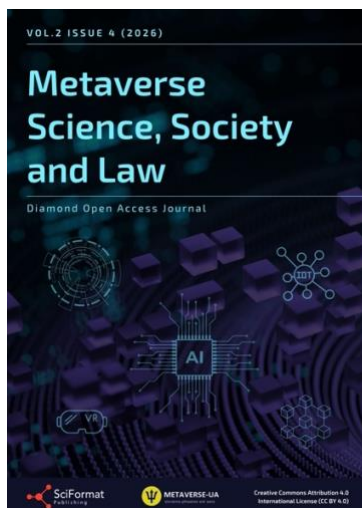




Metaverse Science, Society and Law

Vol. 2, Issue 4 (2026)



Publisher:
SciFormat Publishing Inc.

ISNI: 0000 0005 1449 8214
2734 17 Avenue Southwest, Calgary,
Alberta, Canada, T3E0A7

+15878858911
✉ editorial-office@sciformat.ca

ARTICLE TITLE	A RISK-BASED QUALITY ASSURANCE FRAMEWORK FOR FHIR-BASED CLINICAL APPLICATIONS: ALIGNING VERIFICATION EFFORT WITH CLINICAL CONSEQUENCE
---------------	---

DOI	https://doi.org/10.69635/mssl.2026.2.4.65
-----	---

RECEIVED	25 June 2026
----------	--------------

ACCEPTED	30 September 2026
----------	-------------------

PUBLISHED	09 October 2026
-----------	-----------------

LICENSE	
---------	--



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

© The author(s) 2026.

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

A RISK-BASED QUALITY ASSURANCE FRAMEWORK FOR FHIR-BASED CLINICAL APPLICATIONS: ALIGNING VERIFICATION EFFORT WITH CLINICAL CONSEQUENCE

Pavlo Palamarchuk

Palamarchuk Technologies LLC, Sacramento, CA, USA

ORCID ID: 0009-0001-3089-4813

ABSTRACT

Background: Healthcare software built on the Fast Healthcare Interoperability Resources (FHIR) standard now mediates prescribing, care coordination, and patient access at national scale, yet the quality assurance practices applied to such systems remain largely undifferentiated: a defect in a practitioner-directory lookup receives the same testing attention as a defect in a medication-dispensing workflow.

Objective: This paper proposes a risk-based quality assurance framework for FHIR-based clinical applications that allocates verification effort in proportion to clinical consequence.

Methods: The framework was constructed through a design-science methodology from three inputs: a case series of production failure modes documented on a HIPAA-regulated telehealth platform; quantified harm evidence from federal patient-safety research, including the finding that computerized provider order entry system-related errors occur in 0.3% to 6.3% of all prescriptions; and risk-management logic adapted from ISO 14971 and healthcare failure mode and effect analysis. The framework's behavioral layer was then exercised against four FHIR server implementations using an open-source probe as one example realization.

Results: The resulting model classifies FHIR resources and workflows into four risk tiers along three dimensions — clinical consequence class, data-integrity criticality, and failure detectability — and maps each tier to specific verification obligations spanning resource validation, workflow testing, and operational behavior probing. An assistive artificial-intelligence layer for preliminary risk classification is specified with a mandatory human validation gate. Illustrative application demonstrated that the highest-risk behavior identified by the framework — silent data-completeness failure in paginated synchronization — was present on one of the four tested servers while remaining invisible to profile validation and certification testing; the same exercise surfaced a previously undocumented rate-limiting defect, which was reported upstream.

Conclusions: Verification effort for FHIR-based clinical systems can be allocated by clinical consequence rather than by test-suite convention, extending quality obligations into the operational-behavior layer where specification-conformant systems have been shown to fail. Empirical calibration across sites and defect corpora is the necessary next step, and the instrumentation for it is openly available.

KEYWORDS

FHIR, Risk-Based Testing, Clinical Software Quality Assurance, Patient Safety, Health Information Technology, Interoperability, Verification

CITATION

Pavlo Palamarchuk. (2026) A Risk-Based Quality Assurance Framework for FHIR-Based Clinical Applications: Aligning Verification Effort with Clinical Consequence. *Metaverse Science, Society and Law*. Vol. 2, Issue 4. doi: 10.69635/mssl.2026.2.4.65

COPYRIGHT

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

1. Introduction

The interoperability of health information in the United States is no longer a voluntary engineering aspiration. The 21st Century Cures Act and its implementing regulations require certified health information technology to expose standardized application programming interfaces based on Health Level Seven (HL7) FHIR Release 4 (45 C.F.R. § 170.315(g)(10)), and parallel rules extend FHIR obligations to payers and prior-authorization workflows. A measurable consequence of this policy program is that clinical operations — ordering, dispensing, scheduling, triage, document exchange — increasingly execute across FHIR interfaces operated by hundreds of vendors and thousands of deployments (Ayaz et al., 2021; Vorisek et al., 2022), with a substitutable application ecosystem layered on the SMART on FHIR platform model (Mandel et al., 2016).

Software defects in these systems are not an abstract quality concern; they are a documented cause of patient harm. Adverse drug events account for approximately 39% of all adverse events in the inpatient setting, and system-related errors attributable to computerized provider order entry (CPOE) itself — not to its users — have been observed in 0.3% to 6.3% of all prescriptions, representing between 6.1% and 77.7% of prescribing errors across the studies reviewed (Agency for Healthcare Research and Quality [AHRQ], 2024). Electronic health records were identified as potential contributors to 61.3% of diagnostic errors in ambulatory care, concentrated in the testing stage of the diagnostic process (AHRQ, 2025a). Diagnostic error more broadly accounts for 6% to 17% of hospital adverse events (AHRQ, 2020). The economics point in the same direction from the opposite side: correctly functioning CPOE with clinical decision support was projected to save \$133 billion and 201,000 quality-adjusted life years over a five-year horizon, and its deployment was associated with a greater than 50% decline in preventable adverse drug events (AHRQ, 2025a), a figure consistent with the foundational trial evidence of Bates et al. (1998), in which serious medication error rates fell 55% following CPOE introduction.

Against this harm profile, the quality assurance practices applied to FHIR-based applications display a structural mismatch. The testing ecosystem that has matured around FHIR concentrates on conformance: resource validators verify instances against profiles, terminology bindings, and invariants; certification harnesses verify that a server implements the capabilities regulation requires. Walonoski et al. (2018b) demonstrated that use of such conformance tooling correlates with specification compliance across the Crucible and Touchstone platforms. What conformance tooling does not do — and, on the evidence of the specification's own validation documentation, does not attempt to do — is differentiate testing effort by the clinical consequence of failure. The FHIR validation page enumerates eight aspects of validation, every one of which is a property of a resource instance; it concedes that these methods “can only validate the computable aspects of conformance” (HL7 International, 2024). A Patient demographic record and a MedicationRequest carrying a controlled-substance order pass through the same validation machinery with the same priority.

Two further features of the current environment sharpen the mismatch. The first is the modality shift toward telehealth, which grew into a standing channel of American care delivery faster than its safety evidence base. A federal rapid review of video-based telehealth screened 7,155 citations, found only 23 eligible studies, and reported no studies at all evaluating patient-safety practices for real-time telehealth encounters (AHRQ, 2023). Where mature inpatient settings can lean on decades of incident-reporting infrastructure, virtual-care platforms operate in an evidentiary vacuum; the quality assurance function inside those platforms is, for practical purposes, the safety apparatus. The second feature is the threat environment. The 2024 ransomware attack on Change Healthcare produced estimated losses of \$874 million alongside widespread disruption of provider operations, and hospitals had at that point adopted an average of 70.7% of the applicable National Institute of Standards and Technology cybersecurity framework functions (U.S. Government Accountability Office [GAO], 2024b). Verification of authentication behavior, authorization boundaries, and data-handling contracts is therefore not a separable security concern; it belongs inside the same risk-allocated quality program as clinical correctness, because the failure pathways converge on the same records.

It is worth being precise about what the existing conformance apparatus achieves, since the argument here depends on respecting rather than dismissing it. Profile validation catches structural and terminological deviation with high reliability, and its wide availability across the Firely and HAPI ecosystems has plausibly raised the floor of FHIR data quality; the correlation reported by Walonoski et al. (2018b) between test-platform usage and specification compliance supports that reading. Certification testing, institutionalized through the Inferno test kits for § 170.315(g)(10), forces a shared capability baseline across a market that would otherwise fragment. The deficiency is not in what these instruments test but in what their pass signal is taken to mean. A certification pass is routinely treated, in procurement and in engineering culture alike, as a quality assurance endpoint; the empirical record — capability-statement divergence (Boufahja & Verma, 2025), eight-fold operational variance among

certified endpoints (Jones et al., 2024), production servers whose supported queries had to be discovered by runtime probing (Hauptfeld et al., 2026) — shows it is a starting condition. Risk-based allocation is the discipline that converts that recognition into an actionable testing economy.

The software-testing discipline resolved the analogous problem two decades ago under the name risk-based testing: the allocation of test design, execution, and maintenance effort in proportion to the probability and impact of failure (Felderer & Schieferdecker, 2014). International standards embed the same logic. ISO 14971 requires medical-device manufacturers to estimate and control risk as a function of harm severity and probability; IEC 62304 assigns software safety classes (A through C) that scale lifecycle rigor with injury potential; and healthcare failure mode and effect analysis (HFMEA) operationalizes prospective risk scoring for clinical processes (DeRosier et al., 2002). Yet a review of the FHIR literature locates no published framework that translates these risk disciplines into the specific vocabulary of FHIR resources, workflows, and server behaviors. The two systematic reviews of FHIR research each observe that testing and validation are under-studied relative to the standard's adoption (Ayaz et al., 2021; Vorisek et al., 2022); neither identifies a risk-stratified approach. Empirical work meanwhile keeps demonstrating that FHIR implementations diverge from their declared behavior in ways that matter operationally: Boufahja and Verma (2025) generated exhaustive search-parameter tests and found discrepancies between servers' capability statements and observed conduct, and Jones et al. (2024) measured an eight-fold cross-vendor variation in bulk-export throughput against production endpoints of certified systems.

This paper addresses that gap with three contributions. It proposes a four-tier risk classification for FHIR resources and workflows constructed along three dimensions: clinical consequence class, data-integrity criticality, and failure detectability. It maps each tier to concrete verification obligations, extending beyond resource validation to workflow testing and operational behavior probing — the layer where, as the motivating case series shows, spec-conformant systems still fail. And it specifies a bounded role for artificial intelligence as an assistive classifier within the framework, with a mandatory human validation gate, reflecting both the promise and the documented hazards of automation in clinical contexts, where alert override rates between 46.2% and 96.2% illustrate how poorly calibrated automated signals degrade rather than improve safety behavior (AHRQ, 2024).

2. Methods

The framework was developed through a design-science approach (Hevner et al., 2004; Peffers et al., 2007) in three stages: problem evidence assembly, model construction, and illustrative application. Design science was selected over purely empirical methods because the research objective is an engineered artifact — a classification and mapping model — whose initial validity rests on grounding in documented evidence rather than on a single-site trial.

2.1 Problem evidence assembly

Three evidence streams were assembled. The first is a case series of four production failure modes documented between 2025 and 2026 in the operation of a HIPAA-regulated telehealth platform built on an open-source, FHIR-native server platform, where the author led quality automation. The four incidents — a paginated patient-synchronization job silently stranding records behind a server-side search-offset cap; exhaustion of a shared API rate-limit budget by background jobs, causing clinician-facing failures; an environment-dependent divergence in Binary resource serialization triggered by an Accept-header subtlety; and a platform-wide authentication outage precipitated by enforcement of token-audience validation consistent with JSON Web Token best practices (Sheffer et al., 2020) — are reported here in generalized, confidentiality-preserving form. Each was reproduced after the fact on a clean local FHIR deployment using only synthetic data (Walonoski et al., 2018a), converting proprietary incidents into portable evidence.

The second stream is published harm and performance evidence, drawn from federal patient-safety syntheses (AHRQ, 2020, 2023, 2024, 2025a, 2025b), federal oversight reports on health-system cybersecurity and interoperability (GAO, 2015, 2024a, 2024b), and peer-reviewed FHIR studies identified through a structured search of PubMed, the Journal of the American Medical Informatics Association, JMIR journals, and the ACM and IEEE digital libraries, restricted where possible to the period 2016–2026. Search terms combined “FHIR” and “HL7” with “testing,” “validation,” “conformance,” “quality assurance,” “risk-based testing,” “verification,” and “patient safety”; no result described a risk-stratified verification framework mapping FHIR resources and workflows to clinical consequence, which is the gap this work addresses. The third stream comprises normative risk frameworks: ISO 14971, IEC 62304, ISO/IEC/IEEE 29119, HFMEA (DeRosier et al., 2002), and the sociotechnical safety model of Sittig and Singh (2010), whose eight-dimension analysis was used to check that the classification dimensions do not collapse technology risk into purely technical variables.

2.2 Model construction

Risk dimensions were derived by thematic alignment between the case series and the harm literature: each documented incident was decomposed into the clinical function affected, the data pathway compromised, and the mechanism by which the failure escaped detection. Candidate dimensions were retained only where both an incident and independent published evidence supported their discriminating power. Tier boundaries were set by consequence class rather than by numeric scoring, following the qualitative severity logic of ISO 14971 and the software safety classes of IEC 62304, because numeric probability estimates for novel software failure modes are rarely defensible (Felderer & Schieferdecker, 2014).

A threat to validity deserves acknowledgment at the methods stage rather than the limitations stage, because it shaped a construction decision. Design-science artifacts built by a practitioner from that practitioner's own incidents risk circularity: the framework could simply re-describe the cases that motivated it. Two controls were applied. Each candidate dimension was required to find independent support in evidence unconnected to the case series — the detectability dimension, for instance, stands on the HFMEA tradition and on the published record of undetected capability divergence, not on the synchronization incident alone. And the model was checked against disconfirming territory: workflows the case series never touched, such as bulk population export, were classified using the dimensions and the resulting assignments compared against the independent operational evidence of Jones et al. (2024); the model's placement of boundary-crossing export under elevated data-integrity scrutiny is consistent with the cross-vendor completeness variation that study measured, which it could not have been fitted to.

2.3 Illustrative application

The behavioral component of the framework is implementation-agnostic: any scripted probe capable of a bounded HTTP interaction can realize it. To demonstrate practicality, it was exercised against four FHIR server implementations — two public sandboxes (HAPI FHIR; Firely Server), one hosted sandbox (SMART Health IT), and one fresh local deployment (Medplum v5.1.37) — using one example implementation, *fhir-verify*, an open-source harness written by the author that encodes the six behavior classes and operates on synthetic data exclusively. The framework does not depend on this particular tool, and equivalent probes are reproducible from the behavior definitions in Section 3.3 alone. This application is illustrative, not comparative benchmarking: sandbox configurations may differ from production defaults, and single-run observations carry no statistical weight. Its purpose is to demonstrate that the framework's highest-tier behavioral risks are detectable in practice by portable, low-cost probes. That the exercise was not merely illustrative is evidenced by its yield: run against the local Medplum deployment, the rate-limit probe surfaced a previously undocumented behavior — HTTP 429 responses omitting the standard *Retry-After* header, causing off-the-shelf clients to disregard the server's computed wait interval — which the author reported upstream with root-cause analysis and a proposed fix (Medplum issue #10518).

3. Results

3.1 Risk dimensions

Three dimensions emerged as jointly necessary and individually insufficient for classifying FHIR quality risk.

Clinical consequence class (C). The direct question: if this resource or workflow is wrong, what reaches the patient? The harm literature supports a four-level ordering. At the top sit medication and ordering pathways: MedicationRequest, MedicationDispense, and the CPOE workflows around them, where software-originated errors are documented at up to 6.3% of prescriptions (AHRQ, 2024) and where the foundational interruption evidence shows the largest preventable-harm deltas (Bates et al., 1998). One level down sit diagnostic-data pathways — Observation, DiagnosticReport, ServiceRequest — implicated in the 61.3% figure for EHR contribution to ambulatory diagnostic error (AHRQ, 2025a) and in the 6%–17% share of hospital adverse events attributable to diagnostic failure (AHRQ, 2020). Third, coordination and routing pathways: Encounter, Task, Appointment, CareTeam, the workflows that allocate scarce clinical labor. Their failure consequence is indirect but nationally material: the projected 8% registered-nurse shortage in 2028 rising to 24% in nonmetropolitan areas (Health Resources and Services Administration [HRSA], 2025) prices every misrouting defect in clinician-hours the system does not have, and computerized handoff interventions have shown a 42% reduction in handoff-related adverse events when coordination tooling works as intended (AHRQ, 2025b).

Fourth, administrative and descriptive data — Practitioner directories, Organization records — where failure degrades operations without a proximate patient pathway.

Data-integrity criticality (D). Orthogonal to consequence is the question of whether a failure corrupts, loses, or duplicates clinical data versus merely interrupting service. An outage is loud; a completeness failure is quiet and compounding. The motivating incident is exact here: a synchronization job following server-issued pagination links failed at an offset cap, and every record beyond the cap was silently excluded from downstream dispensing — a defect whose severity derived not from the error (which was visible) but from the unsignaled incompleteness behind it. Published evidence echoes the pattern at national scale: HHS operated 99 pandemic-era data systems without a comprehensive inventory, and 15 of the 68 systems holding personally identifiable information lacked required privacy assessments — data-governance failures that no availability metric would surface (GAO, 2024a).

Failure detectability (F). The third dimension asks how a failure announces itself, adapting the detectability factor of HFMEA (DeRosier et al., 2002). Three levels suffice in practice: self-announcing failures (hard errors, outages); operator-detectable failures (wrong values a trained user can notice); and silent failures (data loss, misrouting, or authorization drift observable only through deliberate verification). The case series suggests, and the framework asserts, that detectability should function as a risk multiplier: a moderate-consequence failure that is silent may warrant more verification investment than a high-consequence failure that is self-announcing, because the latter triggers organizational response and the former does not. The specification-conformance literature offers a supporting observation: servers' declared capabilities and actual behavior diverge in ways their operators had not detected until systematically probed (Boufahja & Verma, 2025; Hauptfeld et al., 2026).

3.2 The four-tier model

Combining the dimensions yields four tiers, summarized in Table 1. Figure 1 shows the framework's overall structure: the evidence inputs feeding classification, the three-dimensional tiering, the tier-scaled verification obligations those tiers drive, and the closed evidence loop that returns production incidents to the classification layer.

Table 1. Risk Tiers for FHIR Resources and Workflows With Documented Consequence Evidence

Tier	Definition	Representative FHIR resources / workflows	Documented consequence evidence	Primary source
1 — Critical	Direct medication, ordering, or treatment pathway; silent failure plausible	MedicationRequest, MedicationDispense; prescription transmission; e-prescribing sync	System-related errors in 0.3%–6.3% of all prescriptions; 39% of inpatient adverse events are ADEs	AHRQ (2024)
2 — High	Diagnostic data or its movement; results delivery; clinical documents	Observation, DiagnosticReport, Binary (clinical documents), ServiceRequest	EHRs potential contributors to 61.3% of ambulatory diagnostic errors; diagnostic error in 6%–17% of hospital adverse events	AHRQ (2020, 2025a)
3 — Elevated	Coordination, routing, scheduling of care and clinicians	Encounter, Task, Appointment, CareTeam; practitioner-routing; handoff tooling	42% reduction in handoff-related adverse events when coordination tooling verified and adopted; RN shortage 8% (2028), 24% nonmetro	AHRQ (2025b); HRSA (2025)
4 — Standard	Administrative and descriptive data without proximate clinical pathway	Practitioner, Organization, Location directories	Operational degradation; no direct documented patient-harm pathway	—

Note. ADE = adverse drug event; RN = registered nurse; EHR = electronic health record. Tier assignment is raised one level when the dominant failure mode is silent (detectability level 3) and the workflow crosses a system boundary (synchronization, messaging, or export). A resource's tier is contextual: Binary is Tier 2 when carrying diagnostic documents and Tier 4 when carrying administrative attachments.

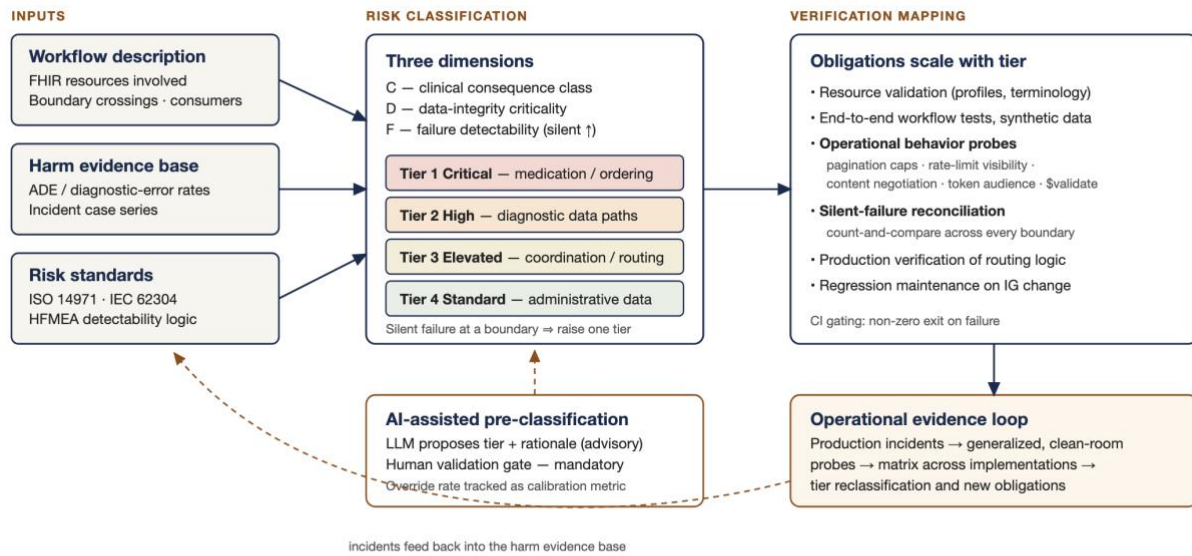


Figure 1. Conceptual structure of the risk-based quality assurance framework for FHIR-based clinical applications: evidence-grounded classification along three dimensions into four tiers; tier-scaled verification obligations including operational behavior probing; and a closed evidence loop returning production incidents to the classification layer. AI assistance is advisory and human-gated.

Fig. 1. Conceptual structure of the risk-based quality assurance framework for FHIR-based clinical applications: evidence-grounded classification along three dimensions into four tiers; tier-scaled verification obligations including operational behavior probing; and a closed evidence loop returning production incidents to the classification layer. AI assistance is advisory and human-gated.

The tier boundaries reward closer inspection, because each encodes a specific evidentiary judgment. Tier 1 is defined by the presence of a pharmacological pathway rather than by resource type: a MedicationRequest displayed read-only in a history view participates in a lower-consequence workflow than the same resource inside an active prescribing loop, and the framework classifies the loop, not the record. The quantitative case for the tier's primacy is unusually direct — medication pathways combine the highest documented harm share (39% of inpatient adverse events) with the best-documented software contribution (0.3%–6.3% of prescriptions affected by system-related error) and the strongest interventional evidence that engineering quality changes outcomes (Bates et al., 1998; AHRQ, 2024). Tier 2 rests on attribution evidence of a different character: the 61.3% figure for EHR contribution to ambulatory diagnostic error is a contributing-factor estimate, not a causal share, and the framework treats it accordingly — as justification for elevated verification of the data pathways feeding diagnosis, particularly the testing stage where 91.8% of the implicated cases concentrated (AHRQ, 2025a). Tier 3's evidence is interventional in the positive direction: coordination tooling, when verified and adopted, produced a 42% reduction in handoff-related adverse events and a 30% reduction in preventable readmissions in the re-engineered discharge program (AHRQ, 2025b); the tier's risk derives from the dependency those gains create. Where coordination software allocates clinicians, its silent failure converts a workforce shortage from a staffing problem into a patient-access incident, and the shortage itself is federally projected rather than hypothesized (HRSA, 2025). Tier 4 is deliberately unglamorous, and its function in the model is economic: an explicit low tier is what makes the high tiers affordable, by licensing teams to spend less where consequence evidence is absent.

Two properties of the model deserve emphasis. Tier membership attaches to workflows first and resources second; the same resource type participates in workflows of different tiers, and classifying by resource alone reproduces the undifferentiated testing the framework exists to correct. And boundary-crossing amplifies risk: every incident in the motivating case series occurred at an integration boundary — server to synchronization job, client to authentication service, environment to environment — which is consistent with the interoperability literature's persistent finding that exchange, not storage, is where health information systems underperform (GAO, 2015; Vorisek et al., 2022).

3.3 Verification mapping

The framework's operative content is the mapping from tier to verification obligation, given in Table 2. The mapping spans three verification layers. Resource validation — profile, terminology, and invariant checking — is necessary at every tier and is well served by existing validators. Workflow verification exercises end-to-end clinical paths with synthetic data. Operational behavior probing tests the server-boundary behaviors that neither of the first two layers reaches: pagination integrity under caps, rate-limit transparency, content-negotiation contracts, token-audience handling, and the presence and honesty of the server's own validation gate.

Table 2. Verification Obligations by Risk Tier

Verification obligation	Tier 1	Tier 2	Tier 3	Tier 4
Profile/terminology validation of resources	Every build	Every build	Every build	Every release
End-to-end workflow tests (synthetic data)	Every build; production-parity environment	Every build	Every release	Risk-triggered
Operational behavior probes (pagination, limits, negotiation, auth)	Every release and on any platform upgrade	Every release	Quarterly and on upgrade	Annual
Silent-failure reconciliation (record counts across boundaries)	Continuous, automated	Daily	Weekly	—
Production verification of routing/assignment logic	Before reliance; after every change	After relevant change	After relevant change	—
Regression suite maintenance trigger	Any IG/profile version change	Any IG/profile version change	Scheduled review	Scheduled review

Note. IG = implementation guide. “Every build” denotes inclusion in continuous-integration gating with non-zero exit on failure. Silent-failure reconciliation compares record counts and identifiers across integration boundaries to detect completeness failures that produce no errors.

The distinctive rows are the third and fourth. Operational behavior probing is absent from published FHIR quality practice, yet it is precisely where the case-series incidents lived; all four were invisible to resource validation by construction, because every payload involved was profile-conformant. Silent-failure reconciliation operationalizes the detectability dimension: for Tier 1 pathways the framework requires an independent count-and-compare mechanism across every boundary a record crosses, on the argument that the cost of reconciliation is trivial against the documented cost of quiet data loss in dispensing pathways.

3.4 The assistive AI layer

Classification effort is the practical obstacle to adoption: a mature clinical platform exposes hundreds of workflow variants. The framework therefore specifies an assistive role for large language models in preliminary tier assignment. The classifier receives a workflow description — participating resources, boundary crossings, downstream consumers — and returns a proposed tier with a stated rationale referencing the three dimensions. Two constraints are mandatory. Proposals are advisory only; a qualified reviewer confirms or overrides every assignment, and the override rate is itself tracked as a calibration metric. And the classifier is never applied to the verification results themselves; it assigns where to look, not whether what was found is acceptable. The design reflects the alert-fatigue evidence: automated signals that are wrong often enough are not merely useless but corrosive, training operators to ignore the channel — override rates for medication alerts reach 96.2% in the published range (AHRQ, 2024). An AI layer that mis-tiers workflows would reproduce the same failure one level up. The framework's position is deliberately conservative relative to the enthusiasm in current literature for generative testing: adaptive REST-testing techniques have shown strong fault-finding performance even against an evolving healthcare system (Sartaj et al., 2025), exploratory studies report meaningful productivity gains from LLM-assisted test authoring (Augusto et al., 2025), and machine-learning prioritization frameworks improve regression ordering on conventional codebases (Khan et

al., 2024), but tier assignment in a clinical context is a judgment about harm, and harm judgments remain human obligations under both ISO 14971 and the FDA’s decision-support guidance (FDA, 2026). Establishing strict human-gated oversight for AI-assisted risk classification is consistent with contemporary scholarship advocating for structured AI law models and ethical legislative frameworks tailored to high-stakes decision-making domains (Kostenko, 2025).

3.5 Illustrative application

Applying the behavioral layer of the framework to four server implementations produced the pattern summarized in Table 3 and reported in full in the example implementation’s repository. On CRUD roundtrip integrity, all four servers passed. On the framework’s Tier 1 amplifier — silent data-completeness hazard in paginated synchronization — one server enforced a hard offset cap (HTTP 400 beyond 10,000 records) while issuing offset-based continuation links, the exact composition that produced the motivating incident; the other three issued non-offset continuation links and exhibited no cap at the probed offset. On rate-limit transparency, three of the four servers exposed no budget information on normal responses, leaving clients to discover throttling by failing. On the Binary content-negotiation contract, the same probe — one synthetic Binary read back under three Accept headers (application/fhir+json, application/json, and */*) — produced divergent results on three of the four servers: only the full FHIR media type (application/fhir+json) returned the FHIR resource identically everywhere, while application/json yielded raw bytes on some servers and the FHIR resource on others. Server-side validation gates, where present, correctly accepted a well-formed synthetic Patient and flagged a malformed one on all servers tested for that behavior. The sixth behavior class, token-audience handling, requires OAuth client credentials that the public sandboxes and the local deployment were not provisioned with for this illustrative run; it is therefore recorded as not exercised rather than passed or failed, and remains part of the framework’s Tier 1 operational-probe set (Table 2).

Table 3. Operational Behavior Probe Results Across Four FHIR Server Implementations (single run, 2026-09-11; synthetic data only)

Behavior probe	HAPI FHIR	Firely Server	SMART Health IT	Medplum v5.1.37 (local)
CRUD roundtrip integrity	Pass	Pass	Pass	Pass
Paginated-sync completeness (offset cap + link style)	No cap at probed offset; non-offset links	No cap; non-offset links	No cap; non-offset links	Hard cap (HTTP 400) with offset-based links
Rate-limit budget visible on normal responses	No	No	No	Yes
Binary content negotiation (application/json vs application/fhir+json)	Divergent ¹	Consistent	Divergent	Divergent
Server \$validate gate (accepts valid, flags invalid)	Pass	Pass	Pass	Pass
Token-audience handling	Not exercised ²	Not exercised ²	Not exercised ²	Not exercised ²

Note. Results are a single dated observation, not a benchmark; public-sandbox configurations may differ from production defaults and change between runs (see Limitations). The Medplum column is a fresh, synthetic-data local installation used only to demonstrate probe feasibility, not the production platform of the motivating case series. ¹In preliminary runs, HAPI Binary support was intermittently restricted (HTTP 412); where Binary creation was permitted, the application/json response shape diverged from application/fhir+json. ²Requires OAuth client credentials not provisioned for this illustrative run.

The probes merit brief description, less as features of the example tool than as evidence of the framework’s cost profile: any team could reimplement them in an afternoon. Each probe is a single, deliberately bounded interaction: the offset-cap probe issues one search request at an offset beyond the

documented cap and classifies the response; the rate-limit probe inspects headers on an ordinary response without ever attempting to induce throttling; the content-negotiation probe uploads one synthetic Binary and reads it back under three Accept headers; the validation-gate probe submits one well-formed and one deliberately malformed synthetic Patient to the server's \$validate operation, where present, and checks that the second is flagged. All created resources carry a dedicated synthetic identifier system and are deleted at run end. A full six-pack run completes in under a minute per server, emits console, JSON, and archival reports, and returns a non-zero exit code on failure, which is what allows the framework's "every build" and "every release" obligations in Table 2 to be enforced mechanically in continuous integration rather than procedurally in checklists.

The observation of practical significance is not any single cell but the distribution: every server was specification-conformant in the behaviors observed, and every server presented a different operational risk surface. A quality strategy calibrated to conformance alone would rate the four deployments equivalent; the framework rates them materially different for a Tier 1 synchronization workload, and the difference was measurable with probes costing under one minute per server.

4. Discussion

4.1 Relation to existing practice

The framework does not compete with conformance testing; it sequences and prioritizes it. Certification harnesses answer whether a server implements required capabilities, and the correlation between conformance-tool use and specification compliance is established (Walonoski et al., 2018b). What certification does not carry is a theory of consequence. The bulk-export study of Jones et al. (2024) is instructive: all five production endpoints studied were certified systems, yet export throughput varied from roughly 1,500 to 12,000 resources per minute — an operational spread with direct consequences for any population-health workload, invisible to certification status. The present framework generalizes that lesson from export performance to the wider class of operational behaviors and attaches it to clinical consequence, which neither the performance literature nor the conformance literature does.

The framework's nearest methodological relatives are outside FHIR. HFMEA supplies the severity-and-detectability logic (DeRosier et al., 2002); IEC 62304's safety classes supply the principle that process rigor scales with injury potential; risk-based testing supplies the allocation economics (Felderer & Schieferdecker, 2014). The contribution here is the translation into FHIR's specific ontology — resources, workflows, server behaviors — with tier assignments grounded in the quantified harm literature rather than in analyst intuition. That grounding is the framework's main defense against the standard criticism of qualitative risk models, that tier boundaries are arbitrary: a reviewer can trace the Tier 1 boundary directly to the prescribing-error and ADE evidence base.

4.2 The silent-failure argument

The framework's most consequential departure from conventional practice is the treatment of detectability as a first-class dimension. Health-IT safety scholarship has long argued that the field lacks "the ability of health IT systems to identify a failure and prevent it from affecting the patient" (Singh & Sittig, 2016, p. 229). The case series sharpens the point: the most dangerous defect among the four was also the least dramatic — no data was corrupted, no user saw an error attributable to the true cause, and the records that failed to synchronize did so in silence. Practitioner literature corroborates the pattern independently: a 2026 architectural guide for FHIR implementers warns that in synchronized (hybrid) deployments "consumers can be overconfident in the quality of the data if they are unaware of the issues with the syncing process" and that "wider data quality issues are not always noticeable immediately" (Devitt, 2026) — an experience-based statement of exactly the boundary-drift hazard the framework's reconciliation obligation addresses. Telehealth amplifies the concern, because the safety-practice evidence base for the modality is nearly empty: a federal rapid review screening 7,155 citations found 23 eligible studies and none evaluating patient-safety practices for real-time telehealth (AHRQ, 2023). In an evidence vacuum, engineered detectability — reconciliation, behavioral probing, count-and-compare — is the available substitute for the incident-reporting infrastructure mature care settings rely on.

4.3 A sociotechnical reading

The framework's dimensions map naturally onto the sociotechnical model of Sittig and Singh (2010), and the mapping guards against a narrow reading of what is being classified. Clinical consequence is a property of the clinical content and workflow dimensions, not of software alone; detectability is jointly produced by the technology and by the people and measurement infrastructure around it; and the verification obligations of Table 2 are organizational rules, which the model treats as a first-class system component. The practical import is that tier assignment sessions should include clinical operations staff, not engineers alone: the clinical significance of a silent routing or assignment failure where coordination software allocates clinicians is typically evident to operations staff before it surfaces in any engineering artifact, which is precisely why detectability must be assessed jointly rather than left to the technical layer. Regulatory currents point the same way.

The FDA's clinical decision support guidance asks developers to disclose algorithm development and validation practices and warns explicitly against automation bias, the human tendency to over-rely on automated output (FDA, 2026). A risk-tiered verification regime, documented and auditable, is the engineering-side complement to that disclosure posture, and early adopters may find that it shortens rather than lengthens their path through oversight conversations. Aligning FHIR verification with clinical consequence reflects broader interdisciplinary frameworks designed to maintain health security, data integrity, and operational resilience across complex digital ecosystems (Kostenko & Drobakha, 2024).

4.4 Limitations

Five limitations bound the claims. The case series originates from a single organization and platform; the failure modes generalize by mechanism, but their frequencies do not. The tier model is qualitative; it inherits the defensibility of its evidence base but has not been calibrated against a defect corpus, and inter-rater reliability of tier assignment is untested. The illustrative application covers four implementations, two of them public sandboxes whose configurations may not represent production defaults; the results demonstrate feasibility, not prevalence. The AI-assist layer is specified but not yet evaluated; its override-rate calibration metric is a design commitment, not a result. And the framework addresses verification, not the organizational economics of acting on verification findings — a tier model changes where teams look, and only indirectly what they fix.

4.5 Implications and research agenda

Industry measurement has recently confirmed how empty the post-launch verification space is. A 2026 market study by an ONC-authorized certification body found that 51% of organizations never test after release, that 67% of connectathon participants judge that preparation inadequate for production, and that time, partner access, and cost — cited by 50%, 67%, and 43% of respondents respectively — are the operative barriers (Drummond Group, 2026), as summarized in Figure 2. Read against the present framework, those figures carry two implications: the “every release” obligations of Table 2 target a practice a majority of the market currently omits entirely, and any framework that ignores the measured barriers will fail on adoption regardless of merit.

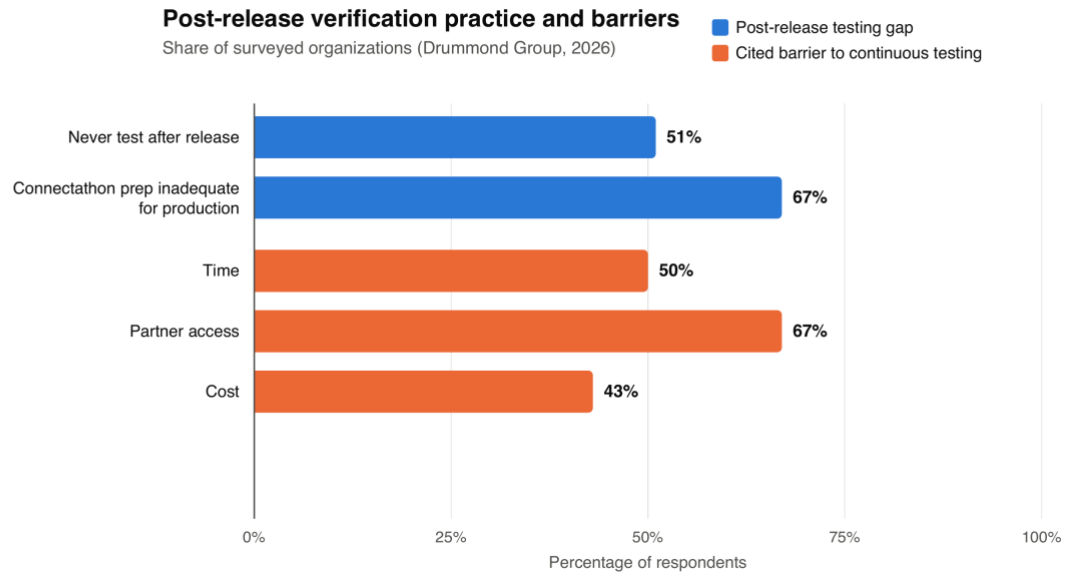


Fig. 2. Post-release verification practice and barriers among surveyed healthcare-software organizations (Drummond Group, 2026). Blue bars show the share of respondents reporting each post-release testing gap; orange bars show the share citing each operative barrier to continuous FHIR testing. Values are percentages of survey respondents.

The adoption economics deserve one further comment, because risk frameworks in software organizations fail more often from cost than from disagreement. The framework's obligations were deliberately designed against three cost ceilings. Classification cost is bounded by the workshop format and the AI-assisted pre-sort; an initial pass over a platform of moderate complexity is a matter of days, not quarters. Execution cost is bounded by automation: every recurring obligation in Table 2 is expressible as a continuous-integration gate, and the behavioral probes execute in seconds. Maintenance cost — historically the silent killer of test regimes, as suites decay against evolving implementation guides — is addressed structurally by tying regression review to profile version changes rather than to calendar schedules, so effort concentrates exactly when the specification surface a suite encodes has actually moved. None of this eliminates the organizational work of acting on findings, but it removes the standing excuse that risk-proportionate testing is a luxury of large teams.

For practitioners, the framework offers an adoptable discipline whose entry cost is low: tier classification of existing workflows is a workshop exercise, and the behavioral probes are open source and take minutes. For vendors, the four-server pattern suggests that operational transparency — visible rate-limit budgets, honest continuation links, documented negotiation contracts — is a competitive quality attribute that certification does not currently reward. For researchers, three studies follow naturally: a multi-site calibration of the tier model against retrospective defect and incident data; an inter-rater reliability study of tier assignment with and without the AI assist; and a prevalence survey of the operational behavior classes across certified production endpoints, extending the single-behavior evidence of Jones et al. (2024) to the full behavioral surface. The last of these would convert the framework's illustrative matrix into population evidence, and is a study whose entry cost is deliberately low: the behavior definitions of Section 3.3 are tool-independent, and at least one open implementation of them already exists.

5. Conclusions

FHIR-based clinical applications inherit an undifferentiated testing culture from general software practice at the same time as they inherit consequence profiles from medicine. The mismatch is resolvable. A four-tier risk model — grounded in documented prescribing-error rates, diagnostic-error attribution, and coordination-failure evidence, and sharpened by a detectability dimension that privileges silent failures — allows verification effort to follow clinical consequence rather than test-suite convention. The framework's mapping extends quality obligations beyond resource validation into operational behavior probing and boundary reconciliation, the layers where specification-conformant systems have been shown to fail. Its assistive AI component is deliberately subordinated to human judgment, in line with the alert-fatigue evidence and prevailing regulatory guidance. Application to four server implementations demonstrated that the framework discriminates operational risk where conformance status cannot. Empirical calibration across sites and defect corpora is the necessary next step, and the instrumentation for it is openly available.

Data and Code Availability: The example probe implementation used in Section 3.5 (fhir-verify), together with the probe definitions for all six behavior classes and the complete run reports, is openly available under the Apache-2.0 license at <https://github.com/pavlopalamarchuk/fhir-verify> (companion preprint: Palamarchuk P., *Beyond Certification Conformance*, Zenodo, 2026, <https://doi.org/10.5281/zenodo.22948377>); the framework itself is implementation-agnostic and reproducible from the definitions in Section 3.3. All test data are synthetic; no patient data were used or accessed in this study.

Ethics Approval: This study did not involve human participants, identifiable human data, or animal subjects; all data used were synthetic. Institutional review board approval was therefore not applicable.

Funding: This research received no external, commercial, or not-for-profit funding.

REFERENCES

1. Agency for Healthcare Research and Quality. (2020). *Making healthcare safer III: A critical analysis of existing and emerging patient safety practices* (Holmes A, Long A, Wyant B, et al.) (AHRQ Publication No. 20-0029-EF). AHRQ.
2. Agency for Healthcare Research and Quality. (2023). *Making healthcare safer IV: Potential harms resulting from patient-clinician real-time clinical encounters using video-based telehealth: A rapid evidence review* (Rosen M, Stewart CM, Kharrazi H, Sharma R, Vass M, Zhang A, Bass EB) (AHRQ Publication No. 23-EHC019-2). AHRQ. https://doi.org/10.23970/AHRQEPC_MHS4TELEHEALTH
3. Agency for Healthcare Research and Quality. (2024). *Making healthcare safer IV: Computerized clinical decision support to prevent medication errors and adverse drug events: Rapid review* (Syrowatka A, Motala A, Lawson E, Shekelle PG) (AHRQ Publication No. 24-EHC019-6). AHRQ. https://doi.org/10.23970/AHRQEPC_MHS4MEDERROR
4. Agency for Healthcare Research and Quality. (2025a). *AHRQ-funded patient safety project highlights: Improving healthcare safety by enhancing health information technology and health information exchange*. AHRQ. <https://www.ahrq.gov/sites/default/files/wysiwyg/patient-safety/highlights/ps-project-highlights-hit-hie-2025.pdf>
5. Agency for Healthcare Research and Quality. (2025b). *AHRQ-funded patient safety project highlights: Improving patient safety by enhancing care coordination*. AHRQ. <https://psnet.ahrq.gov/issue/ahrq-funded-patient-safety-project-highlights-improving-patient-safety-enhancing-care>
6. Augusto, C., Morán, J., Bertolino, A., de la Riva, C., & Tuyá, J. (2025). Software system testing assisted by large language models: An exploratory study. In *Testing Software and Systems: 36th IFIP WG 6.1 International Conference (ICTSS 2024), London, UK* (Lecture Notes in Computer Science, Vol. 15383, Ch. 17). Springer. https://doi.org/10.1007/978-3-031-80889-0_17
7. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7), e21929. <https://doi.org/10.2196/21929>
8. Bates, D. W., Leape, L. L., Cullen, D. J., Laird, N., Petersen, L. A., Teich, J. M., Burdick, E., Hickey, M., Kleefield, S., Shea, B., Vander Vliet, M., & Seger, D. L. (1998). Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *JAMA*, 280(15), 1311–1316. <https://doi.org/10.1001/jama.280.15.1311>
9. Boufahja, A., & Verma, T. (2025). Extensive conformance testing and validation of FHIR data exchange variabilities. In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOS/TEC/HEALTHINF)* (pp. 297–307). SCITEPRESS. <https://doi.org/10.5220/0013370000003911>

10. DeRosier, J., Stalhandske, E., Bagian, J. P., & Nudell, T. (2002). Using health care failure mode and effect analysis: The VA National Center for Patient Safety's prospective risk analysis system. *The Joint Commission Journal on Quality Improvement*, 28(5), 248–267. [https://doi.org/10.1016/S1070-3241\(02\)28025-6](https://doi.org/10.1016/S1070-3241(02)28025-6)
11. Devitt, D. (2026). *FHIR architecture decisions: Selecting a FHIR server and choosing between facade, hybrid, and FHIR-native approaches* (v1.0) [Self-published practitioner guide]. <https://darrendevitt.com/fhir-architecture-decisions-book/>
12. Drummond Group. (2026, August 17). *Market study finds barriers to continuous FHIR testing* [Industry report]. <https://www.drummondgroup.com/blog/market-study-finds-barriers-to-continuous-fhir-testing/>
13. Felderer, M., & Schieferdecker, I. (2014). A taxonomy of risk-based testing. *International Journal on Software Tools for Technology Transfer*, 16(5), 559–568. <https://doi.org/10.1007/s10009-014-0332-3>
14. Hauptfeld, L., Csarmann, A., Grob, M., Liepold, J., Rappelsberger, A., & Adlassnig, K.-P. (2026). Ask not what your server can do for you: A fast query compatibility strategy for navigating FHIR implementation gaps. *Studies in Health Technology and Informatics*, 338, 658–662. <https://doi.org/10.3233/shti260927>
15. Health Resources and Services Administration. (2025). *Nurse workforce projections, 2023–2038*. National Center for Health Workforce Analysis. <https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/data-research/nursing-projections-factsheet.pdf>
16. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
17. HL7 International. (2024). *FHIR specification: Validating resources*. <https://hl7.org/fhir/validation.html>
18. International Electrotechnical Commission. (2015). *IEC 62304: Medical device software — Software life cycle processes* (Ed. 1.1).
19. International Organization for Standardization. (2019). *ISO 14971: Medical devices — Application of risk management to medical devices*.
20. ISO/IEC/IEEE. (2022). *ISO/IEC/IEEE 29119-1: Software and systems engineering — Software testing — Part 1: General concepts* (2nd ed.).
21. Jones, J. R., Gottlieb, D., McMurry, A. J., Atreja, A., Desai, P. M., Dixon, B. E., Payne, P. R. O., Saldanha, A. J., Shankar, P., Solad, Y., Wilcox, A. B., Ali, M. S., Kang, E., Martin, A. M., Sprouse, E., Taylor, D. E., Terry, M., Ignatov, V., & Mandl, K. D. (2024). Real world performance of the 21st Century Cures Act population-level application programming interface. *Journal of the American Medical Informatics Association*, 31(5), 1144–1150. <https://doi.org/10.1093/jamia/ocae040>
22. Khan, M. A., Azim, A., Liscano, R., Smith, K., Chang, Y. K., Seferi, G., & Tauseef, Q. (2024). An end-to-end test case prioritization framework using optimized machine learning models. In *2024 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 1–8). IEEE. <https://doi.org/10.1109/ICSTW60967.2024.00014>
23. Kostenko, O. (2025). AI law model for ethical legislation: Strategic recommendations for the regulation of artificial intelligence. SciFormat Publishing Inc. <https://doi.org/10.69635/978-1-0690482-5-7>
24. Kostenko, O., & Drobakha, A. (2024). Metaverse: A simulacrum platform for digital transformations of the metaverse e-state society. *Journal of Educational and Health Studies*, (3), 12–20. <https://doi.org/10.61345/2734-8873.2025.1.5>
25. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5), 899–908. <https://doi.org/10.1093/jamia/ocv189>
26. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
27. Sartaj, H., Ali, S., & Gjøby, J. M. (2025). REST API testing in DevOps: A study on an evolving healthcare IoT application. *ACM Transactions on Software Engineering and Methodology*. Advance online publication. <https://doi.org/10.1145/3765744>
28. Sheffer, Y., Hardt, D., & Jones, M. (2020). *JSON Web Token best current practices* (RFC 8725). Internet Engineering Task Force.
29. Singh, H., & Sittig, D. F. (2016). Measuring and improving patient safety through health information technology: The Health IT Safety Framework. *BMJ Quality & Safety*, 25(4), 226–232. <https://doi.org/10.1136/bmjqs-2015-004486>
30. Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality and Safety in Health Care*, 19(Suppl. 3), i68–i74. <https://doi.org/10.1136/qshc.2010.042085>
31. U.S. Food and Drug Administration. (2026, January 6). *Clinical decision support software: Guidance for industry and Food and Drug Administration staff* (revised). U.S. Department of Health and Human Services. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>

32. U.S. Government Accountability Office. (2015). *Electronic health records: Nonfederal efforts to help achieve health information interoperability* (GAO-15-817).
33. U.S. Government Accountability Office. (2024a). *COVID-19: HHS needs to identify duplicative pandemic IT systems and implement key privacy requirements* (GAO-24-106638).
34. U.S. Government Accountability Office. (2024b). *Healthcare cybersecurity: HHS continues to have challenges as lead agency* (GAO-25-107755).
35. Vorisek, C. N., Lehne, M., Klopfenstein, S. A. I., Mayer, P. J., Bartschke, A., Haese, T., & Thun, S. (2022). Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: Systematic review. *JMIR Medical Informatics*, 10(7), e35724. <https://doi.org/10.2196/35724>
36. Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018a). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3), 230–238. <https://doi.org/10.1093/jamia/ocx079>
37. Walonoski, J., Scanlon, R., Dowling, C., Hyland, M., Ettema, R., & Posnack, S. (2018b). Validation and testing of Fast Healthcare Interoperability Resources standards compliance: Data analysis. *JMIR Medical Informatics*, 6(4), e10870. <https://doi.org/10.2196/10870>