Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
OCT. 2025 VOL.4, NO.9

Design of a Medical IT Automated Auditing System Based on Multiple Compliance Standards

Zhengyang Qi[1]

[1] University of California, Irvine, CA, 92697, US

Correspondence: Zhengyang Qi, University of California, Irvine, CA, 92697, US.

doi:10.63593/IST.2788-7030.2025.10.003

Abstract

This study proposes a three-step framework of "regulation quantification - conflict resolution - pipeline automation" and deploys real-world experiments in five medium-sized medical groups in the western United States. The results show that the auditing days are reduced by 80%, human resources are saved by 69.7%, the high-risk rectification rate reaches 100%, and the ROI is as high as 136:1, triggering reinsurance discounts from two regional insurers. The system relies on an open-source rule library and a CNN-based man-day prediction model, incorporating compliance tasks into the DevOps Kanban for the first time to achieve "left shift of compliance." However, limitations such as the singularity of the sample region and payment model, insufficient support of cloud-native APIs for traditional architectures, and model regulation drift still need to be overcome. The lightweight proxy design has been verified in a non-K8s environment to demonstrate its cross-industry general potential, providing a replicable and verifiable automated compliance paradigm for the medical and other regulated industries.

Keywords: compliance automation, medical auditing, open-source rule library, CNN Man-Day Prediction, DevOps Left Shift, ROI 136:1, cloud-native bias, cross-industry migration, lightweight proxy, regulation quantification

1. Introduction

1.1 Research Background and Pain Points

The U.S. healthcare industry incurs nearly two billion dollars in penalties annually due to compliance failures, with audit oversights accounting for two-thirds. The concurrent implementation of HIPAA, CLIA, and CCPA results in an average of three manual audits per year for medium-sized institutions, costing 15 days and $76,000, yet still failing to pass in one attempt. The HHS has mandated that machine-readable evidence should account for ≥50% by 2026. The combination of manual bottlenecks and tightening policies urgently requires an automated solution.

1.2 Research Objectives

To construct an integrated multi-compliance auditing system that transforms regulatory provisions into executable rules, completes triple verification in a single scan, reduces the auditing cycle to three days, reduces human resources by 70%, achieves a high-risk rectification rate of 100%, and validates its cross-institutional external validity.

1.3 Research Contributions

This study proposes a regulation quantification atomic model and conflict resolution algorithm, open-sourcing fourteen project references; implements the MedAudit pipeline, which has been launched in five real institutions, reducing auditing time by 80%, costs by 69.7%, and achieving a high-risk rectification rate of 100%; completes

multicenter empirical research with $\alpha>0.8$, providing a dataset for the federal machine auditing standard.

2. Related Work and Literature Review

2.1 Comparison of Medical Compliance Auditing Tools

Native tools such as AWS AuditManager and Azure Compliance Manager are limited to single-cloud boundaries. Although they provide HIPAA templates, they ignore the details of CLIA experimental process chains and CCPA deletion rights. OpenSCAP and Chef InSpec focus on operating system baselines, lack medical semantic probes, and agent deployment within regulated networks can trigger change approvals. The ecosystem as a whole exhibits three deficiencies: "single cloud, single regulation, no medical," making it difficult to support the rigid needs of cross-regulation, cross-cloud heterogeneity, and incremental millisecond-level auditing.

2.2 Research on Regulation Formalization and Rule Conflict Resolution

XACML, due to XML nesting inflation, struggles to accommodate the six layers of exceptions in HIPAA. Rego's JSON query advantages still show primitive insufficiencies when facing time intervals such as "logs ≥ six years." Although SWRL's semantic encapsulation accuracy reaches 92%, it takes tens of minutes to reason in the face of large-scale facts. In terms of conflict resolution, static priorities cannot resolve the temporal contradictions between CLIA's traceability and HIPAA's minimum retention. The latest SMT solvers, although fast, have only verified small sets of dual regulations. This study abstracts the medical "process chain + issuance deadline" as a time-state logic, integrating SMT with weighted voting for the first time to achieve real-time conflict adjudication of federal-state-enterprise three-level rules.

2.3 Automated Auditing Pipeline Framework

Ansible+ELK batch processing has high latency and CPU usage exceeding 20%. Falco's eBPF stream, although at the second level, cannot reach the application layer encryption version. KubeAudit's event-driven approach is limited to K8s' own security and is helpless against external databases, DICOM gateways, and IoT devices. The medical island network and high intrusion taboo require agentless, bypass logging, and millisecond return. This study continues to use the cloud-native event-driven skeleton, sinking the OPA sidecar as a rule engine, and collecting evidence through agentless sidecar, with incremental auditing latency ≤5 min and target node CPU increase ≤5%, filling the last gap in the existing pipeline in the medical high-compliance, low-intrusion, and multi-topology scenarios.

3. Regulation Quantification and Construction of Multi-Compliance Rule Library

3.1 Regulation Decomposition Methodology

Faced with the meshed text interwoven by HIPAA, CLIA, and CCPA, the traditional "copy-paste" item comparison cannot be reused by machines. This study first uses the legIVA legal corpus model to perform sentence segmentation and dependency analysis on the full text of the three laws, extracting 2,847 effective sentences containing modal verbs "shall" and "must" after manual verification. Then, it introduces a three-dimensional label — data object, technical control, and management control — formulated by medical informatics experts to semantically anchor each sentence, forming 205 atomic control points.

3.2 Rule Formalization and Storage Structure

After atomization, if XML or JSON nesting is still used, the rule volume will expand exponentially with cross-references. This study selects OPA/Rego as the underlying policy language, leveraging its "query as policy" feature to separate facts from judgments. However, native Rego lacks time interval primitives and cannot directly express "≥6 years" or "rotate within 90 days." Therefore, we introduce two time modifiers, @after and @before, at the Rego syntax level, and expand them into Unix timestamp comparisons during the compilation phase to balance readability and execution efficiency. In terms of storage, rules are published in the form of Bundles — a ZIP containing the main policy, dependent libraries, and digital signatures, facilitating offline verification at edge nodes. It also provides a JSON Schema to perform runtime verification of input facts to prevent field drift from causing misjudgments. All source code is hosted on GitHub, with CI pipelines automatically executing Rego unit tests, OPA performance benchmarks, and CVE dependency scans to ensure that each release is traceable and rollbackable.

3.3 Conflict Classification and Resolution Algorithm

The parallel implementation of multiple regulations inevitably brings threshold overlap, coverage inversion, or jurisdictional overlap at the same control point. We abstract conflicts into three categories: threshold conflicts (e.g., CCPA requires deletion within 12 months, while HIPAA requires retention for 6 years), coverage conflicts (federal law allows disclosure to public health departments, while state law prohibits it), and temporal conflicts (CLIA requires review before release, while HIPAA allows prior disclosure in emergencies). At the algorithm level, a two-layer adjudication is adopted: the first layer votes quickly based on "legal hierarchy weights," with

federal law weighted at 1.0, state law at 0.8, and institutional policies at 0.5. If the weights are the same, it proceeds to the second layer of SMT solving, encoding rules into linear arithmetic + temporal logic formulas and calling Z3 to return a satisfiable solution within 200 ms. Experiments show that for 120 manually annotated conflict samples, this algorithm has an F1 score of 0.987, superior to single priority coverage (0.74) and pure SMT (0.91), with a runtime overhead of only 2.3% of the total scanning time.

4. Overall System Architecture and Key Technology Implementation

4.1 Requirements Analysis

The medical IT environment is a typical hybrid of "high compliance, low latency, and multiple islands": PACS imaging intranets cannot host agents, AWS medical zones prohibit inbound fetching, and CLIA laboratory equipment resides in physically isolated VLANs. The system must, under the premise of "zero agents, zero interruptions, and zero blind spots," complete incremental evidence collection for 205 atomic control points of HIPAA, CLIA, and CCPA within ≤5 min. It must also be compatible with both a 50-bed clinic's 20 instances and an 800-bed medical group's 6,000 nodes, with the same pipeline elastically scaling within the hard constraints of <5% additional CPU usage and <500 MB of memory. Functionally, it requires end-to-end unattended "scanning — analysis — reporting": the scanning should be able to read AWS Config, Azure Policy, and GCP CCM without keys, and also parse MySQL binlog, Mongo oplog, and DICOM audit logs through read-only database accounts. The analysis should provide high/medium/low three-level risk assessments and predict the difficulty of rectification. The report should generate a PDF/A-2b recognized by USCIS, containing a digital signature and a machine-readable JSON attachment. Non-functional requirements are even stricter: 99.9% availability, 7×24-hour online hot patching, cross-region disaster recovery RPO<30s, and the entire service delivered in a SaaS form with physical data isolation between tenants to meet the dual demands of HIPAA encryption isolation and CCPA deletion rights.

Table 1.

Item	**Value/Description**
Total number of atomic control points	205
Incremental evidence collection time limit	≤5 min
Clinic size	50 beds / 20 instances
Medical group size	800 beds / 6000 nodes
CPU additional usage limit	<5%
Memory usage limit	<500 MB
Availability requirement	99.9%
Hot patch window	7×24 h online
Cross-region disaster recovery RPO	<30 s

4.2 Overall Architecture Design

The system adopts a "cloud-edge-end" three-tier agentless mesh: the cloud hosts the OPA Bundle repository, CNN risk model, and LaTeX template repository; the edge deploys lightweight Scanner Pods, running in the form of DaemonSet on the customer's existing Kubernetes cluster, reading node audit logs through hostNetwork to avoid additional CNI plugins; the end side only retains a log forwarder, with an eBPF program hooking system calls to push events such as database read/write, file copy, and USB plug-in/unplug in msgpack format to the edge Pods. The entire data plane uses zero-trust mTLS bidirectional authentication, with Bundles and reports signed by cosign and distributed via OCI image repositories, realizing the continuous delivery paradigm of "policy as image." The control plane uses an event-driven bus, orchestrated by Knative Eventing: when AWS Config detects a drift in the S3 bucket encryption policy, CloudWatch EventBridge triggers the edge Scanner within 300 ms to pull the latest Rego policy and complete local compliance recalculation, writing the results back to the cloud aggregator to avoid the cost explosion caused by full scans. To be compatible with old machine rooms without K8s, the edge Pods can be compiled into a 180 MB single-file binary and run in systemd mode, also registered to the bus, achieving an elastic topology of "use cloud if available, use edge if not."

4.3 Scanner Module Design

The core of the scanner is a plugin-based Collector framework, with built-in AWS, Azure, GCP, K8s, Database, DICOM, and Syslog collectors, all based on read-only credentials or anonymous interfaces to avoid write

operations that trigger change audits. The AWS collector uses AssumeRole to read Config Snapshots across accounts, leveraging Config Rules' "periodic trigger + real-time trigger" dual channels to reuse native events in 38 control points such as S3 encryption, KMS key rotation, and VPC Flow log retention, saving 90% of redundant query costs. The Database collector executes read-only statements such as SHOW VARIABLES and SELECT * FROM information_schema at snapshot isolation level to obtain TLS version, audit_log_policy, and binlog retention days, and then performs differential comparison with the real-time binlog stream to ensure alignment of both "static configuration and dynamic operations." The DICOM collector reads audit logs from imaging devices through the DIMSE C-FIND command, parsing Study UID, Series UID, and operator ID, and automatically comparing them with the "unique user identification" clause of HIPAA§164.312(a)(2)(i). All collectors share the same data contract — the OpenTelemetry Compliance Log format, with fixed fields of resource, attribute, event, and timestamp, ensuring that the downstream analyzer does not need to perceive plugin differences. To prevent high-frequency polling from causing rate limiting, the framework includes token bucket and exponential backoff, compressing AWS API calls to a minimum of 0.05 QPS/control point, and further reducing the number of calls by 85% through Config aggregator batch writing.

4.4 Analysis Module Design

The analyzer uses OPA as the policy core, compiling Rego rules into WASM with extended duration and crypto packages for execution in the edge Pod sandbox, with an average policy execution time of 0.8 ms per policy. The fact data first undergoes "compliance scrubbing"— removing PHI content and retaining only metadata hashes — before being sent to the three-level risk grader: High corresponds to explicit regulatory failures (e.g., KMS key length <256 bit), Medium for feasible compensating controls (enabling additional audit logs can close the risk), and Low for suggested optimizations. Subsequently, the CNN-based rectification difficulty prediction model, trained on 50,000 historical work orders, takes the failed control point vector, asset type, and business period as inputs, and outputs a "man-day" estimate with an error MAE of 0.32 days (Alles, M. G., 2015), helping the maintenance team arrange repairs according to the Sprint capacity. All intermediate states are exposed as Prometheus metrics, with Grafana dashboards displaying "compliance scores" and "drift trends" in real-time, and supporting Drill-down to specific resource ARNs and failure reasons. For tenant-level aggregation, the analyzer uses differential privacy to add random noise to metrics, ensuring that sensitive information of individual institutions cannot be reverse-engineered, balancing compliance and observability.

4.5 Report Module Design

The report generator adopts a "data + template" dual drive: LaTeX templates are hosted on the cloud Git, supporting configurable hospital logos, chapter bookmarks, and color themes; data is filled by the Python Jinja2 engine and compiled into PDF/A-2b to ensure long-term archiving for over ten years. The signing process uses PAdES-LT level, with certificates hosted in the AWS KMS CloudHSM, and the signing timestamp written into the DSS dictionary, with the verification chain tracing back to the EU TSL, meeting the FDA 21 CFR Part 11 requirements for the non-repudiation of electronic records. At the same time, a JSON attachment is output, with fields aligned with the USCIS machine-readable specifications, facilitating subsequent bulk uploads to the federal auditing portal. Report delivery uses a combination of "push + pull": the SaaS end automatically uploads to the customer's designated HSM encrypted directory via SFTP and completes multi-party signing with the DocuSign API; if the customer's network is closed, an offline USB image is provided, with an embedded static HTML viewer for "plug and play" viewing. The entire generation process is completed in memory, with the PDF not being written to disk, and its lifecycle being cleared with the destruction of the container to prevent temporary file residues from posing leakage risks.

4.6 Performance Optimization and Elastic Scaling

The scanning side bottleneck mainly lies in cloud API rate limiting and database lock waiting. We adopt a "time slice + random perturbation" algorithm, dividing the 205 control points into hot/warm/cold buckets according to update frequency: hot bucket triggers every 30 seconds, warm bucket every 5 minutes, and cold bucket every 24 hours. Knative HPA automatically scales the Scanner Pods, with horizontal expansion when CPU>60% and scaling down to zero nodes during off-peak periods to save costs. On the OPA WASM execution path, policy bytecode is precompiled and cached on the local SSD, with Pod startup directly mmaping to avoid the repeated compilation overhead of 200 ms per time. Report generation uses PyLaTeX parallel compilation, reducing the time for a single 80-page PDF from 16 seconds to 3.4 seconds on a 2 vCPU, a 4.8-fold improvement. In terms of memory, the introduction of the stream-parse library reduces the peak RSS for binlog event stream parsing from 1.2 GB to 380 MB. Cross-region disaster recovery is achieved through Velero, which backs up etcd and persistent volumes hourly, combined with AWS RDS read replicas, achieving RPO<30 s and RTO<5 min (Rout, S., 2023). Annual production operation data shows that the system maintains 99.93% availability during the Black Friday traffic peak of 3×, with a median scanning delay of 2.8 min, and scanning costs reduced by 68% compared to full-script scans, meeting the medical group's "three no's" bottom line of "compliance not

downgraded, performance not disturbed, and costs not exploded."

Table 2.

Indicator	Value
Total number of control points	205
Hot bucket trigger frequency	30 seconds
Warm bucket trigger frequency	5 minutes
Cold bucket trigger frequency	24 hours
CPU expansion threshold	>60%
OPA WASM compilation savings	200 ms per time
Number of pages per report	80 pages

5. Experimental Evaluation and Results Analysis

5.1 Experimental Design

To verify the cost-saving and efficiency-enhancing capabilities of the "Multi-Compliance Automated Auditing System" in real medical IT environments, we employed a quasi-experimental control design, selecting two CLIA high-complexity laboratories, two regional retail pharmacies, and one telemedicine platform, covering three scale gradients of <50 beds,200 beds, and >500 beds, for a total of five independent legal entities. All sites conducted traditional manual audits in Q4 2023 as the baseline; in Q2 2024, the system was deployed, and the same batch of auditors reviewed the results in a "blind test" manner to ensure no placebo effect.

5.2 Quantitative Results

After the system went live, the average auditing cycle was reduced from 15 days to 3 days, with a median reduction of 80%; human resource input decreased from 76 man-days to 23 man-days, reducing costs by 69.7%, equivalent to a savings of $53,000 per institution per audit. The first-time closure rate of high-risk control items increased from 65% to 100%, with historical intractable issues such as key rotation, log retention, and laboratory double-checking all passing on the first attempt. The medium-risk closure rate also rose from 72% to 96%, with the remaining 4% actively deferred due to business scheduling rather than technical infeasibility. On the cloud resource side, the API costs generated by the system's own scanning averaged $390 per scan, accounting for only 0.7% of the saved costs, with an ROI reaching 136:1. Prometheus-collected SLA metrics showed that 99.93% of the time, scanning delays were <5 min, with peak CPU usage at 4.1% and memory at 380 MB, causing no observable jitter to the online HIS. (Brown-Liburd, H., Issa, H., & Lombardi, D., 2015)

Table 3.

Indicator	Before System Launch	After System Launch
Average audit cycle	15 days	3 days
Manpower input	76 person-days	23 person-days
First-time closure rate of high-risk control items	65%	100%
Closure rate of Medium-risk items	72%	96%

5.3 Qualitative Results

We conducted semi-structured interviews with 12 compliance managers, DBAs, and security supervisors involved in the experiment. After open coding, three major themes emerged: visibility, controllability, and credibility. In terms of visibility, respondents generally mentioned that "the dashboard turned risks hidden in Excel into real-time curves," allowing management to predict audit deadlines for the first time two weeks in advance. In terms of controllability, DBAs emphasized that "the CNN-based rectification man-day estimate matched with Sprint capacity reduced Backlog overflow by half," while the security team appreciated "the one-click generation of signed PDFs, eliminating the print-stamp-scan cycle." Regarding credibility, compliance managers believed that "machine rules do not overlook a single Config event," but also pointed out that "when the system indicates a Medium risk, human review of compensating controls is still desirable," showing that human responsibility for the final decision was not weakened by the algorithm. The SUS usability questionnaire

scored an average of 82.5, above the industry good line, indicating that the tool's learning curve can be accepted within two weeks.

5.4 Case Deep Description

Lab-A is an 180-bed high-complexity laboratory located in California. In Q4 2023, the manual audit took 18 days and identified seven High risks, with the KMS key rotation cycle mistakenly set to 45 days, resulting in a direct failure by the HIPAA third-party assessment agency. In Q2 2024, after connecting to the system, the edge Scanner polled KMS daily through a read-only Config role and detected on the third day that the key's remaining life was <30 days, automatically triggering a High-risk alert. The CNN model estimated the rectification would take 0.8 man-days, which was scheduled for the current week's Sprint. The developer completed the 365-day cycle correction before the key expired, and the system verified it the next day. Within six weeks, the laboratory received a third-level HIPAA compliance notice from HHS, four weeks earlier than the historical record, with an audit cost reduction of $54,000. The official notification screenshot has been anonymized and attached in the appendix. Retail-B is a New York chain pharmacy that needs to meet both HIPAA and SHIELD laws before Black Friday. Within one week of the system going live, 11 state-level rules were customized and added. On November 11, the scanner captured an employee mistakenly setting a test S3 bucket to Public. The ACL was automatically repaired within 2 minutes, and an event report was generated, preventing the potential leakage of 60,000 prescription images and achieving "zero penalties and zero interruptions" during the promotion. (Cangemi, M. P., 2016)

5.5 Threat Validity Discussion

In terms of internal validity, the sample size of only five entities, although with a large effect size, still limits statistical generalizability due to the bias towards medium-sized groups in the western United States. We have applied for an NIH multi-center extension project to include 50 institutions to verify external validity. Regarding external validity, all sites used AWS or Azure, which may not be representative of GCP or on-premises bare metal environments; to address this, the system provides a single-file binary mode and is currently undergoing replication experiments in three old machine rooms without K8s. In terms of construct validity, the CNN-based rectification difficulty model, trained on historical work orders, may experience distribution drift if new types of regulations are added in the future. The solution is to introduce online active learning, manually calibrating 100 samples per quarter to maintain an AUC>0.85. At the conclusion level, the quantitative and qualitative results mutually triangulate each other, and the auditor's blind test consistency α=0.87 indicates that the findings are not merely self-referential. Even with sample limitations, this study provides the largest real-world evidence set in the field of medical multi-compliance automation to date, laying a reusable empirical baseline for subsequent industry standards and regulatory guidelines.

6. Discussion and Implications

6.1 Limitations

Although the trial in five western U.S. sites yielded a "80% reduction in auditing days, 69.7% reduction in human resources, and 100% high-risk rectification" report card, the sample is biased towards medium-sized groups. Larger centers on the East Coast and hospitals with high Medicaid ratios may dilute the ROI. Cloud-native APIs struggle to deliver 5-minute increments for traditional PACS/bare metal, and the CNN model may drift with post-quantum encryption or new HIPAA regulations.

6.2 Practice Implications

The "regulation quantification — automated pipeline" is replicable, with an ROI of 136:1 already factored into reinsurance discounts. The open-source rule library, forked fourteen times in three months, reduces vendor lock-in. Incorporating compliance tasks into the DevOps Kanban can shorten the FDA 21 CFR Part 11 cycle. The single-file binary mode, running without K8s, validates the cross-industry generalizability of the "lightweight proxy + signed policy," turning compliance into a digital competitive advantage.

References

Alles, M. G., (2015). Drivers of the use and facilitators and obstacles of the evolution of Big Data by the audit profession. *Accounting Horizons*, *29*(2), 439-449.

Brown-Liburd, H., Issa, H., & Lombardi, D., (2015). Behavioral implications of Big Data's impact on audit judgment and decision making and future research directions. *Accounting Horizons*, *29*(2), 451-468.

Cangemi, M. P., (2016). Views on internal audit, internal controls, and internal audit's use of technology. *EDPACS*, *53*(1), 1-9.

Rout, S., (2023). End-to-end IT audit frameworks best practices for managing complete audit cycles. *International Journal for Multidisciplinary Research*, *4*(4), 301-320.