluation contains anomaly types or data sources unseen during training, and whether anomalies are real or synthetic. The domains are:

(i) **D1 Reproducibility and reporting completeness** (PROBAST+AI analysis domain; TRIPOD+AI items on data and code availability).
(ii) **D2 Performance overestimation** (PROBAST+AI analysis domain): held-out testing, testing on unseen anomaly types or sources, leakage and circular labels.
(iii) **D3 Clinical applicability and validation** (PROBAST+AI participants and applicability domains): real health data, clinically or administratively grounded anomaly definitions, and the validation setting.

Each domain was judged low, unclear or high risk with the GENERIC template of the robvis tool [50]. The ROBINS-I template was not used because it addresses non-randomized studies of interventions. In the original submission the threshold was described with a numerical score. The rule that was actually applied, and that reproduces every decision in Supplementary S5, is the following: *a report was retained when at least two of its three domains were judged low risk, and excluded otherwise*. We have corrected the description accordingly.

Using risk of bias as an exclusion criterion is accepted in systematic reviews in software engineering and computing [57], but it can introduce selection bias, and Cochrane-style reviews generally avoid it. We therefore report a sensitivity analysis of the threshold (Section 3.3), compare the domain-level profiles of retained and excluded reports, and, in the synthesis, treat risk of bias as a stratification variable rather than a further filter.

2.6. Data collection and data items

Data were first extracted with a piloted form in Microsoft Excel (reference, year, objective, methodology, key findings, performance metrics, limitations and recommendations), and a random 20% sample of the included studies (n = 10) was extracted a second time, independently, for verification. Discrepancies were resolved by discussion, but agreement was not quantified. During the revision the free-text form was replaced by an operational codebook with explicit decision rules (Supplementary S7), and all 49 reports were re-extracted against their full texts. Cases in which the rules were ambiguous, for example pseudo-labels produced by an unsupervised detector or attack labels from a network testbed, were resolved by adding a rule to the codebook, and these cases are listed in S7. The re-extraction corrected several entries of the original extraction table, and the corrected values are used throughout this paper. For each study the following items were recorded: year, venue type and peer-review

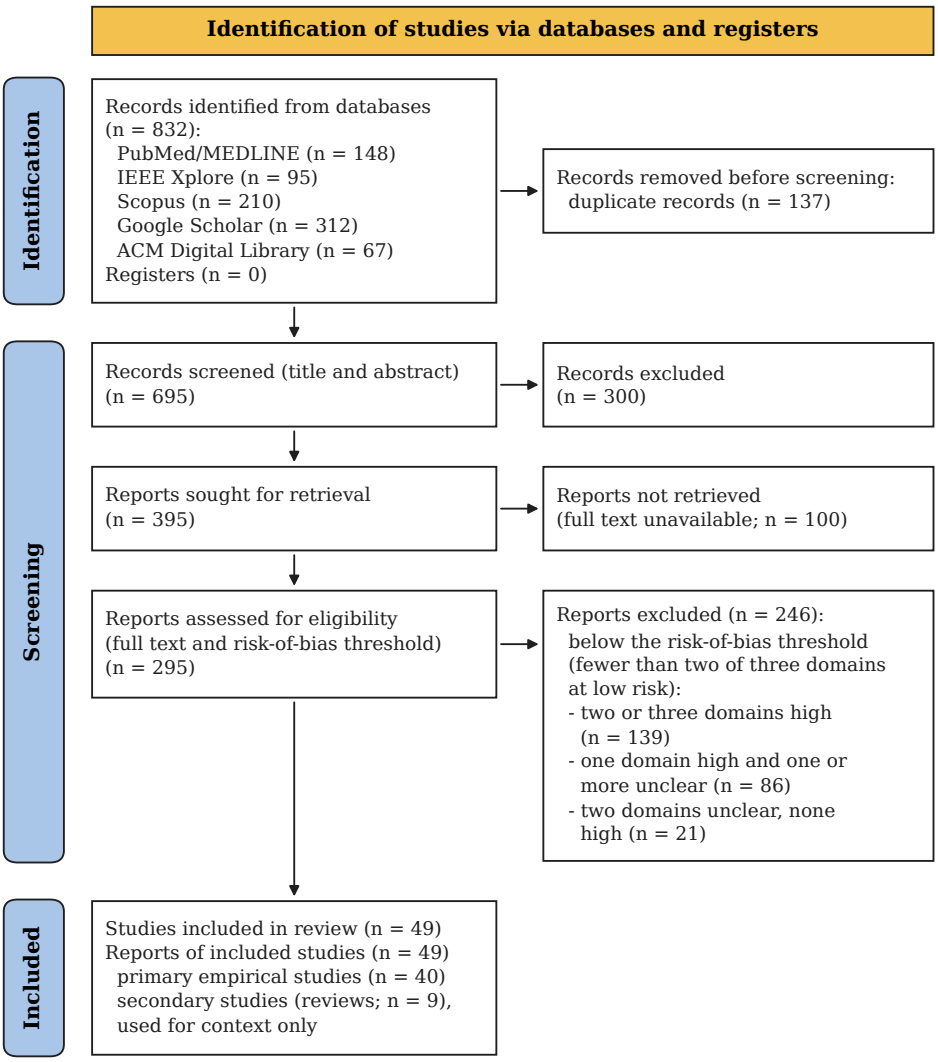


Figure 1. PRISMA 2020 flow diagram. Exclusion reasons at full text are derived from the domain-level judgments in Supplementary S5

status; study design; domain and sub-domain; data modality; mechanism family; zero-shot degree (Table 2) and use of a foundation model; datasets and their public availability; code availability; validation level (V0 to V4); whether anomalies were real (clinically or administratively confirmed) or synthetic, injected or attack-label based; the headline value of each reported metric (AUROC, F1, accuracy and others); the use of explainability methods; and whether computational cost was reported. The study-level dataset, with a note on the evaluation setting behind each value, is provided in Supplementary S1 and as a spreadsheet, so that every summary in this paper can be recomputed.

2.7. Full-text re-appraisal of the included studies

The eligibility-stage appraisal was coarse, and its overall label meant only that a report had passed the threshold. To give the reader domain-level judgments that can be traced to the reports, the 40 primary studies were re-appraised

during revision with the signalling questions in Table 5. The overall judgment follows the usual convention in risk-of-bias tools: it equals the most severe domain judgment. The answers to each signalling question are given in Supplementary S5. Secondary studies were not appraised with this tool, because it is designed for primary evaluations.

Table 5. Signalling questions and judgment rules for the full-text re-appraisal, with their source in PROBAST+AI or TRIPOD+AI

Domain	Signalling questions	Judgment rule	Source
D1 Reproducibility and reporting	1.1 Is the code publicly available? 1.2 Are the data public or available under credentialed access? 1.3 Are data provenance and the evaluation protocol reported?	Low: 1.1 and 1.2 yes. Unclear: one of them yes. High: neither, or no reproducible evaluation (1.3 no).	TRIPOD+AI items on data, code and protocol
D2 Performance overestimation	2.1 Is performance estimated on data held out from training and model selection? 2.2 Is the model tested on a dataset, site, modality or anomaly type not used in training? 2.3 Are evaluation labels independent of the detector, and are anomalies real rather than simulated?	Low: 2.2 yes with real anomalies. High: evaluation only on simulated or non-health data, no reproducible evaluation, or circular pseudo-labels. Unclear: internal held-out testing only.	PROBAST+AI analysis domain
D3 Clinical applicability and validation	3.1 Are the data from a real healthcare setting? 3.2 Are anomalies defined by clinical or administrative ground truth? 3.3 Is there external, multi-site or real-world validation?	Low: 3.2 yes and 3.3 yes. High: 3.1 no. Unclear: otherwise.	PROBAST+AI participants and applicability domains
Overall	Most severe of D1–D3	Low only if all three domains are low	Convention of risk-of-bias tools

2.8. Synthesis method

Grouping. Following SWiM [53], studies were grouped by domain (fraud, EHR and health-information systems, IoT monitoring, and imaging and diagnostic signals), by zero-shot degree and by validation level. The groups were defined before the quantitative summaries were computed, but after data extraction, so they are post hoc with respect to the original protocol.

Metric harmonization. Different metrics were never pooled. AUROC, F1 and accuracy were summarized separately. For each study we took the headline value of the proposed or best-performing model on its primary test set, as reported, on the proportion scale. When a study reported several datasets, the first dataset in its results table was used and the others are listed in Supplementary S1. A study reporting only a range (for example, F1 from 0.80 to 0.99 across thresholds) was tabulated but not included in the medians. Missing metrics were not imputed and were not recomputed from figures.

Summary statistics. Because a meta-analysis was not appropriate, we report for each stratum the number of studies, the unweighted median and the range of the reported values. These summaries describe what studies reported and are not estimates of a common effect. No weighting was applied, because sample sizes and variances were rarely comparable or reported.

Heterogeneity and sensitivity. Heterogeneity was explored by stratification (domain, zero-shot degree, validation level and supervision). Six sensitivity analyses were defined during revision, before the stratified summaries were computed. They excluded, in turn, studies without real-data evaluation (V0 and V1), preprints, studies at high risk of performance overestimation (D2), supervised studies, label-free and zero-shot studies, and IoT security studies that rely on testbed attack labels.

Thematic analysis. Qualitative findings were synthesized with reflexive thematic analysis [58]. Initial codes were derived from the extraction fields and refined inductively. The final codebook, with code definitions, inclusion rules, example studies and the number of studies per code, is reported in Supplementary S7. The audit trail is provided by the extraction spreadsheet, in which each code is linked to the study and to the passage that supports it. Coder agreement was not computed, because coding was done by one reviewer with consensus discussion.

Table 6. Distribution of the 49 included studies by year of the version of record, with examples of venues

Year	Studies	%	Primary / secondary	Venues (examples)
2022	5	10.20	3 / 2	Sensors; JMIR Aging; PLOS ONE; Springer book chapter
2023	8	16.33	6 / 2	IEEE Access; Scientific Reports; Healthcare; Health Information Science and Systems; DATA conference
2024	14	28.57	12 / 2	IEEE Access; Journal of Biomedical Informatics; BMC Medical Informatics and Decision Making; CVPR and CVPR workshops; arXiv
2025	22	44.90	19 / 3	Computers in Biology and Medicine; International Journal of Medical Informatics; IEEE Journal of Biomedical and Health Informatics; Sensors; ICCV; Research Square
Total	**49**	**100**	**40 / 9**	

Reporting bias and certainty. A funnel plot or Egger's test was not possible, because there is no common effect measure. Reporting bias was instead assessed qualitatively. We recorded selective reporting of the best model only, the absence of threshold-free metrics, and the 100 reports that could not be retrieved. The certainty of each main finding was rated with an approach adapted from GRADE [59]. Every finding started at high certainty and was downgraded for risk of bias, inconsistency, indirectness (for example, evidence from one domain applied to another), imprecision (few studies) and suspected reporting bias.

Self-appraisal of the review. The conduct of the review was audited with the general CASP checklist for systematic reviews (Supplementary S4). The checklist version for meta-analyses of randomized controlled trials used in the original submission was not appropriate for this design and has been replaced. CASP was used only to audit the review itself and played no part in selecting studies.

3. Results

3.1. Study selection

The searches returned 832 records. After 137 duplicates were removed, 695 records were screened by title and abstract and 300 were excluded. Of the 395 reports sought for retrieval, 100 could not be obtained in full text and were not assessed. The remaining 295 reports were assessed for eligibility, and 246 of them fell below the risk-of-bias threshold: 139 had two or three domains at high risk, 86 had one domain at high risk and at least one unclear, and 21 had two unclear domains and none at high risk (Figure 1). The original flow diagram showed these 246 exclusions under three unlabeled reasons and listed 832 records under "registers". Both errors are corrected in Figure 1. Forty-nine studies were included: 40 primary studies and 9 secondary studies used for context.

3.2. Characteristics of the included studies

Table 6 gives the distribution by year. Counting the year of the version of record, 22 studies were published in 2025, 14 in 2024, 8 in 2023 and 5 in 2022. The original submission reported 23 and 24 studies for 2025 in two different places. That inconsistency arose from counting some articles by their online-first year, and the counts in Table 6 are now used consistently throughout the manuscript and the supplement. Thirty-nine records are journal articles, five are conference papers, three are preprints and two are book chapters.

Among the 40 primary studies, 2 address fraud and billing [14, 16], 14 address EHR and health-information systems, 15 address IoT and wearable monitoring (7 on physiological monitoring and 8 on device or network security) and 9 address medical imaging and diagnostic signals. By data modality, 14 studies analyze network traffic, system logs or event sequences, 9 analyze physiological signals or sensor streams, 9 analyze structured tabular or administrative records, 7 analyze images and 1 combines images with EHR data [12]. Public data were

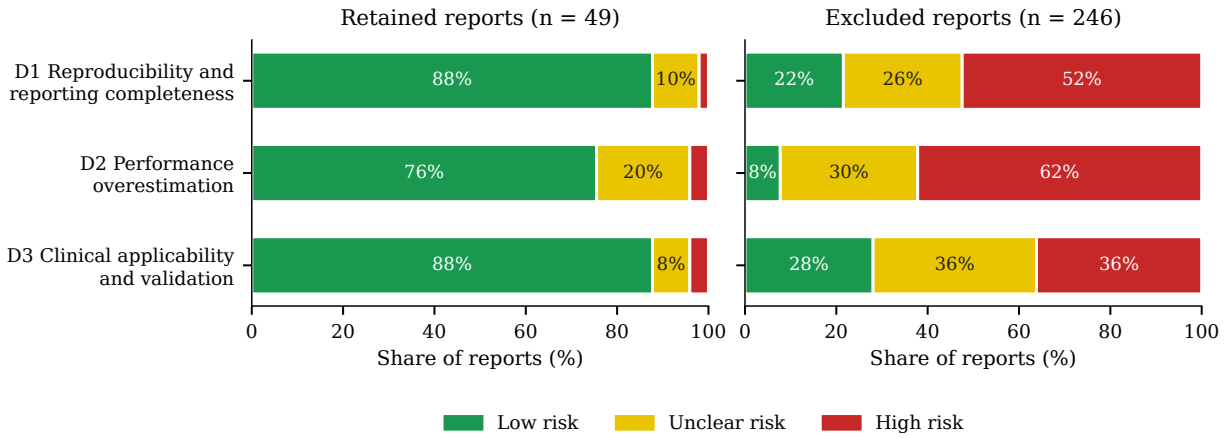


Figure 2. Eligibility-stage risk-of-bias judgments for the 49 retained and 246 excluded full-text reports, by domain

used in 22 studies, public code was released in 10, and both were available in 8 [1, 3, 9, 24, 28, 30, 39, 40]. Table 8 lists every primary study with its coding.

3.3. Risk of bias

Eligibility stage. Figure 2 compares the domain-level judgments of the 49 retained and the 246 excluded reports. Retained reports were judged at low risk in D1 in 43 cases (88%), in D2 in 37 (76%) and in D3 in 43 (88%). Among excluded reports, D2 (performance overestimation) was the domain most often at high risk (153 of 246, 62%), followed by D1 (129, 52%) and D3 (89, 36%). Table 7 shows how the size of the evidence base depends on the threshold. Requiring all three domains at low risk would have retained 25 reports. The applied rule retained 49. The score-based wording of the original submission would have retained 70, the 21 additional reports being those with two unclear domains. Excluding a report only when two or more domains were at high risk would have retained 156.

Table 7. Sensitivity of the eligible evidence base to the risk-of-bias threshold (295 full-text reports; Supplementary S5)

Decision rule	**Retained**	**Excluded**
All three domains low (strict worst-domain rule)	25	270
At least two domains low (rule applied in this review)	**49**	**246**
No domain high (unclear allowed)	65	230
Total score ≤ 5 and at most one high domain (wording of the original submission)	70	225
Excluded only if two or more domains high	156	139
No risk-of-bias exclusion	295	0

The threshold therefore removed mainly reports whose evaluation was likely to overstate performance. If anything, this should make the performance values summarized below less optimistic than those of the full literature. The threshold may also have removed relevant studies from smaller venues and lower-resource settings, where reporting is often less complete. Performance data were not extracted from the 246 excluded reports, so a direct comparison of results with and without them is not possible. We treat this as a limitation of the review (Section 5).

Full-text re-appraisal. Figure 3 shows the re-appraisal of the 40 primary studies with the rules in Table 5. In D1, 8 studies were at low risk, 16 unclear and 16 high, mainly because 30 studies released no code. In D2, only the two studies with external or cross-dataset evaluation were at low risk [19, 30]. Twenty-nine were unclear because they reported internal testing only, and nine were at high risk because they used simulated, synthetic or non-health data,



Figure 3. Full-text risk-of-bias re-appraisal of the 40 primary studies (traffic-light plot and summary). Symbols: + low, − unclear, × high. Numbers in brackets are reference numbers

reported no reproducible evaluation, or derived their labels from another detector [32]. In D3, four studies were at low risk [16, 19, 30, 43], 28 were unclear and 8 were at high risk. Overall, one study was at low risk [30], 21 were unclear and 18 were at high risk. This is consistent with appraisals of machine-learning prediction studies in other clinical areas, which also find most studies at high or unclear risk [61]. The eligibility-stage figure in the original submission labeled all 49 studies as low risk overall. That label meant that a study had passed the threshold, and it should not be read as a worst-domain judgment. Because risk of bias now serves for stratification, no study was removed after re-appraisal.

3.4. RQ1: ZSAD mechanisms and their zero-shot degree

Applying the scale in Table 2 shows that very few studies test zero-shot generalization (Figure 4). One study evaluated strict zero-shot detection together with few-shot adaptation [30]. MVFA adapts the CLIP image encoder with multi-level residual adapters trained on other medical datasets and is tested on an unseen dataset, modality and anatomical region in a leave-one-dataset-out design. One study combined a pathology vision-language model with normal-only training (Z1+FM) and tested it on an external dataset [19]. Nineteen studies were anomaly-label-free (Z1). They learned normality from normal or unlabeled target data with reconstruction, isolation, distance, temporal or self-supervised sequence models, so they detect anomaly types absent from training but are not zero-shot in the sense of Definition 1. Fourteen studies were supervised (Z0), including two that used pseudo-labels or active learning [32, 47]. Five did not report which data and labels were used for training [4, 5, 12, 25, 34].

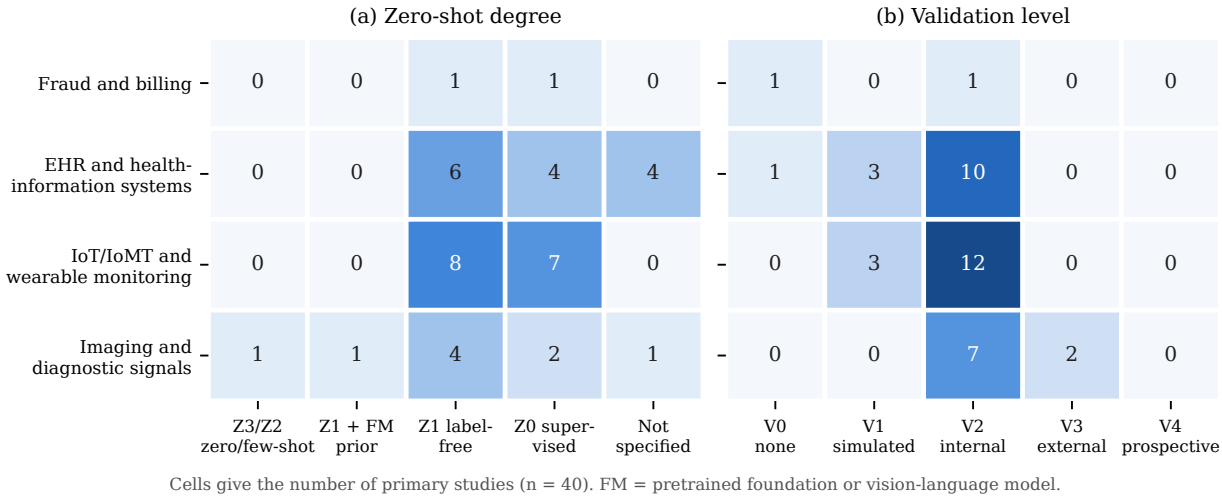


Figure 4. Evidence map of the 40 primary studies: (a) domain by zero-shot degree; (b) domain by validation level

Pretrained foundation or vision-language models were used in three studies, all in imaging [12, 19, 30]. No fraud, EHR or IoT study evaluated zero-shot transfer to an unseen target.

By mechanism family, the primary studies used supervised classifiers (9), generative or reconstruction models (5), embedding or representation models including vision-language and BERT-type encoders (4), temporal deep models (4), isolation and density methods (3), hybrid pipelines (3), graph or attention models (2), distance or clustering methods (2), ontology-driven rules (2), benchmarks of many methods (2), one ensemble classifier, one reinforcement-learning method with active learning, and two studies that did not specify a method. Table 8 gives the family of each study, and the codebook in Supplementary S7 defines the families.

Table 8. Study-level characteristics of the 40 primary studies, ordered by domain (full extraction in Supplementary S1)

Ref.	Year	Domain	Setting	Family	Z	V	Code / data	Headline result (as reported)	RoB D1D2D3 / overall
[14]	2025	Fraud	US billing and insurance	Supervised	Z0	V0	N/N	Acc 0.92	HHH/H
[16]	2025	Fraud	Insurance claims (GP billing)	Isolation, density	Z1	V2	N/N	AUROC 0.82	HUL/H
[3]	2025	EHR/HIS	EHR ML pipeline (outcome prediction)	Supervised	Z0	V2	Y/Y	NR	LUU/U
[4]	2025	EHR/HIS	Big-data pipeline security	Not specified	NS	V1	N/N	Acc 0.97	HHH/H
[5]	2025	EHR/HIS	Ontology-driven access control	Ontology	NS	V1	N/Y	DR 0.94	UHH/H
[7]	2025	EHR/HIS	HIT network traffic across hospitals	Hybrid	Z1	V2	Y/N	NR	UUU/U
[10]	2025	EHR/HIS	ECG abnormality from EHR summary data	Supervised	Z0	V2	N/N	F1 0.63; Acc 0.67	HUU/H
[20]	2024	EHR/HIS	EHR event sequences (HIT)	Embedding	Z1	V2	Y/N	F1 0.80–0.99	UUU/U
[25]	2024	EHR/HIS	Healthcare data warehouses	Not specified	NS	V0	N/N	NR	HHH/H
[32]	2024	EHR/HIS	EHR access audit logs	Hybrid	Z0	V2	N/N	F1 0.99; Acc 0.992	HHU/H

Ref.	Year	Domain	Setting	Family	Z	V	Code / data	Headline result (as reported)	RoB D1D2D3 / overall
[34]	2024	EHR/HIS	Cloud EHR access control	Ontology	NS	V1	N/Y	NR	UHH/H
[39]	2023	EHR/HIS	Disease-surveillance records	Isolation, density	Z1	V2	Y/Y	NR	LUU/U
[41]	2023	EHR/HIS	Clinical tabular data	Isolation, density	Z1	V2	N/Y	AUROC 0.78	UUU/U
[47]	2025	EHR/HIS	Cardiovascular clinical records	RL+AL	Z0	V2	N/N	AUROC 0.75; F1 0.56	HUU/H
[48]	2025	EHR/HIS	NHS quality-indicator time series	Distance, clustering	Z1	V2	N/N	NR	HUU/H
[49]	2025	EHR/HIS	Bioinformatics pipeline syslogs	Temporal DL	Z1	V2	N/N	F1 0.91	HUU/H
[1]	2025	IoT	Wearable single-lead ECG	Generative	Z1	V2	Y/Y	F1 0.94; Acc 0.91	LUU/U
[6]	2025	IoT	IoT medical-device streams	Temporal DL	Z1	V2	N/N	AUROC 0.91; F1 0.86	HUU/H
[13]	2025	IoT	IoT security and integrity	Generative	Z1	V1	N/N	AUROC 0.92; F1 0.84; Acc 0.95	HHH/H
[17]	2025	IoT	Elderly monitoring, non-wearable sensors	Hybrid	Z1	V2	N/N	AUROC 0.95; F1 0.86; Acc 0.92	HUU/H
[18]	2025	IoT	IoMT cybersecurity	Graph, attention	Z0	V2	N/Y	AUROC 0.998; F1 0.991; Acc 0.98	UUU/U
[21]	2024	IoT	IoMT intrusion detection	Supervised	Z0	V2	N/Y	AUROC 0.93; Acc 0.98	UUU/U
[22]	2024	IoT	WBAN interference/malicious nodes	Supervised	Z0	V1	N/N	NR	HHH/H
[27]	2024	IoT	IoT/IoMT intrusion detection	Ensemble	Z0	V2	N/Y	Acc 0.98	UUU/U
[31]	2024	IoT	WBAN vital signs	Generative	Z1	V2	N/Y	F1 0.96; Acc 0.93	UUU/U
[33]	2024	IoT	IoT intrusion detection	Supervised	Z0	V2	N/Y	Acc 0.996	UUU/U
[35]	2023	IoT	Healthcare IoT security (sensor data)	Supervised	Z0	V2	N/Y	F1 0.97; Acc 0.99	UUU/U
[37]	2023	IoT	Medical IoT traffic	Supervised	Z0	V1	N/N	NR	HHH/H
[43]	2022	IoT	Home sensors, dementia	Distance, clustering	Z1	V2	N/N	Recall 0.84	HUL/H
[44]	2022	IoT	Cloud ECG monitoring	Temporal DL	Z1	V2	N/Y	Acc 0.98	UUU/U
[46]	2022	IoT	WBAN physiological data	Temporal DL	Z1	V2	N/Y	F1 0.98; Acc 0.990	UUU/U
[9]	2025	Img/sig	Digital pathology (benchmark)	Benchmark	Z1	V2	Y/Y	NR	LUU/U
[11]	2025	Img/sig	EEG	Graph, attention	Z0	V2	N/Y	Acc 0.93	UUU/U
[12]	2025	Img/sig	Multimodal (imaging + EHR)	Embedding	NS	V2	N/Y	AUROC 0.93; F1 0.88	UUU/U
[19]	2025	Img/sig	Pathology WSIs	Embedding	Z1+FM	V3	N/Y	AUROC 0.97 (int.); 0.997 (ext.)	ULL/U
[24]	2024	Img/sig	Chest X-ray, fundus	Generative	Z1	V2	Y/Y	AUROC 0.92	LUU/U
[28]	2024	Img/sig	Six imaging datasets (benchmark)	Benchmark	Z1	V2	Y/Y	AUROC 0.92	LUU/U
[30]	2024	Img/sig	Six imaging datasets, five modalities	Embedding	Z3/Z2	V3	Y/Y	AUROC 0.80 (ZS); 0.89 (4-shot)	LLL/L

Ref.	Year	Domain	Setting	Family	Z	V	Code / data	Headline result (as reported)	RoB D1D2D3 / overall
[40]	2023	Img/sig	Chest X-ray, multi-annotator	Supervised	Z0	V2	Y/Y	NR	LUU/U
[42]	2023	Img/sig	Mammography	Generative	Z1	V2	N/N	AUROC 0.70; Acc 0.64	HUU/H

Z: zero-shot degree (Table 2); +FM: pretrained foundation-model prior; NS: not specified. V: validation level (Definition 3). Code/data: public availability (Y/N). Acc: accuracy; DR: detection rate; ZS: zero-shot; int./ext.: internal/external test; NR: no extractable headline metric. RoB: L low, U unclear, H high.

3.5. RQ2: Comparative performance across domains and zero-shot degrees

Metric reporting. Headline AUROC values were reported by 14 primary studies, F1 by 13 and accuracy by 17. Twelve studies reported accuracy without any threshold-free metric, even though accuracy is uninformative when anomalies are rare. Precision-recall summaries (AUPRC or average precision) were reported or discussed in 7 studies, and 10 studies reported no extractable headline metric for anomaly detection. Because the metrics are not interchangeable, Table 9 summarizes each one separately. The original submission stated that ensemble and deep-learning models achieved 85–95% accuracy and reported pooled means and standard deviations across metrics. Those statements mixed AUROC, F1 and accuracy and could not be traced to a reproducible rule, so they have been removed and replaced by the study-level values in Table 8 and the stratified summaries in Table 9.

Table 9. Reported performance by zero-shot degree and by domain: median [range] (number of studies reporting the metric). Values are unweighted descriptive summaries of headline results and not pooled estimates

Stratum (primary studies)	AUROC	F1	Accuracy
By zero-shot degree			
Z3/Z2 zero/few-shot (n = 1)	0.80 zero-shot; 0.89 at k = 4 (1)	–	–
Z1+FM label-free with FM prior (n = 1)	0.97 internal; 0.997 external (1)	–	–
Z1 anomaly-label-free (n = 19)	0.91 [0.70–0.95] (8)	0.91 [0.84–0.98] (7)	0.93 [0.64–0.99] (7)
Z0 supervised (n = 14)	0.93 [0.75–0.998] (3)	0.97 [0.56–0.991] (5)	0.98 [0.67–0.996] (9)
Not specified (n = 5)	0.93 (1)	0.88 (1)	0.97 (1)
By domain			
Fraud and billing (n = 2)	0.82 (1)	–	0.92 (1; V0)
EHR and health-information systems (n = 14)	0.77 [0.75–0.78] (2)	0.77 [0.56–0.99] (4)	0.97 [0.67–0.99] (3)
IoT and wearable monitoring (n = 15)	0.93 [0.91–0.998] (5)	0.95 [0.84–0.991] (8)	0.98 [0.91–0.996] (11)
Imaging and diagnostic signals (n = 9)	0.92 [0.70–0.97] (6)	0.88 (1)	0.78 [0.64–0.93] (2)
All primary studies (n = 40)	**0.92 [0.70–0.998] (14)**	**0.91 [0.56–0.991] (13)**	**0.97 [0.64–0.996] (17)**

Three patterns stand out. First, the highest values come from supervised classification of known attack classes on IoT testbed data. Examples are accuracy of 0.996 on CICIoT2023 [33] and AUROC of 0.998 on CICIoMT2024 [18]. These are closed-set tasks with random internal splits, so they say little about detection of attacks that were absent from training. Second, the only strict zero-shot evaluation produced lower values (mean AUROC 0.80 over six unseen datasets) under the most demanding design [30]. The level of a reported number therefore depends at least as much on the evaluation design as on the method. Third, EHR and health-information-system studies reported the lowest threshold-free values (AUROC 0.75 and 0.78 [47, 41]), and fraud contributed a single AUROC (0.82 [16]). Both domains lack shared benchmarks, which leaves their evidence thin.

3.5.1. Sensitivity analyses Table 10 reports the six sensitivity analyses. Removing studies without real-data evaluation, removing preprints, or removing studies at high risk of performance overestimation changed the median

AUROC and F1 by at most 0.015. The median accuracy fell from 0.973 to 0.930 when IoT security studies with testbed attack labels were removed, and it was 0.930 among label-free studies against 0.983 among supervised ones. The high accuracies in the literature therefore come mainly from closed-set supervised tasks, and the threshold-free summaries are stable.

Table 10. Sensitivity analyses: median of reported headline values (number of studies reporting the metric)

Analysis	**Studies**	**AUROC**	**F1**	**Accuracy**
S0 All primary studies	40	0.918 (14)	0.910 (13)	0.973 (17)
S1 Excluding studies without real-data evaluation (V0, V1)	32	0.916 (13)	0.925 (12)	0.979 (14)
S2 Excluding preprints	38	0.916 (13)	0.910 (13)	0.973 (17)
S3 Excluding studies at high risk in D2	31	0.916 (13)	0.910 (11)	0.979 (13)
S4 Label-free, few-shot and zero-shot only (Z1–Z3)	21	0.913 (10)	0.910 (7)	0.930 (7)
S5 Supervised only (Z0)	14	0.929 (3)	0.967 (5)	0.983 (9)
S6 Excluding IoT security studies with testbed attack labels	36	0.913 (12)	0.897 (12)	0.930 (13)

3.5.2. Zero-shot versus few-shot and hybrid adaptation Only MVFA [30] reports zero-shot and few-shot results for the same model and targets, so it is the only source for a direct comparison. In the zero-shot setting its image-level AUROC on the six unseen datasets was 0.78 (histopathology), 0.71 (chest X-ray), 0.95 (OCT17), 0.79 (brain MRI), 0.76 (liver CT) and 0.83 (RESC), with a mean of 0.80. With four labeled shots per target ($k = 4$, normal and abnormal images) the values rose to 0.83, 0.82, 0.99, 0.92, 0.81 and 0.96, with a mean of 0.89. The gain ranged from 0.04 to 0.14 AUROC across targets. Four labeled images cost far less than full annotation, so few-shot adaptation is the better choice whenever a handful of confirmed examples can be collected. Zero-shot detection remains the only option when no confirmed example exists yet, as with a new fraud scheme or a previously unseen device fault, and its performance then sets the floor that a deployed system must tolerate. The original submission described hybrids as typically adding 0.03 to 0.07 AUROC. That range could not be traced to the included studies and has been replaced by the values above.

3.5.3. Failure modes The included studies also document conditions under which label-efficient detection fails or its results cannot be interpreted:

- *Domain gap in vision-language priors.* Under zero-shot transfer, WinCLIP, which relies on prompts designed for natural images, scored below chance on two retinal datasets (AUROC 0.47 on OCT17 and 0.43 on RESC). Its average AUROC fell from 0.95 on industrial images to 0.76 on medical images in the few-shot setting [30].
- *Inadequate representation of categorical data.* For insurance claims with high-cardinality categorical variables, one-hot encoding gave AUROC between 0.51 and 0.60 for every detector, against 0.69 to 0.82 with learned embeddings [16].
- *Weak signal in coarse features.* Accuracy was 0.67 for ECG abnormality detection from summary measurements [10], a StyleGAN2 detector reached AUROC 0.70 with a specificity of 0.52 in mammography [42], and reinforcement learning with active learning reached F1 0.56 and AUROC 0.75 on cardiovascular records [47].
- *Evaluation artefacts.* When classifiers were trained and tested on labels produced by Isolation Forest, their accuracy of 0.99 measured agreement with that detector and not detection of true anomalies [32]. Six studies evaluated only on simulated, synthetic or non-health data [4, 5, 13, 22, 34, 37], and one evaluated on injected anomalies in real system data [7].
- *Unmeasured alarm burden.* Only one study reported an alert rate (84.3% recall at an alert rate of 5.1%) [43]. Without this information, the effect of a detector on staff workload cannot be judged.

3.5.4. Benchmark datasets and evaluation protocols Table 11 lists the datasets used by the primary studies. Imaging is the only domain with shared public benchmarks and reproducible splits, such as the BMAD collection [28] and the leave-one-dataset-out benchmark used for MVFA [30]. Physiological monitoring relies mainly on MIMIC and PhysioNet signals, and the anomaly labels there are often derived from thresholds or injected faults rather than from clinical events. IoT security relies on laboratory testbeds with attack labels (CICIoMT2024, CICIoT2023, WUSTL-EHMS-2020). Evaluating the same detector on several intrusion datasets, as done for recurrent models in [89], is a first step toward cross-dataset testing, but training and testing on the same attack families remains the norm. EHR and fraud studies used private institutional data, which cannot be shared. No public healthcare-claims dataset with confirmed anomalies was used by any included study. As a result, evaluation of zero-shot transfer is currently possible only in imaging, and the absence of shared benchmarks is the main obstacle to cumulative evidence in the other three domains.

Table 11. Datasets used by the primary studies and the resulting evaluation gaps

Domain	**Datasets (studies)**	**Evaluation gap**
Fraud and billing	Private Belgian insurer claims [16]; dataset not reported [14]	No public claims benchmark with confirmed anomalies
EHR and health-information systems	MIMIC-IV and eICU [3]; MIMIC-III with simulated access patterns [34]; private VA EHR sequences and hospital-network traffic [7, 20]; private hospital audit logs [32]; regional ECG summary records [10]; public malaria-surveillance data [39]; ten public medical tabular datasets [41]; NHS primary-care indicators [48]; cardiovascular records [47]; simulated or non-health data [4, 5]; real and simulated pipeline logs [49]	Mostly private data; anomalies often injected or simulated; no shared EHR anomaly benchmark
IoT: physiological monitoring	Five open ECG datasets and Icentia11k [1]; MIMIC and MIMIC-II via PhysioNet [31, 46]; MIT-BIH and other ECG data [44]; UK Dementia Research Institute home-sensor cohort [43]; smart-home sensor data [17]; unnamed device streams [6]	Clinically confirmed events only in [43]; heterogeneous anomaly definitions
IoT: device and network security	CICIoMT2024 [18]; CICIoT2023 [33]; WUSTL-EHMS-2020 [21]; two public IoT security datasets [27]; PhysioNet sensor data [35]; simulated traffic or networks [13, 22, 37]	Closed-set attack labels; random splits; no test on unseen attack families
Imaging and diagnostic signals	BMAD: BraTS2021, BTCV and LiTS, RESC, OCT2017, RSNA, Camelyon16 [28]; BrainMRI, LiverCT, RESC, OCT17, ChestXray, HIS [30]; GastricLN with external Camelyon16 [19]; five pathology datasets [9]; RSNA, VinDr-CXR, LAG [24]; VinDr-CXR [40]; IU X-ray and VinDr-CXR [12]; TUH Abnormal EEG [11]; institutional mammograms [42]	Shared benchmarks exist; external and cross-modality tests possible but rarely run

3.5.5. Computational requirements and deployment feasibility Sixteen primary studies reported hardware, latency, energy or run-time information. The vision-language approaches were designed to be light at deployment. MVFA adds lightweight residual adapters to a pretrained CLIP image encoder and was trained on a single consumer GPU [30], and Ano-NAViLa trains a small multilayer perceptron on frozen pathology embeddings to cope with gigapixel slides [19]. Edge and fog architectures were proposed for monitoring, either to keep data local or to reduce latency [17, 6]. A lightweight matrix-profile method was designed for continuous home monitoring [43], and an energy-efficient classifier was designed for body-area sensor nodes [22]. Streaming security frameworks reported detection latencies of about 500 ms [4]. None of the studies compared accuracy and computational cost across method families under the same conditions, so we cannot rank the families by deployment feasibility. Two qualitative conclusions are supported. Frozen-backbone vision-language methods move most of the cost to pretraining and

keep target-side cost low. Temporal and one-class models are small enough for edge devices but must be retrained or recalibrated locally, which conflicts with the goal of avoiding target-specific training.

3.6. Explainability: what the evidence shows

The original submission stated that SHAP or LIME were used in 28.57% of the studies and that they reduced false positives by 12–18%. The re-extraction does not support either statement, and both have been removed. Post hoc attribution with SHAP was applied in two primary studies [16, 18]. Two more mentioned SHAP or LIME without reporting attribution results [6, 14]. Five used intrinsic or model-specific interpretability: disentangled latent factors [1], knowledge-guided attention [12], text-based explanations from a vision-language model [19], outlying-aspect mining [41] and matrix-profile motifs reviewed by a clinical monitoring team [43]. Outside the included set, saliency methods such as Grad-CAM are a common route to interpretability in medical image classification [90]. No included study compared false-positive rates, alert burden, user trust or generalization with and without explanations. Any statement that explainability improves these outcomes is therefore a hypothesis for future controlled studies, and the evidence reviewed here does not establish it.

3.7. ZSAD classification framework

To organize heterogeneous approaches, the review proposes a framework with three axes (Figure 5). Axis A is the zero-shot degree of Table 2. Axis B is the mechanism family: embedding and prompt alignment, generative and reconstruction models, isolation, density, distance and temporal models, and graph, knowledge and ontology reasoning, with hybrids recorded as combinations. Axis C is the three-level evaluation taxonomy (bibliographic, methodological and clinical) with a risk-of-bias and certainty overlay. The framework differs from a list of method families in two ways. Its categories are defined by checkable decision rules, and every study receives exactly one code on each axis, so two reviewers can apply it and compare their codes.

Table 12 states what the framework inherits from existing taxonomies, what it revises and what it adds. The method axis inherits the reconstruction, one-class and density categories of general anomaly-detection reviews [69, 70, 71]. It revises them by separating representation learning with external pretrained models from representation learning on target data, because that separation decides whether zero-shot transfer is possible at all. The zero-shot axis is new for healthcare. It adapts the zero-shot learning protocol [73] to anomaly detection and adds the few-shot and label-free levels that ZSL taxonomies do not need. The validation levels inherit the staged logic of DECIDE-AI [60] and add an explicit level for simulated or non-health data, which is common in IoT security studies. The appraisal overlay inherits PROBAST+AI and TRIPOD+AI [55, 56] and adds the unseen-anomaly and synthetic-anomaly questions.

Case-based verification. All 49 included studies were coded on the three axes. On the zero-shot axis, 35 of the 40 primary studies could be coded, and five could not because the regime was not reported [4, 5, 12, 25, 34]. Thirty-one were coded directly from the definitions. Three required a decision rule that the original categories did not contain: pseudo-labels from an unsupervised detector were coded Z0 [32], active-learning label queries were coded Z0 [47], and an unsupervised embedding with a labeled k-nearest-neighbor head was coded Z1 with a note [1]. One more study used unlabeled images that contained both normal and abnormal cases and was coded Z1 [24]. These rules are now part of the codebook (Supplementary S7). The framework could therefore classify every study that reported its training data, and the cases it could not classify point to a reporting gap rather than a gap in the framework. Independent coding by a second reviewer is the next step of validation.

3.8. Thematic synthesis

Table 13 summarizes the themes mapped to the research questions. Each code is defined in the codebook (Supplementary S7), and the number of primary studies supporting it is given, so that the strength of each theme can be judged.

3.9. Certainty of evidence

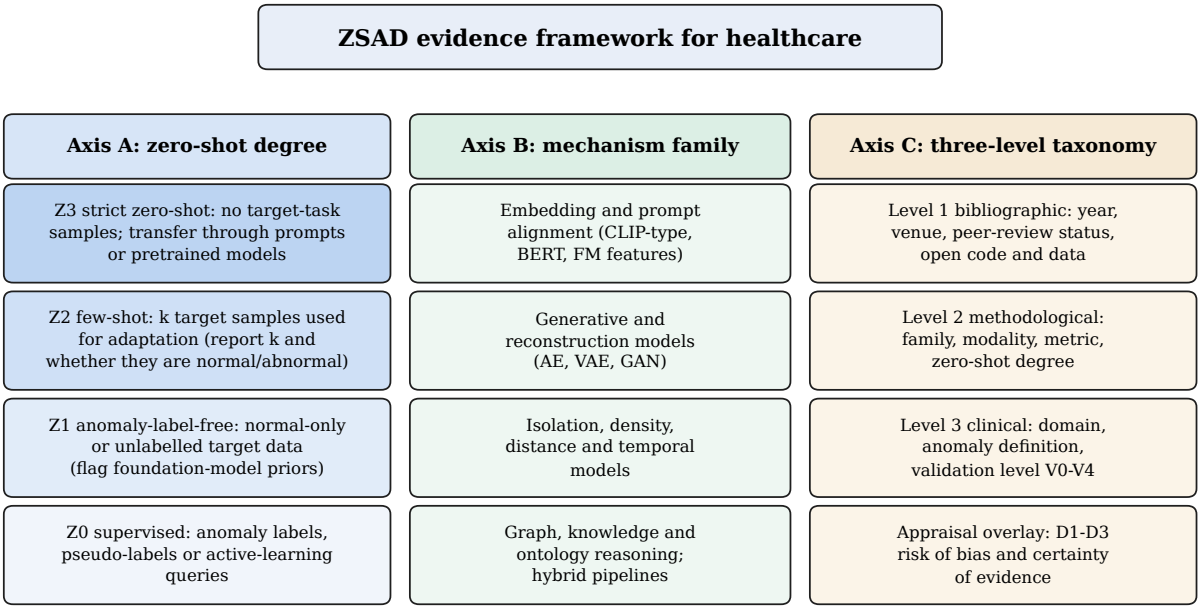


Figure 5. ZSAD evidence framework: zero-shot degree, mechanism family and the three-level evaluation taxonomy with an appraisal overlay

Table 12. Relation of the proposed framework to existing taxonomies and evaluation frameworks

Framework element	**Inherited from**	**Revised**	**Newly added**
Axis A: zero-shot degree	Zero-shot learning protocols [73, 74]; ZSAD and few-shot AD settings [63, 64, 94]	Target defined as dataset, site or modality rather than class	Four-level scale Z0–Z3 with decision rules; foundation-model flag; pseudo-label and active-learning rules
Axis B: mechanism family	Reconstruction, one-class, density and distance families [69, 70, 71]	Representation learning split into external pretraining and target-data learning	Knowledge and ontology reasoning and graph models as explicit families for EHR and claims data
Axis C, level 2–3: validation	DECIDE-AI staged evaluation [60]	Applied to retrospective benchmark studies	V1 level for simulated or non-health data; V3 includes leave-one-domain-out designs
Appraisal overlay	PROBAST, PROBAST+AI, TRIPOD+AI [54, 55, 56]; GRADE [59]	Three domains instead of four	Signalling questions on unseen anomaly types, synthetic anomalies and circular labels
Governance indicators	EU AI Act, FDA GMLP and PCCP, GDPR, HIPAA, NIST AI RMF [80, 81, 82, 83, 84, 86]	Requirements expressed as measurable indicators	Each indicator linked to the number of included studies that already report it (Table 15)

Table 14 rates the certainty of the main findings. No finding reached moderate or high certainty. The commonest reasons for downgrading were risk of bias, indirectness (most evidence comes from one domain or from internal splits) and imprecision (one or two studies).

Table 13. Thematic synthesis aligned to the three research questions (number of primary studies supporting each code)

Research question	Theme	Codes (number of studies)	Representative evidence
RQ1	Label-free normality modeling dominates; strict zero-shot transfer is rare	Anomaly-label-free training (19); foundation-model semantic transfer (3); strict zero-shot evaluation (1); few-shot evaluation (1); supervised closed-set classification (14); regime not reported (5)	[30, 19, 12, 16, 20, 43]
RQ2	Reported performance depends on evaluation design	Internal split only (30); simulated or non-health data (6); external or cross-dataset test (2); accuracy without a threshold-free metric (12); circular labels (1); domain gap in vision-language priors (1)	[30, 33, 18, 32, 5]
RQ2	Data structure decides which mechanism works	Temporal dependence modeling of physiological signals and sensor streams (9); relational and graph structure for EHR events, EEG channels and networks (4); embeddings for high-cardinality categorical claims (1)	[16, 20, 7, 44, 46]
RQ3	Translational and governance gaps	No prospective evaluation (40 of 40 without); no subgroup or fairness analysis (40 of 40 without); alert burden reported (1); privacy or regulation discussed (HIPAA or GDPR 7; federated or differential privacy 12, none evaluated); public code (10)	[43, 4, 5, 34, 17]

Table 14. Certainty of the main findings (GRADE-adapted)

Finding	Studies	Reasons for downgrading	Certainty
Strict zero-shot detection on unseen medical-imaging datasets reaches moderate discrimination (mean AUROC 0.80; 0.71–0.95)	1	Imprecision (single study); indirectness for fraud, EHR and IoT	Low (imaging); very low (other domains)
Four labeled shots improve on zero-shot detection (mean AUROC +0.085)	1	Imprecision (single study)	Low
Anomaly-label-free detectors report AUROC 0.70–0.95 (median 0.91)	8	Risk of bias (internal validation); inconsistency; indirectness to unseen targets	Low
Supervised IoT intrusion detectors report accuracy of 0.97 or higher	5	Risk of bias; indirectness (closed-set testbed attacks rather than unseen anomalies)	Very low
Explainability reduces false alarms or improves trust	0 comparative	No comparative evidence	Very low (no evidence)
Detectors improve clinical or financial outcomes in practice	0 prospective	No prospective or deployed evaluations	Very low (no evidence)

4. Discussion

4.1. Principal findings

This review set out to compare zero-shot anomaly detection across fraud detection, EHR and health-information systems, and IoT monitoring. Its main finding is that the healthcare literature labeled as zero-shot or label-efficient rarely tests zero-shot transfer. Of 40 primary studies, one evaluated strict zero-shot detection, and it did so in medical imaging [30]. Another combined a pathology foundation model with normal-only training and an external test [19]. The remaining studies were label-free one-class detectors (19), supervised classifiers (14) or did not report their regime (5). Two studies tested on data from a source not used in training, and none was prospective.

The original submission described ZSAD as having transformative potential for healthcare integrity and patient safety. The evidence supports a narrower statement: zero-shot transfer is feasible for medical images with moderate discrimination, and in fraud, EHR and IoT data it has not yet been tested.

These results do not show that zero-shot methods are ineffective. They show that the evaluations needed to judge them have mostly not been done. Reported performance was high across all strata (median AUROC 0.92), but most values come from internal splits, and the highest come from closed-set supervised tasks. The only leave-one-dataset-out evaluation reported a mean AUROC of 0.80. This gap between internal and cross-domain results agrees with the general observation that machine-learning studies in medicine are at high or unclear risk of bias and tend to overstate performance [61].

4.2. Comparative insights across domains

The included studies suggest that data structure, more than the domain label, decides which mechanism works. For physiological streams, temporal models such as LSTM autoencoders and ConvLSTM networks reported F1 of 0.86 to 0.98 [6, 46] and accuracy of 0.98 [44] under internal validation, because anomalies in these data are deviations in dynamics and in correlations between channels. For EHR event sequences and hospital-network traffic, contextual and relational models (EHR-BERT and graph analysis across hospitals) captured irregularities that single-event rules miss [20, 7], but they were evaluated on private Veterans Affairs data with injected or unconfirmed anomalies. For insurance claims, the decisive step was the representation of high-cardinality categorical variables. Learned embeddings raised AUROC from at most 0.60 to 0.82, and SHAP attributions were used to explain flagged providers to investigators [16]. For images, vision-language priors were the only route to zero-shot transfer, and they worked better when adapted to medical features [30, 19].

Ensemble and deep models did not consistently outperform simpler detectors. In claims data, classical one-class SVM was among the best methods once the representation was fixed [16]. In clinical tabular benchmarks, Isolation Forest had the best average AUROC (0.78) [41]. Random forests were competitive with a graph-transformer hybrid on IoMT traffic [18]. Heterogeneity, temporal dependence and class imbalance in healthcare data favor methods that model normality well. They do not by themselves favor deeper models, and the original claim of a general superiority of ensembles and deep learning has therefore been withdrawn.

How the three domains can reinforce one another. The evidence for integration across domains is conceptual, since no included study linked anomalies across claims, records and devices. Three mechanisms are nevertheless plausible and testable. First, a shared model of normality: a device reading, the EHR entry it produces and the claim that bills it describe the same care event, so an inconsistency between them is itself an anomaly signal that no single stream contains. Second, shared drift monitoring: software updates, coding changes and new devices shift all three streams at once, and drift detected in one stream can trigger recalibration in the others. Third, shared semantic priors: text descriptions of normal and abnormal clinical states, which drive zero-shot scoring in imaging [30, 19], could be reused to score EHR notes and claims narratives. Testing any of these mechanisms requires linked data from the same patients, which none of the current benchmarks provides.

4.3. Zero-shot, few-shot and hybrid approaches: costs and benefits

The single direct comparison available shows that four labeled shots per target raised mean AUROC from 0.80 to 0.89 and that the gain varied from 0.04 to 0.14 between targets [30]. In practice the choice depends on whether confirmed examples exist. When a new anomaly type has already been observed a few times, for example a confirmed fraud pattern or a known device fault, the labeling cost of few-shot adaptation is small and the expected gain is substantial. When the anomaly type has not yet been observed, few-shot adaptation is impossible, and the zero-shot performance is what the system will deliver. A hybrid design that runs zero-shot scoring from day one and switches to few-shot adaptation as confirmed cases accumulate combines the two. It also creates a governance requirement, because every adaptation changes the deployed model (Section 4.5).

4.4. Foundation models beyond CLIP and deployment feasibility

All three foundation-model studies used contrastive vision-language encoders [30, 19, 12]. Other encoders are relevant but untested in the included healthcare studies. These include self-supervised visual features (DINOv2) [65], sigmoid-loss language-image models (SigLIP) [66], biomedical vision-language models (BiomedCLIP) [67], pathology-specific models (CONCH, used as the backbone of Ano-NAViLa) [68, 19] and in-context generalist detectors [94]. For fraud, EHR and IoT data the bottleneck is different. Pretrained models for claims, event logs and multivariate physiological streams are much less mature than image-text models, so zero-shot transfer in these domains will depend on such models and on benchmarks that test them. Regarding computation, frozen-backbone methods shift cost to pretraining and are inexpensive at the target site, whereas one-class models are small but need local training data and periodic recalibration. Deployment decisions should weigh these costs against the validation evidence, which is currently strongest for frozen-backbone imaging models.

4.5. Ethical, institutional, and governance considerations

The original discussion of ethics and governance listed principles without saying how compliance could be checked. Table 15 translates the main requirements of the EU Artificial Intelligence Act [80], the FDA, Health Canada and MHRA Good Machine Learning Practice principles [81], the FDA guidance on predetermined change control plans [82], the GDPR [83], the HIPAA Security Rule [84], the NIST AI Risk Management Framework [86] and the WHO guidance on AI for health [87] into audit indicators. For each indicator it reports how many of the 40 primary studies already provide the evidence the indicator needs, assigns a responsible role and proposes an illustrative deployment threshold. The thresholds are examples to be set by each institution, and they have not been validated.

Two accountability boundaries follow from the table. Developers are accountable for documentation, validation evidence and change-control plans. Deploying institutions are accountable for local validation, alert handling, subgroup monitoring and the decision to act on an alert. In every domain reviewed here an alert should flag a case for human review and should never trigger an automatic sanction, denial of payment or change in treatment. Most indicators in Table 15 are not met by the current evidence. Closing these gaps is a precondition for responsible deployment, and the indicators can serve as a reporting checklist for future studies.

4.6. Comparison with previous reviews

Compared with the reviews in Table 1, this review differs in four respects. It applies one protocol to fraud, EHR and IoT monitoring, while earlier reviews covered one data type or method family [29, 36, 38, 45, 85]. It separates zero-shot, few-shot, label-free and supervised detection with checkable rules, which none of the earlier reviews did. It appraises risk of bias with criteria adapted to machine-learning evaluations and rates the certainty of its conclusions. Finally, it links governance requirements to measurable indicators and to the evidence that studies actually report. Its conclusions are also more cautious than those of earlier narrative reviews, because the appraisal shows how much of the reported performance rests on internal validation.

5. Limitation

The review process has several limitations. The review was not registered and no protocol was published in advance. The stratification by zero-shot degree and validation level, the full-text re-appraisal and the sensitivity analyses were added during revision, so they are post hoc and were applied by authors who knew the studies. Screening was done by one reviewer, and extraction agreement was checked on a 20% sample without a formal statistic, so selection and coding errors cannot be excluded. To mitigate this, every coding decision is reported in Supplementary S1 and S5 for independent checking. One hundred reports could not be retrieved, and their content is unknown. The risk-of-bias threshold excluded 246 of 295 full-text reports without extraction of their results, so the effect of the threshold on the performance summaries can only be reasoned about and cannot be measured (Section 3.3). The search used zero-shot terms with domain and enabling terms but not the vocabulary of out-of-distribution detection, open-set recognition or novelty detection. It covered NeurIPS, ICML and ICLR only through Google Scholar, which was

Table 15. Governance requirements mapped to audit indicators, evidence in the included studies, responsibility and illustrative deployment thresholds

Requirement (source)	**Audit indicator**	**Evidence in primary studies (of 40)**	**Responsible role**	**Illustrative deployment threshold**
Representative data and independent test sets (AI Act Art. 10; GMLP 3–5)	Test on a dataset, site or anomaly type not used in training; dataset description	External or cross-dataset test: 2	ML lead with data steward	Local retrospective validation before go-live, with AUROC loss within a pre-specified margin (for example, no more than 0.05) of the development estimate
Technical documentation and record-keeping (Art. 11–12; TRIPOD+AI)	Public or escrowed code, model version log, data provenance	Public code: 10; code and data: 8	Developer or vendor	Model card and version log approved by the oversight committee
Transparency and human oversight (Art. 13–14; GMLP 7, 9)	Explanation available per alert; human review before action; override logging	Post hoc attribution: 2; any interpretability: 7	Clinical or audit owner	Every alert reviewable by a named human before any action; overrides logged
Accuracy, robustness and cybersecurity (Art. 15)	Performance on unseen anomaly types; robustness to drift and attack	Unseen-target test: 2; drift addressed in design: 1	ML lead with IT security	Monitored metrics with an agreed recalibration trigger (for example, a sustained drop in precision at the operating threshold)
Workload and alarm burden (GMLP 7–8; WHO)	Alert rate per patient-day or per 1,000 claims; precision at the operating point	Alert rate reported: 1	Clinical or audit owner	Alert volume within staffed review capacity agreed in advance
Fairness and equity (WHO; NIST AI RMF)	Performance by sex, age, ethnicity and provider type	Subgroup analysis: 0	Oversight committee	Pre-specified maximum gap between subgroups before deployment
Change control and post-market monitoring (Art. 72; FDA PCCP; GMLP 10)	Documented change-control plan for retraining and few-shot adaptation	Prospective or deployed evaluation: 0	Developer with oversight committee	Adaptations only within the bounds of an approved change-control plan, each followed by revalidation
Privacy and lawful processing (GDPR Art. 5, 25, 35; HIPAA Security Rule)	Data-protection impact assessment; data minimization; local or federated processing	HIPAA or GDPR discussed: 7; edge processing designed: 2; federated evaluation: 0	Data protection officer	Impact assessment completed before processing; no identifiable data leave the institution without a legal basis

screened up to 200 results per query, and it did not search technical reports or industry white papers. Studies in other languages and studies published before 2022 or after 2025 were not included, and the field is changing quickly.

The evidence base has its own limitations. Only one study evaluated strict zero-shot transfer and only two used external validation, so the main findings rest on very few studies. Fraud detection contributed two primary studies, one of which did not report its data, so conclusions for fraud are of very low certainty. Metrics, datasets and anomaly definitions were heterogeneous, which prevented meta-analysis. Selective reporting of the best model was common, and publication bias could not be assessed formally.

6. Conclusion

6.1. Conclusion

This systematic review examined ZSAD in healthcare across fraud and billing, EHR and health-information systems, and IoT monitoring, with medical imaging as a fourth stratum. Among 40 primary studies, one tested strict zero-shot detection and one combined foundation-model priors with normal-only training. The rest relied on label-free one-class models, supervised classifiers or unreported regimes. Reported performance was high under internal validation, but the only cross-dataset zero-shot evaluation reached a mean AUROC of 0.80, and four labeled shots raised it to 0.89. No study was prospective, none reported subgroup performance, and none measured the effect of explanations on false alarms. The certainty of every main finding is low or very low.

The review contributes an operational definition and a four-level zero-shot scale that make claims of zero-shot detection testable, a study-level evidence map with risk-of-bias and certainty ratings, a framework that states its relation to existing taxonomies, and governance indicators linked to the evidence. The evidence supports cautious optimism for zero-shot transfer in medical imaging. For fraud, EHR and IoT monitoring it identifies what still has to be shown, and it describes how such evaluations should be designed and reported.

6.2. Implications for practice and policy

- *Healthcare practitioners and auditors* should treat current ZSAD and label-free detectors as triage tools that rank cases for human review. Before relying on a detector, they should ask for local validation and for the alert rate at the intended operating point.
- *Healthcare institutions* should set up an oversight committee that applies the indicators in Table 15, including subgroup performance checks and drift monitoring, and should start with silent-mode pilots in which alerts are recorded but not acted on, before any scaled rollout [60].
- *Policymakers and regulators* can require that claims of zero-shot or generalizable detection state the zero-shot degree and the validation level (Definitions 2 and 3). They can also treat few-shot adaptation in deployed systems as a model change that falls under a change-control plan [82].
- *Technology developers* should report AUROC together with AUPRC and the alert rate, release code or provide escrowed access, and test on at least one external site or unseen anomaly type before claiming generalization.
- *Privacy and data governance* require impact assessments, data minimization and, where data cannot leave an institution, local or federated designs. Such designs should be evaluated rather than only proposed [83, 84].

6.3. Future research

The following proposals are specific enough to be tested.

i. *Public benchmarks for the non-imaging domains.* EHR event logs and claims datasets with confirmed anomalies, released with leave-one-site-out and leave-one-anomaly-type-out splits, would make zero-shot evaluation possible in the domains where it is now missing. Linked device, EHR and claims data for the same patients would allow the cross-domain mechanisms in Section 4.2 to be tested.
ii. *Shot-response curves.* Studies should report performance at $k \in \{0, 1, 2, 4, 8, 16\}$ shots on the same targets, so that the cost-benefit trade-off between zero-shot and few-shot detection can be estimated across domains rather than from one imaging study.
iii. *Prospective silent-mode evaluations* in at least one domain, reporting precision at the operating point, alert burden and the time staff spend on review.
iv. *Controlled tests of explainability*, comparing alert handling, false-positive dismissal and trust with and without explanations.
v. *Subgroup reporting* by sex, age and ethnicity for clinical data, and by provider type for claims.
vi. *Non-image foundation models*, such as models pretrained on clinical event sequences or multivariate physiological signals, evaluated for zero-shot anomaly scoring with leave-one-site-out designs.

vii. *Validation of the composite generalization index* proposed below against expert rankings or downstream outcomes.

6.3.1. A composite generalization index (theoretical proposal) The original submission proposed a composite metric of zero-shot robustness. Its weights were not justified and its components were on different scales. We therefore reformulate it here as a theoretical proposal with normalized components, a default weighting rule, a weight-sensitivity analysis and a worked example. Validation remains future work. For a detector evaluated on T unseen targets, let

$$\mathrm{AGI} = w_1 Z + w_2 S + w_3(1 - D) + w_4 E, \qquad \sum_{i=1}^{4} w_i = 1,\ w_i \geq 0, \tag{1}$$

where all four components lie in $[0, 1]$:

- $Z = (\overline{\mathrm{AUROC}}_0 - 0.5)/0.5$, clipped to $[0, 1]$, is the mean zero-shot AUROC over the T targets rescaled so that chance level equals 0;
- S is the semantic alignment score, defined as the proportion of test samples whose embedding has a higher cosine similarity with the prompt for its true state (normal or abnormal) than with the prompt for the other state. It applies only to embedding-based methods;
- $D = (\mathrm{AUROC}_{\mathrm{seen}} - \mathrm{AUROC}_{\mathrm{unseen}})/(\mathrm{AUROC}_{\mathrm{seen}} - 0.5)$, clipped to $[0, 1]$, is the relative drop from seen to unseen targets. It enters as $1 - D$, so that a smaller drop raises the index;
- $E = (\mathrm{AUROC}_k - \overline{\mathrm{AUROC}}_0)/(1 - \overline{\mathrm{AUROC}}_0)$, clipped to $[0, 1]$, is the share of the remaining headroom recovered with k labeled shots (k must be reported).

The original formulation subtracted D; writing $1 - D$ keeps the index in $[0, 1]$ without changing the ranking it produces. In the absence of empirical evidence for unequal importance, the default weights are equal ($w_i = 1/4$), the usual default for composite indicators [79]. When a component is not reported, its weight is set to zero, the remaining weights are renormalized, and the missing component is stated. Because any choice of weights is a value judgment, results should be accompanied by a sensitivity analysis in which weights are drawn uniformly from the simplex (a Dirichlet$(1, \ldots, 1)$ distribution) and the stability of the ranking is reported. Stakeholder-elicited weights, for example from a Delphi panel, are an alternative for specific deployments.

Worked example. Table 16 applies the index to the three methods compared under the same protocol in [30]. Z uses the zero-shot AUROC on six unseen medical datasets. D uses the drop from the in-domain industrial benchmark to the out-of-domain medical benchmark at $k = 4$. E uses the gain from zero-shot to $k = 4$ on the medical datasets. S was not reported, so the weights were renormalized to $1/3$. MVFA obtains the highest index (0.63), followed by APRIL-GAN (0.50) and WinCLIP (0.40). In 10,000 random weightings MVFA ranked first every time, and the full ranking was unchanged in 87% of draws. Two caveats apply. WinCLIP's few-shot setting uses normal shots only, whereas the other two methods use normal and abnormal shots, so E is not strictly comparable across them. In addition, Z and E are negatively related by construction, because a high zero-shot score leaves less headroom for adaptation.

Table 16. Worked example of the composite generalization index using published results of [30] (AUROC in %; components on $[0, 1]$)

Method	**Mean zero-shot AUROC (6 datasets)**	**In-domain AUROC, k = 4**	**Medical AUROC, k = 4**	Z	S	D	E	**AGI**
WinCLIP	60.09	95.16	76.38	0.202	NR	0.416	0.408	0.398
APRIL-GAN	74.19	92.77	81.65	0.484	NR	0.260	0.289	0.504
MVFA	80.43	96.19	88.97	0.609	NR	0.156	0.436	0.630

The index has limitations. It rewards discrimination and adaptation but ignores calibration, alert burden and clinical utility. Its components depend on how "seen" and "unseen" targets are defined. It has not yet been checked

against expert judgment or patient outcomes. It should therefore be used to structure comparisons and reporting, and it should not be read as a validated measure of clinical readiness.

7. Supplementary material and data availability

The supplementary material contains: S1, the study-level extraction for all 49 studies (also provided as a spreadsheet); S2, the PRISMA 2020 checklist and the full search strings for each database; S3, the PRISMA 2020 flow diagram; S4, the CASP systematic-review checklist applied to this review; S5, the risk-of-bias judgments at eligibility (295 reports) and the full-text re-appraisal of the 40 primary studies with signalling-question answers; S6, the protocol as applied, with deviations; and S7, the thematic and classification codebook, including the decision rules added during coding. The analysis scripts that reproduce every table and figure from S1 and S5 are available from the corresponding author.

REFERENCES

1. M. Kapsecker, M. C. Möller, and S. M. Jonas, "Disentangled representational learning for anomaly detection in single-lead electrocardiogram signals using variational autoencoder," *Computers in Biology and Medicine*, vol. 184, p. 109422, 2025, doi: 10.1016/j.compbiomed.2024.109422.
2. J. Hussain, M. Båth, and J. Ivarsson, "Generative adversarial networks in medical image reconstruction: A systematic literature review," *Computers in Biology and Medicine*, vol. 191, p. 110094, 2025, doi: 10.1016/j.compbiomed.2025.110094.
3. Y. Ramakrishnaiah, N. Macesic, G. I. Webb, A. Y. Peleg, and S. Tyagi, "EHR-ML: A data-driven framework for designing machine learning applications with electronic health records," *International Journal of Medical Informatics*, vol. 196, p. 105816, 2025, doi: 10.1016/j.ijmedinf.2025.105816.
4. A. O. Ezeogu and A. Emmanuel, "Securing big data pipelines in healthcare: A framework for real-time threat detection in population health systems," *Research Corridor Journal of Engineering Science*, vol. 2, no. 1, 2025. https://researchcorridor.org/index.php/RCJES/article/view/531
5. S. Raghavan, "Ontology-driven cross-domain access control framework for anomaly detection in cloud-based healthcare system," *Research Square*, preprint, Version 1, 2025, doi: 10.21203/rs.3.rs-6128795/v1.
6. S. T. Amin, "Anomaly detection in IoT-connected medical devices using machine learning for disease monitoring," *Journal of Information Systems Engineering and Management*, vol. 10, no. 33s, pp. 1068–1079, 2025, doi: 10.52783/jisem.v10i33s.5833.
7. H. Niu, O. A. Omitaomu, M. A. Langston, S. K. Grady, M. Olama, O. Ozmen, H. B. Klasky, A. Laurio, M. Ward, and J. Nebeker, "Anomaly detection in electronic health records across hospital networks: Integrating machine learning with graph algorithms," *IEEE Journal of Biomedical and Health Informatics*, 2025, doi: 10.1109/JBHI.2025.3527752.
8. T. Shehzadi, "Challenges and innovations in anomaly detection for healthcare data," preprint, 2025, doi: 10.13140/RG.2.2.22491.68644.
9. C. Cui, X. Zheng, R. Deng, Q. Liu, T. Yao, K. T. Wilson, L. A. Coburn, B. A. Landman, H. Yang, Y. Wang, and Y. Huo, "Quantitative benchmarking of anomaly detection methods in digital pathology images," *Machine Learning: Health*, vol. 1, no. 1, p. 010502, 2025, doi: 10.1088/3049-477X/ae0f9f.
10. K. J. Basco, A. Singh, D. Nasef, C. Hartnett, M. Ruane, J. Tagliarino, M. Nizich, and M. Toma, "Electrocardiogram abnormality detection using machine learning on summary data and biometric features," *Diagnostics*, vol. 15, no. 7, p. 903, 2025, doi: 10.3390/diagnostics15070903.
11. X. Zhang and Q. Wang, "EEG anomaly detection using temporal graph attention for clinical applications," *Journal of Computational Technology and Software*, vol. 4, no. 7, 2025. https://ashpress.org/index.php/jcts/article/view/200
12. Z. Tang, Z. Tang, and J. Wu, "Anomaly detection in medical via multimodal foundation models," *Frontiers in Bioengineering and Biotechnology*, vol. 13, p. 1644697, 2025, doi: 10.3389/fbioe.2025.1644697.
13. A. K. Rai, D. K. Verma, and R. K. Dwivedi, "Machine learning based approach for anomaly detection in healthcare IoT systems with variational autoencoders for data integrity and security enhancement," *Journal of Information Systems Engineering and Management*, vol. 10, no. 37s, 2025. https://jisem-journal.com/index.php/journal/article/view/6429
14. R. Dey, A. Roy, J. Akter, A. Mishra, and M. Sarkar, "AI-driven machine learning for fraud detection and risk management in U.S. healthcare billing and insurance," *Journal of Computer Science and Technology Studies*, vol. 7, no. 1, 2025, doi: 10.32996/jcsts.2025.7.1.14.
15. N. Sriraam, B. Chinta, S. Seshadri, and S. Suresh, "A comprehensive review of artificial intelligence-based algorithm towards fetal facial anomalies detection (2013–2024)," *Artificial Intelligence Review*, vol. 58, art. no. 152, 2025, doi: 10.1007/s10462-025-11160-7.
16. H. De Meulemeester, F. De Smet, J. van Dorst, E. Derroitte, and B. De Moor, "Explainable unsupervised anomaly detection for healthcare insurance data," *BMC Medical Informatics and Decision Making*, vol. 25, p. 14, 2025, doi: 10.1186/s12911-024-02823-6.
17. A. Ali, T. Montanaro, I. Sergi, S. Carrisi, D. Galli, C. Distante, and L. Patrono, "An innovative IoT and edge intelligence framework for monitoring elderly people using anomaly detection on data from non-wearable sensors," *Sensors*, vol. 25, no. 6, p. 1735, 2025, doi: 10.3390/s25061735.
18. M. Z. Khan, A. Sabur, and H. A. Ghandorh, "A novel Internet of Medical Things hybrid model for cybersecurity anomaly detection," *Sensors*, vol. 25, p. 6501, 2025, doi: 10.3390/s25206501.

19. J. Song, J. Wang, A. T. Nguyen, K. Byeon, S. Ahn, S. H. Lee, and J. T. Kwak, "Normal and abnormal pathology knowledge-augmented vision-language model for anomaly detection in pathology images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025. https://arxiv.org/abs/2508.15256
20. H. Niu, O. A. Omitaomu, M. A. Langston, M. Olama, O. Ozmen, H. B. Klasky, A. Laurio, M. Ward, and J. Nebeker, "EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records," *Journal of Biomedical Informatics*, vol. 150, p. 104605, 2024, doi: 10.1016/j.jbi.2024.104605.
21. G. Lazrek, K. Chetioui, Y. Balboul, S. Mazer, and M. El Bekkali, "An RFE/Ridge-ML/DL based anomaly intrusion detection approach for securing IoMT system," *Results in Engineering*, vol. 23, p. 102659, 2024, doi: 10.1016/j.rineng.2024.102659.
22. M. Kaleem and G. G. Devarajan, "Energy-efficient classification strategy for detecting interference and malicious nodes in wireless body area networks," *Cyber Security Applications*, vol. 2, p. 100048, 2024, doi: 10.1016/j.csa.2024.100048.
23. M. E. Tschuchnig and M. Gadermayr, "Anomaly detection in medical imaging: A mini review," in *Data Science–Analytics and Applications*. Cham, Switzerland: Springer, 2022, pp. 33–38, doi: 10.1007/978-3-658-36295-9_5.
24. Z. Zhang, Z. Sun, Z. Liu, B. Du, R. Yu, Z. Zhao, and Y. Xu, "Spatial-aware attention generative adversarial network for semi-supervised anomaly detection in medical image," arXiv preprint arXiv:2405.12872v1, 2024. https://arxiv.org/abs/2405.12872
25. A. K. Ramachandran Sumangala Devi, "Anomaly detection in healthcare data warehouses using ML," *International Journal of Multidisciplinary Research*, vol. 6, no. 5, 2024. https://www.ijfmr.com/papers/2024/5/28705.pdf
26. S. Jayabharathi and V. Ilango, "Machine learning and deep learning approaches for anomaly detection in medical image analysis: Advancing diagnostic healthcare," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 5415–5421, 2024. https://ijisae.org/index.php/IJISAE/article/view/7391
27. D. Alsalman, "A comparative study of anomaly detection techniques for IoT security using adaptive machine learning for IoT threats," *IEEE Access*, vol. 12, pp. 14720–14730, 2024, doi: 10.1109/ACCESS.2024.3359033.
28. J. Bao, H. Sun, H. Deng, Y. He, Z. Zhang, and X. Li, "BMAD: Benchmarks for medical anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2024, pp. 4042–4053, doi: 10.48550/arXiv.2306.11876.
29. A. Nawaz, S. S. Khan, and A. Ahmad, "Ensemble of autoencoders for anomaly detection in biomedical data: A narrative review," *IEEE Access*, vol. 12, pp. 17274–17289, 2024, doi: 10.1109/ACCESS.2024.3360691.
30. C. Huang, A. Jiang, J. Feng, Y. Zhang, X. Wang, and Y. Wang, "Adapting visual-language models for generalizable anomaly detection in medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 11375–11385, doi: 10.48550/arXiv.2403.12570.
31. M. A. Rassam, "Autoencoder-based neural network model for anomaly detection in wireless body area networks," *IoT*, vol. 5, no. 4, pp. 852–870, 2024, doi: 10.3390/iot5040039.
32. M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud, and F. Khalique, "Anomaly-based threat detection in smart health using machine learning," *BMC Medical Informatics and Decision Making*, vol. 24, p. 347, 2024, doi: 10.1186/s12911-024-02760-4.
33. M. M. Khan and M. Alkhathami, "Anomaly detection in IoT-based healthcare: Machine learning for enhanced security," *Scientific Reports*, vol. 14, p. 5872, 2024, doi: 10.1038/s41598-024-56126-x.
34. S. K. B. Sangeetha, C. Selvarathi, S. K. Mathivanan, J. Cho, and S. V. Easwaramoorthy, "Secure healthcare access control system (SHACS) for anomaly detection and enhanced security in cloud-based healthcare applications," *IEEE Access*, vol. 12, pp. 164544–164559, 2024, doi: 10.1109/ACCESS.2024.3492024.
35. V. V. Raje, S. Goel, S. V. Patil, M. D. Kokate, D. A. Mane, and S. Lavate, "Realtime anomaly detection in healthcare IoT: A machine learning-driven security framework," *Journal of Electrical Systems*, vol. 19, no. 3, 2023, doi: 10.52783/jes.700.
36. M. Cekić, "Anomaly detection in medical time series with generative adversarial networks: A selective review," in *Anomaly Detection—Recent Advances, AI and ML Perspectives and Applications*. London, U.K.: IntechOpen, 2023, doi: 10.5772/intechopen.112582.
37. L. Aversano, M. Bernardi, M. Cimitile, D. Montano, R. Pecori, and L. Veltri, "Anomaly detection of medical IoT traffic using machine learning," in *Proc. 12th Int. Conf. Data Sci., Technol. Appl. (DATA)*, 2023, pp. 173–182, doi: 10.5220/0012132000003541.
38. X. Yang, X. Qi, and X. Zhou, "Deep learning technologies for time series anomaly detection in healthcare: A review," *IEEE Access*, vol. 11, pp. 117789–117799, 2023, doi: 10.1109/ACCESS.2023.3325896.
39. P. U. Eze, N. Geard, I. Mueller, and I. Chadès, "Anomaly detection in endemic disease surveillance data using machine learning techniques," *Healthcare*, vol. 11, no. 13, p. 1896, 2023, doi: 10.3390/healthcare11131896.
40. K. H. Le, T. V. Tran, H. H. Pham, H. T. Nguyen, T. T. Le, and H. Q. Nguyen, "Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis," *IEEE Access*, vol. 11, pp. 14105–14114, 2023, doi: 10.1109/ACCESS.2023.3243845.
41. D. Samariya, J. Ma, S. Aryal, and X. Zhao, "Detection and explanation of anomalies in healthcare data," *Health Information Science and Systems*, vol. 11, no. 1, p. 20, 2023, doi: 10.1007/s13755-023-00221-2.
42. S. Park, K. H. Lee, B. Ko, and N. Kim, "Unsupervised anomaly detection with generative adversarial networks in mammography," *Scientific Reports*, vol. 13, p. 2925, 2023, doi: 10.1038/s41598-023-29521-z.
43. N. Bijlani, R. Nilforooshan, and S. Kouchaki, "An unsupervised data-driven anomaly detection approach for adverse health conditions in people living with dementia: Cohort study," *JMIR Aging*, vol. 5, no. 3, p. e38211, 2022, doi: 10.2196/38211.
44. M. Nawaz and J. Ahmed, "Cloud-based healthcare framework for real-time anomaly detection and classification of 1-D ECG signals," *PLoS ONE*, vol. 17, no. 12, p. e0279305, 2022, doi: 10.1371/journal.pone.0279305.
45. J. S. Sunny, C. P. K. Patro, K. Karnani, S. C. Pingle, F. Lin, M. Anekoji, L. D. Jones, S. Kesari, and S. Ashili, "Anomaly detection framework for wearables data: A perspective review on data concepts, data analysis algorithms and prospects," *Sensors*, vol. 22, no. 3, p. 756, 2022, doi: 10.3390/s22030756.
46. A. Albattah and M. A. Rassam, "A correlation-based anomaly detection model for wireless body area networks using convolutional long short-term memory neural network," *Sensors*, vol. 22, no. 5, p. 1951, 2022, doi: 10.3390/s22051951.
47. M. Briskilla and T. D. Rajkumar, "Adaptive anomaly detection in cardiovascular time series through deep reinforcement and active learning," *Procedia Computer Science*, vol. 252, pp. 43–52, 2025, doi: 10.1016/j.procs.2024.12.005.

48. F. Khanizadeh, A. Ettefaghian, G. Wilson, A. Shirazibeheshti, T. Radwan, and C. Luca, "Smart data-driven medical decisions through collective and individual anomaly detection in healthcare time series," *International Journal of Medical Informatics*, vol. 194, p. 105696, 2025, doi: 10.1016/j.ijmedinf.2024.105696.
49. S. R. Madamanchi, "Real-time anomaly detection with AI-driven syslog monitoring for bioinformatics reliability," *International Journal of Innovative Science, Engineering and Management*, vol. 4, no. 2, pp. 403–408, 2025, doi: 10.69968/ijisem.2025v4i2403-408.
50. L. A. McGuinness and J. P. T. Higgins, "Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments," *Research Synthesis Methods*, pp. 1–7, 2020, doi: 10.1002/jrsm.1411. https://mcguinlu.shinyapps.io/robvis/
51. M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.
52. M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, and PRISMA-S Group, "PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews," *Systematic Reviews*, vol. 10, p. 39, 2021, doi: 10.1186/s13643-020-01542-z.
53. M. Campbell, J. E. McKenzie, A. Sowden, S. V. Katikireddi, S. E. Brennan, S. Ellis, J. Hartmann-Boyce, R. Ryan, S. Shepperd, J. Thomas, V. Welch, and H. Thomson, "Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline," *BMJ*, vol. 368, p. l6890, 2020, doi: 10.1136/bmj.l6890.
54. R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Annals of Internal Medicine*, vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.
55. K. G. M. Moons, J. A. A. Damen, T. Kaul, L. Hooft, C. Andaur Navarro, P. Dhiman, A. L. Beam, B. Van Calster, L. A. Celi, S. Denaxas, A. K. Denniston, M. Ghassemi, G. Heinze, A. P. Kengne, L. Maier-Hein, X. Liu, P. Logullo, M. D. McCradden, N. Liu, L. Oakden-Rayner, K. Singh, D. S. Ting, L. Wynants, B. Yang, J. B. Reitsma, R. D. Riley, G. S. Collins, and M. van Smeden, "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, p. e082505, 2025, doi: 10.1136/bmj-2024-082505.
56. G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, L. A. Celi, S. Denaxas, A. K. Denniston, B. Glocker, R. M. Golub, H. Harvey, G. Heinze, M. M. Hoffman, A. P. Kengne, E. Lam, N. Lee, E. W. Loder, L. Maier-Hein, B. A. Mateen, M. D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, and P. Logullo, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024, doi: 10.1136/bmj-2023-078378.
57. B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, Tech. Rep. EBSE-2007-01, 2007.
58. V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.
59. G. H. Guyatt, A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H. J. Schünemann, "GRADE: An emerging consensus on rating quality of evidence and strength of recommendations," *BMJ*, vol. 336, no. 7650, pp. 924–926, 2008, doi: 10.1136/bmj.39489.470347.AD.
60. B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, X. Liu, B. A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D. S. W. Ting, P. Watkinson, W. Weber, P. Wheatstone, P. McCulloch, and the DECIDE-AI expert group, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nature Medicine*, vol. 28, pp. 924–933, 2022, doi: 10.1038/s41591-022-01772-9.
61. C. L. Andaur Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft, "Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review," *BMJ*, vol. 375, p. n2281, 2021, doi: 10.1136/bmj.n2281.
62. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, pp. 8748–8763, 2021. https://arxiv.org/abs/2103.00020
63. J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023. https://arxiv.org/abs/2303.14814
64. Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024. https://arxiv.org/abs/2310.18961
65. M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024. https://arxiv.org/abs/2304.07193
66. X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023. https://arxiv.org/abs/2303.15343
67. S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon, "A multimodal biomedical foundation model trained from fifteen million image–text pairs," *NEJM AI*, vol. 2, no. 1, p. AIoa2400640, 2025, doi: 10.1056/AIoa2400640.

68. M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang, and F. Mahmood, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, pp. 863–874, 2024, doi: 10.1038/s41591-024-02856-4.
69. G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, art. no. 38, 2021, doi: 10.1145/3439950.
70. L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021, doi: 10.1109/JPROC.2021.3052449.
71. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, art. no. 15, 2009, doi: 10.1145/1541880.1541882.
72. J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, pp. 5635–5662, 2024, doi: 10.1007/s11263-024-02117-4.
73. Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019, doi: 10.1109/TPAMI.2018.2857768.
74. C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014, doi: 10.1109/TPAMI.2013.140.
75. T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection: A survey," *ACM Computing Surveys*, vol. 54, no. 7, art. no. 141, 2021, doi: 10.1145/3464423.
76. T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30–44, 2019, doi: 10.1016/j.media.2019.01.010.
77. C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction," *BMC Bioinformatics*, vol. 22, suppl. 2, p. 31, 2021, doi: 10.1186/s12859-020-03936-1.
78. N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," *IEEE Access*, vol. 9, pp. 118571–118583, 2021, doi: 10.1109/ACCESS.2021.3107163.
79. OECD and European Commission Joint Research Centre, *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris, France: OECD Publishing, 2008.
80. European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, L series, 12 July 2024.
81. U.S. Food and Drug Administration, Health Canada, and Medicines and Healthcare products Regulatory Agency, "Good machine learning practice for medical device development: Guiding principles," October 2021.
82. U.S. Food and Drug Administration, "Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff," December 2024.
83. European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," *Official Journal of the European Union*, L 119, 4 May 2016.
84. U.S. Department of Health and Human Services, "Health Insurance Portability and Accountability Act (HIPAA) Security Rule," 45 CFR Part 160 and Part 164, Subparts A and C.
85. N. Kumaraswamy, M. K. Markey, T. Ekin, J. C. Barner, and K. Rascati, "Healthcare fraud data mining methods: A look back and look ahead," *Perspectives in Health Information Management*, vol. 19, no. 1, p. 1i, 2022.
86. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.
87. World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva, Switzerland: WHO, 2021.
88. H. Mostafa, M. A. A. Makhlouf, H. Refaat, and B. Hafiz, "Integrating LSTM and LOF: A comprehensive approach to anomaly detection in healthcare data," *Statistics, Optimization & Information Computing*, vol. 13, pp. 1245–1265, 2025, doi: 10.19139/soic-2310-5070-2206.
89. W. Alayash, M. Rahrouh, A. A. Ibrahim, M. H. Mohamed, S. T. Ahmed, M. H. Albarri, and M. H. Ahmed, "Assessing LSTM and GRU for multi-dataset intrusion detection in IoT environments," *Statistics, Optimization & Information Computing*, vol. 15, pp. 3155–3173, 2026, doi: 10.19139/soic-2310-5070-3226.
90. D. A. Ali and H. T. Sadeeq, "An interpretable deep learning framework for multi-class dental disease classification from intraoral RGB images," *Statistics, Optimization & Information Computing*, vol. 14, no. 6, pp. 3380–3397, 2025, doi: 10.19139/soic-2310-5070-2880.
91. R. Abdulkareem and A. M. Abdulazeez, "A comparative study of multi-class classification based on imbalanced data," *Statistics, Optimization & Information Computing*, vol. 15, no. 5, pp. 4337–4357, 2026, doi: 10.19139/soic-2310-5070-2731.
92. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017. https://arxiv.org/abs/1610.02136
93. C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, 2021, doi: 10.1109/TPAMI.2020.2981604.
94. J. Zhu and G. Pang, "Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024. https://arxiv.org/abs/2403.06495