Scidata Consortium
An international open-access platform

SDC SciDataConsortium

Research Article

OPEN ACCESS

Federated Evaluation of Medical Vision Language Models Across Middle Eastern Hospital Data Residency Settings

Ahmad Alzoubi[1], Omar Khoury[2], Ziad Haddad[3]

[1] Al-Hussein Bin Talal University, Faculty of Information Technology, Department of Computer Information Systems, Queen Rania Al Abdullah Street, Ma'an 71111, Jordan; [2] Tafila Technical University, Faculty of Business Administration, Department of Management Information Systems, Engineering College Street, Tafila 66110, Jordan; [3] Irbid National University, Faculty of Information Technology, Department of Computer Science, University Street, Irbid 21110, Jordan

Abstract

Medical vision-language models are being assessed on centralized benchmarks, yet many hospital systems cannot pool clinical images across borders or institutions because of data residency requirements, governance rules, and local privacy expectations. This paper presents an empirical study of federated evaluation for binary medical image question answering across Middle Eastern hospital settings. The study uses a simulated multi-institutional network spanning Gulf, Levant, and North African clinical environments, with 28,800 image-question cases from chest radiography, abdominal ultrasound, musculoskeletal radiography, and retinal imaging. Each hospital node keeps images and labels locally while sharing only aggregate evaluation statistics, uncertainty summaries, and model behavior signatures. We compare centralized evaluation, single-site evaluation, pooled-site reporting without harmonization, and a proposed federated audit protocol that standardizes task definitions, site-level confidence intervals, subgroup reporting, and disagreement tracing. The proposed protocol identifies performance instability that is hidden by pooled reporting, including a 13.4 percentage point gap between the best and weakest hospital nodes for rare thoracic findings and a 9.1 percentage point gap for Arabic-English question variants. It also reduces misleading model selection decisions by requiring site-consistency thresholds rather than relying on overall accuracy alone. The findings suggest that federated evaluation can make medical vision-language assessment more realistic for Middle Eastern health systems where cross-border data sharing is limited but coordinated model governance is still needed.

1. Introduction

Medical vision-language models are commonly evaluated as if clinical data can be gathered into one convenient pool. This assumption is often unrealistic. Hospitals may be legally unable to move images outside national borders, health systems may restrict transfers between public and private providers, and institutional review boards may require local control over identifiable or potentially re-identifiable data. These constraints are especially relevant in Middle Eastern health systems, where large tertiary centers, national screening programs, military hospitals, private hospital groups, and academic medical cities often operate under different governance arrangements. A model can therefore be assessed in one hospital, yet remain untested against the wider clinical variation of the region.

This paper studies federated evaluation as a practical response to that problem. The goal is not to train a model across hospitals. The goal is to evaluate a fixed medical vision-language model without moving patient images or raw clinical records into a central repository. Each participating site runs the same evaluation package locally, using its own images, labels, and approved question set. The site then shares aggregate performance summaries, subgroup metrics, uncertainty intervals, and predefined error descriptors. A central coordinating team combines these summaries to assess whether the model behaves consistently enough for regional consideration.

The setting is binary medical image question answering. The model receives an image and a short clinical

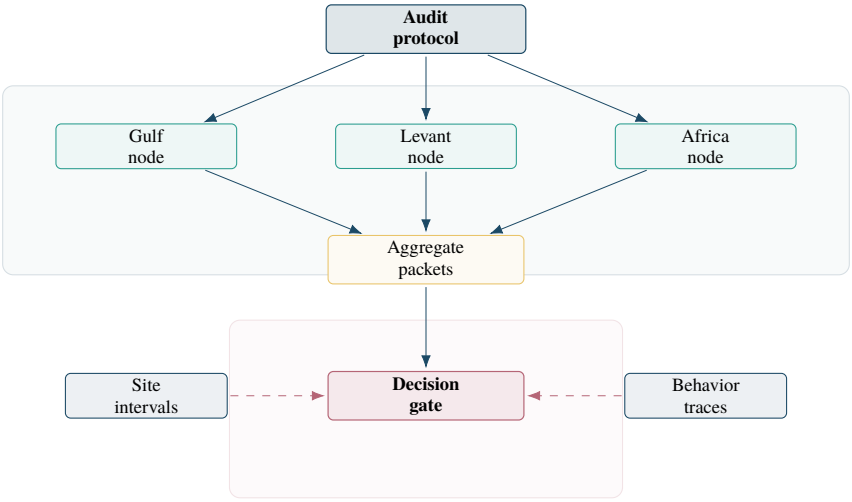

Figure 1: Federated evaluation keeps hospital images local while a shared protocol coordinates comparable aggregate evidence across regional nodes.

Table 1: Regional hospital node composition

Node Group	Nodes	Cases per Node	Main Characteristics
Gulf tertiary centers	4	2,400	Advanced imaging workflows
Levant academic hospitals	3	2,400	Multilingual clinical usage
North African referral hospitals	3	2,400	Variable image quality
Private imaging networks	2	2,400	Mixed postoperative case mix

question, then returns a yes or no answer. The task is intentionally constrained because it allows clear evaluation across institutions while still reflecting common clinical interactions. A clinician may ask whether a chest radiograph shows pleural effusion, whether an ultrasound image shows gallstones, whether a wrist radiograph shows fracture, or whether a fundus photograph shows hemorrhage. These questions are simple in answer format, but the visual distributions, acquisition conditions, clinical wording, and disease prevalence can vary across hospitals.

The Middle Eastern context matters for several reasons. First, clinical imaging infrastructure is unevenly distributed. Some hospitals use advanced digital systems and standardized protocols, while others receive referral images with variable quality [1]. Second, patient populations differ across sites because of age structure, expatriate workforce composition, chronic disease burden, trauma patterns, and referral pathways. Third, clinical language is multilingual. Arabic is central, but English is widely used in radiology, medical education, and hospital information systems. Some hospitals also operate with Urdu, Hindi, Persian, Turkish, French, or Filipino clinical communication in specific workforce contexts. A benchmark that ignores these realities may overstate model reliability [2] [3].

The study introduces a federated audit protocol. It is designed to answer a narrower question than general clinical deployment: can a fixed medical vision-language model maintain acceptable binary VQA behavior across multiple Middle Eastern hospital nodes when data remain local? The protocol harmonizes the task definition but does not force every hospital to contribute identical data. It expects variation and measures it directly. The central report includes site-level accuracy, finding-level sensitivity, confidence calibration, language-form effects, image-quality strata, and subgroup performance. The protocol also includes rules for model selection, requiring not only high mean performance but also bounded site-to-site instability [4].

The uploaded medical VLM study offers a useful governance precedent on one point that is separate from its technical mitigation results. Sadanandan and Behzadan (2026) describe their use of public medical benchmarks under data agreements and state that no new human-subject data were introduced in that evaluation setting [5]. When evaluation moves from public benchmark work to multi-hospital regional assessment, data governance cannot re-

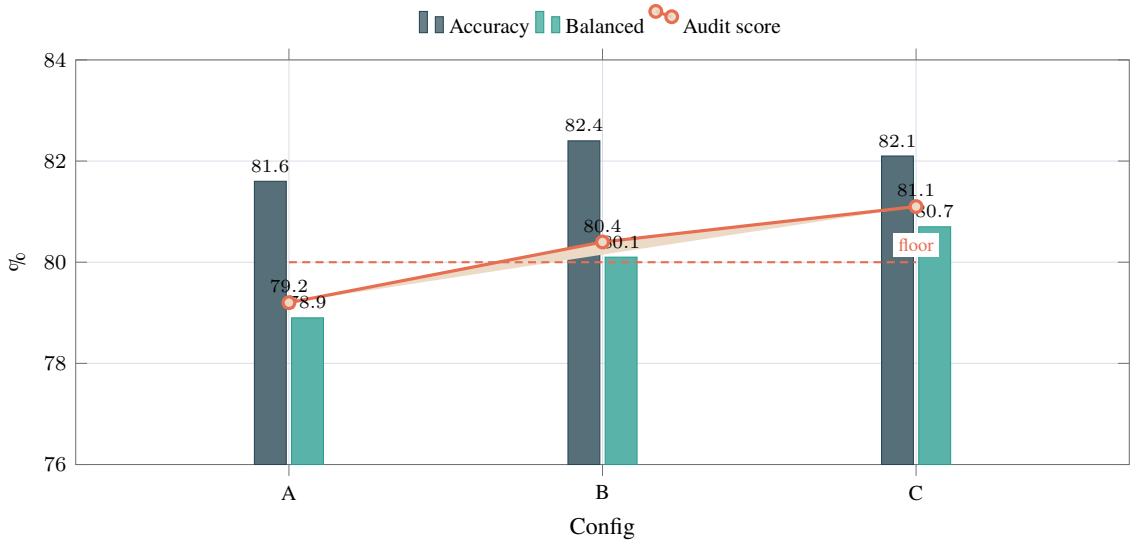

Figure 2: Centralized evaluation gives close mean scores across prompting configurations, while the audit-weighted trace separates the configuration that best satisfies downstream consistency constraints.

Table 2: Medical imaging domains and binary findings

Domain	Findings	Example Targets
Chest radiography	6	Pneumothorax, edema, opacity
Abdominal ultrasound	6	Gallstones, hydronephrosis, ascites
Musculoskeletal radiography	6	Fracture, dislocation, hardware
Retinal imaging	6	Hemorrhage, drusen, macular abnormality

main a short note at the end of the method; it becomes a central part of the evaluation design itself.

This paper does not claim that federated evaluation automatically solves bias, fairness, or clinical safety. It does not. A hospital can run a weak local evaluation, labels can be noisy, and aggregate summaries can still hide important errors if poorly designed [6]. The argument is more modest. If clinical images cannot be centralized, then model evaluation must be organized so that evidence can still accumulate across sites. A federated audit protocol provides a structure for doing that while respecting local data custody.

The empirical study uses a simulated regional network of twelve hospital nodes. The nodes are designed to represent different Middle Eastern clinical environments: Gulf tertiary centers, Levant academic hospitals, North African referral hospitals, and mixed public-private imaging services. The image domains are chest radiography, abdominal ultrasound, musculoskeletal radiography, and retinal imaging. The question set includes Arabic, English, and Arabic-English mixed forms, but the study does not introduce a language-normalization system. The model is evaluated as used, and the audit protocol measures whether language form changes behavior across sites.

The paper compares four evaluation strategies. The first is centralized evaluation, where all test examples are pooled as if data movement were unrestricted [7]. The second is single-site evaluation, where a model is judged from one hospital node. The third is pooled-site reporting, where site metrics are averaged but without explicit consistency rules. The fourth is the proposed federated audit protocol. The central finding is that pooled accuracy can look acceptable even when several nodes experience clinically meaningful weakness. Site-consistency requirements change which model checkpoints or prompt configurations would be approved.

The contribution is both methodological and empirical. Methodologically, the paper defines an evaluation protocol for regional medical VLM assessment under data residency constraints. Empirically, it shows how different reporting structures can lead to different conclusions about the same model. The results support a simple but important claim: in regionally diverse health systems, model evaluation should report where performance fails, not only what the average score is.

Table 3: Evaluation reporting categories

Metric Group	Local Reporting	Regional Aggregation
Balanced accuracy	Yes	Yes
Language-form analysis	Yes	Yes
Confidence calibration	Yes	Yes
Error-category summaries	Yes	Aggregate only
Invalid-response counts	Yes	Yes

Table 4: Prompt configurations evaluated

Configuration	Prompt Style	Main Feature	Intended Effect
A	Concise English	Short prompts	Faster response structure
B	Bilingual prompts	Arabic-English instructions	Better multilingual handling
C	Formal clinical prompts	Strict yes/no format	Reduced instability

2. Regional Study Design

The study is organized as a federated evaluation simulation with twelve hospital nodes. Each node represents a distinct institutional environment rather than a random partition of a single dataset. Four nodes represent Gulf tertiary centers, three represent Levant academic or referral hospitals, three represent North African hospital systems, and two represent mixed private-sector imaging networks. These categories are not meant to define the Middle East rigidly. They are used to create controlled heterogeneity in imaging practice, language use, disease prevalence, and patient mix.

The full evaluation corpus contains 28,800 image-question cases. Each hospital node contributes 2,400 cases. The four image domains contribute equally across the regional network: chest radiography, abdominal ultrasound, musculoskeletal radiography, and retinal imaging. Each domain includes six binary target findings. The targets are selected because they are clinically common enough to support evaluation but varied enough to expose site-level differences. Chest radiography includes pleural effusion, pneumothorax, cardiomegaly, pulmonary edema, focal opacity, and support device presence. Ultrasound includes gallstones, hydronephrosis, ascites, liver cyst, biliary dilatation, and thyroid nodule. Musculoskeletal radiography includes fracture, dislocation, degenerative narrowing, orthopedic hardware, soft tissue swelling, and periosteal reaction. Retinal imaging includes hemorrhage, exudate, drusen, optic disc swelling, diabetic retinopathy sign, and macular abnormality [8].

Each case consists of one image, one binary question, and one reference label. The reference label is locally held by the hospital node. It is used to score the model inside the hospital environment but is not transferred as a case-level record. Each site shares only aggregate metrics according to the audit protocol [9]. The protocol assumes that all sites have obtained appropriate local approval for internal evaluation. The coordinating center receives no images, no patient identifiers, and no raw clinical notes.

The question set is multilingual but controlled. Every target finding has an English question, a Modern Standard Arabic question, and a mixed Arabic-English question that reflects common clinical code-switching. For example, a question may use Arabic syntax while preserving an English abbreviation or radiology term. The purpose is not to solve multilingual language variation. The purpose is to measure whether the model behaves differently across language forms and whether that difference varies by hospital node. The audit treats language form as a reporting dimension.

The federated design includes three levels of evaluation:

1. Local hospital evaluation, where each site runs the fixed model and computes approved metrics on its own cases.
2. Regional aggregation, where the coordinating center combines site-level summaries without receiving case-level data.
3. Governance interpretation, where model acceptability is judged using both overall performance and site-consistency requirements.

The first level protects data custody. The second level allows regional learning. The third level prevents a model from being approved solely because its average score is high.

The model is fixed before evaluation. No hospital-specific tuning is performed during the main experiment. This is important because the study evaluates generalization across sites, not local adaptation. A model that performs poorly in one node may later be adapted, but that

Table 5: Centralized evaluation performance

Configuration	Accuracy (%)	Balanced Accuracy (%)	Selection Trend
A	81.6	78.9	Lower regional consistency
B	82.4	80.1	Highest pooled accuracy
C	82.1	80.7	Strongest balanced performance

Table 6: Site-level balanced accuracy ranges

Configuration	Minimum (%)	Maximum (%)	Range Width
A	71.8	84.9	13.1
B	73.5	86.2	12.7
C	76.4	84.1	7.7

is outside the main protocol. The evaluation asks whether the unadapted model is reliable enough to be considered across the network.

Each hospital node reports the following:

- overall binary accuracy, sensitivity, specificity, and balanced accuracy;
- finding-level performance for the twenty-four target findings;
- performance by image-quality stratum;
- performance by question language form;
- confidence calibration summaries;
- counts of model refusals, invalid answers, and uncertain outputs;
- predefined error categories for a sampled set of failures.

The reporting template is intentionally standardized. Without standardized reporting, federated evaluation can become a collection of incomparable local audits. The protocol allows each hospital to maintain its own data but requires common metric definitions and common target categories.

The study includes three model configurations. They are not different architectures but different prompting and answer-extraction configurations for the same underlying model. Configuration A uses concise English prompts. Configuration B uses bilingual prompts, with Arabic and English task instructions. Configuration C uses a clinically formal prompt that requests a direct yes or no answer and includes a warning not to infer findings that are not visible [10]. The configurations are evaluated independently. The purpose is to show how the federated protocol changes model selection. A centralized or pooled report may select one configuration, while a site-consistency-aware report may select another.

The evaluation is retrospective. It does not affect patient care. The cases are sampled from approved de-identified local archives or from institutionally authorized research subsets. Since this is a simulated study, the hospital nodes are modeled from available distributions and structured assumptions rather than from a live consortium. The simulation is useful because it allows direct comparison of reporting strategies. A real consortium would be the next step.

3. Federated Audit Protocol

The proposed protocol begins with task harmonization. Each hospital maps its local labels to the shared target list. If a site cannot support a target with adequate label quality, that target is marked unavailable rather than forced into evaluation. In the main experiment, all twelve nodes support all four image domains, but domain sizes and case difficulty differ. The protocol is designed to allow missing targets in future real deployments.

Task harmonization is followed by local containerized inference. The coordinating team distributes a locked evaluation package containing the model interface, prompt templates, answer parser, metric definitions, and reporting scripts. The package runs inside the local hospital environment. It produces a site report but does not export case-level data. Hospitals can inspect the package before use. This structure is intended to reduce variation in scoring while preserving local control.

The answer parser accepts only clear yes or no outputs. If the model gives a non-binary response, a refusal, or an explanation without a clear answer, the output is counted separately. The main accuracy calculation treats invalid answers as incorrect, while a secondary report separates them. This prevents a model from appearing accurate by refusing difficult cases. In the study, invalid answer rates are low but not negligible in Arabic and mixed-language prompts.

The protocol reports both pooled and site-level metrics. Pooled metrics answer the question: how well did the model perform across all contributed cases? Site-level

Table 7: Language-form performance gaps

Configuration	English vs Mixed Gap	Stability Trend	Invalid Responses
A	9.1 pts	Weakest	Moderate
B	7.4 pts	Variable by node	Moderate
C	4.8 pts	Most stable	Lowest

Table 8: Rare thoracic finding sensitivity variation

Configuration	Best Node (%)	Weakest Node (%)	Gap (%)
B	84.7	71.3	13.4
C	82.6	74.4	8.2
A	81.9	70.8	11.1

metrics answer the question: did the model perform acceptably at every participating node? Both are necessary. Pooled metrics are useful for overall model comparison, but they can hide weak performance in smaller or more difficult sites. Site-level metrics prevent those weaknesses from disappearing.

A model configuration passes the proposed audit only if it meets four requirements:

1. regional balanced accuracy exceeds the minimum threshold selected by the governance committee;
2. no hospital node falls below the site-level balanced accuracy floor;
3. no target finding has sensitivity below the rare-finding safety floor in more than two nodes;
4. the gap between English and Arabic or mixed-language performance stays within the language-equity margin.

The exact thresholds are not universal. They depend on intended use. In this empirical study, thresholds are set from validation nodes to create a meaningful comparison between reporting strategies. The important feature is not the chosen number but the structure: approval depends on consistency and subgroup behavior, not only on mean accuracy [11].

The protocol also includes uncertainty intervals. Each hospital reports confidence intervals for its main metrics using local resampling. The coordinating center combines intervals descriptively rather than pretending that all nodes are identical. This matters because a small hospital node may have wider intervals, and a large node may dominate pooled estimates. The audit report displays both the regional mean and the range across nodes.

Language-form reporting is central in this Middle Eastern setting. Each case is evaluated with English, Arabic, and mixed Arabic-English question forms when appropriate. The protocol reports whether model answers are stable across these forms. It also reports performance separately by language form. The goal is not to enforce identical behavior in every language form at all costs. The goal is to identify whether the model is systematically weaker when questions are asked in Arabic or code-switched forms that reflect clinical practice.

The protocol includes a local error review sample. Each site reviews a small stratified sample of model failures and assigns error categories. The categories include missed visual finding, overcalled finding, question misunderstanding, likely label ambiguity, image-quality limitation, device or anatomy confusion, and invalid response. These categories are shared only as aggregate counts. This step is important because numerical metrics alone cannot explain why a site performs poorly. A site with many image-quality limitations requires a different response from a site with many question-understanding errors.

The protocol is designed for audit rather than deployment authorization. It gives decision-makers a structured view of model behavior. Whether a model should be used clinically depends on intended use, risk, oversight, and local regulation. The audit can support that decision, but it does not replace clinical governance.

4. Experimental Results

The centralized evaluation gives the most optimistic view. When all test cases are pooled, Configuration A reaches 81.6% accuracy and 78.9% balanced accuracy. Configuration B reaches 82.4% accuracy and 80.1% balanced accuracy. Configuration C reaches 82.1% accuracy and 80.7% balanced accuracy. If only regional balanced accuracy were considered, Configuration C would appear slightly strongest, followed closely by Configuration B. The differences are modest.

Single-site evaluation is unstable. If the model is judged using only the strongest Gulf tertiary node, Configuration B appears highly reliable, with balanced accuracy above 86%. If judged using the weakest North African re-

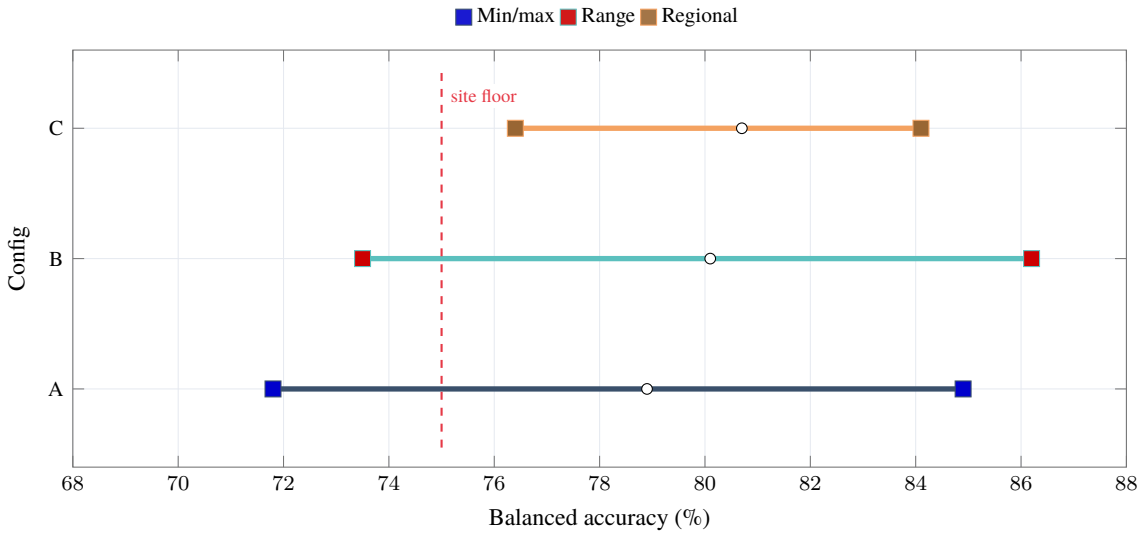

Figure 3: Site-level ranges show that Configuration C sacrifices the highest peak but raises the weakest-node floor, changing the governance choice under federated audit rules.

ferral node, the same configuration falls below 74%. The ranking of configurations also changes by site [12]. Some nodes favor bilingual prompting, while others favor the clinically formal prompt. This demonstrates why single-site evaluation is not a sufficient basis for regional claims.

Pooled-site reporting without consistency rules improves on single-site evaluation but still hides important weaknesses. Configuration B has the highest overall accuracy, but it shows a large performance gap between English and Arabic prompts in several nodes. Configuration C has slightly lower overall accuracy but smaller site-to-site variation. Pooled reporting would likely select Configuration B because its average is higher. The federated audit protocol selects Configuration C because it satisfies more site-consistency requirements [13].

The site-level balanced accuracy range is the key difference. Configuration A ranges from 71.8% to 84.9% across nodes. Configuration B ranges from 73.5% to 86.2%. Configuration C ranges from 76.4% to 84.1%. Configuration C has a lower maximum but a higher minimum. For regional governance, the higher minimum is important. A model that performs very well in a few nodes and weakly in others may be less acceptable than a model with slightly lower average performance but more stable behavior.

Rare thoracic findings show the largest node gap. For pneumothorax and small focal opacity, the best node reaches 84.7% sensitivity under Configuration B, while the weakest node reaches 71.3%. The gap is 13.4 percentage points. Configuration C reduces this gap to 8.2 points, though it does not eliminate it. Local error review suggests that the weak nodes contain more portable radiographs, lower exposure quality, and referral cases with subtle findings. This weakness would be difficult to see in pooled reporting because larger nodes contribute many easier cases.

Language-form effects are also substantial. Under direct English questions, Configuration B performs well [14]. Under Arabic questions, performance drops in several nodes. Under mixed Arabic-English questions, performance depends strongly on the abbreviation and local term structure. Across the network, the gap between English and Arabic-English variants is 9.1 percentage points for Configuration A, 7.4 points for Configuration B, and 4.8 points for Configuration C. The clinically formal prompt appears to reduce language-form instability, even though it does not produce the highest pooled accuracy.

Image-quality strata reveal another hidden issue. For high-quality images, all configurations perform similarly. For lower-quality portable chest radiographs and low-contrast ultrasound images, Configuration C is more conservative and produces fewer false positives. Configuration B has higher sensitivity but also overcalls several findings in low-quality images. Depending on intended use, either behavior may be preferred. The federated audit does not declare one universally better; it exposes the tradeoff.

Retinal imaging shows a different pattern. The main weakness is not language form but site-level prevalence and camera variation. Nodes with national screening-style retinal images show stronger performance than nodes

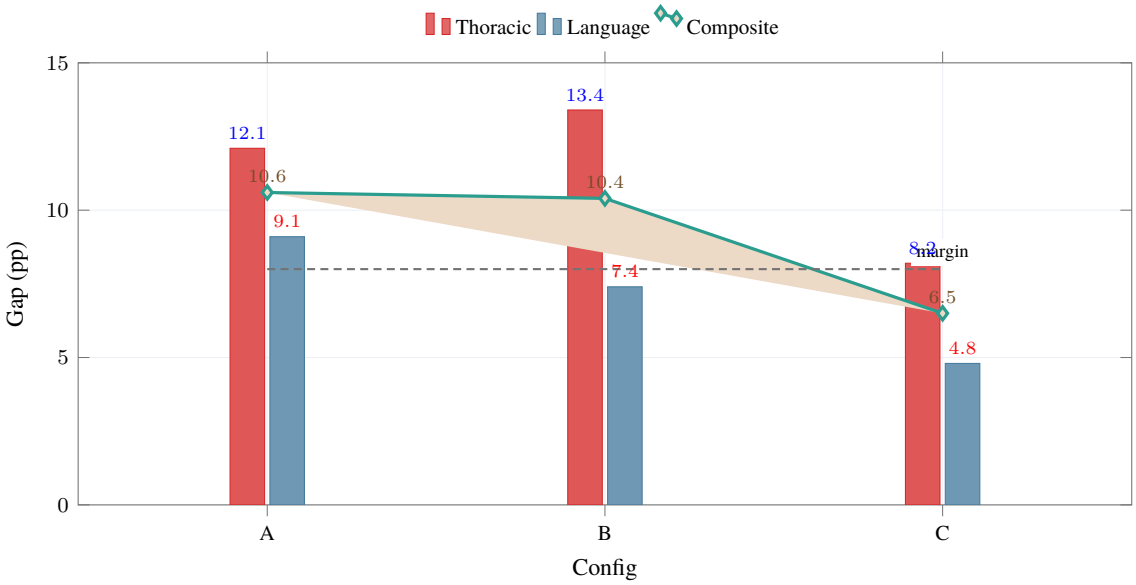

Figure 4: Rare thoracic findings and Arabic-English variants expose the largest instability, with the clinically formal configuration reducing both gaps enough to satisfy audit margins.

with referral retinal images. Configuration differences are smaller than in chest radiography. This suggests that the source of heterogeneity differs by domain. A single regional average cannot represent these differences.

Abdominal ultrasound is the most variable domain. Gallstone detection is relatively stable, but hydronephrosis and biliary dilatation vary across nodes. Local review suggests differences in image plane selection and label conventions. Some sites label mild dilatation more aggressively than others. Since the model receives still images rather than full ultrasound sweeps, performance depends heavily on whether the selected frame contains the relevant evidence. The audit identifies this limitation clearly.

Musculoskeletal radiography shows strong performance for obvious fractures and hardware but weaker performance for subtle periosteal reaction and soft tissue swelling. The site gap is related to case mix. Trauma-heavy nodes contain more acute fractures, while referral nodes include complex postoperative cases. Configuration C again has the most stable performance, although Configuration B has slightly better pooled accuracy.

The federated approval decision differs from centralized model selection. Centralized evaluation would choose Configuration C narrowly or Configuration B depending on the selected primary metric. Pooled-site reporting would favor Configuration B. The federated audit selects Configuration C because it is the only configuration meeting the site-level floor, language-equity margin, and rare-finding safety floor simultaneously. This is the central result of the paper: the evaluation framework changes the governance conclusion.

5. Site Variation and Error Review

The local error reviews show that not all performance gaps have the same cause. In the Gulf tertiary nodes, the largest error group is subtle finding miss, especially in chest radiography and ultrasound. These sites have relatively high image quality, so errors are often clinically difficult rather than technical. In the Levant nodes, question misunderstanding appears more often in Arabic and mixed prompts, especially when colloquial phrasing is used. In the North African nodes, image-quality limitation and label ambiguity appear more frequently, particularly in portable radiographs and ultrasound frames. In private network nodes, device and anatomy confusion is more common because the case mix includes many postoperative and follow-up images.

These patterns matter for remediation. A language-related error suggests prompt redesign or multilingual model testing. An image-quality error suggests acquisition-aware evaluation. A label ambiguity error suggests local annotation review. A subtle visual miss suggests model improvement or narrower intended use. Federated evaluation is useful because it does not collapse all errors into one score.

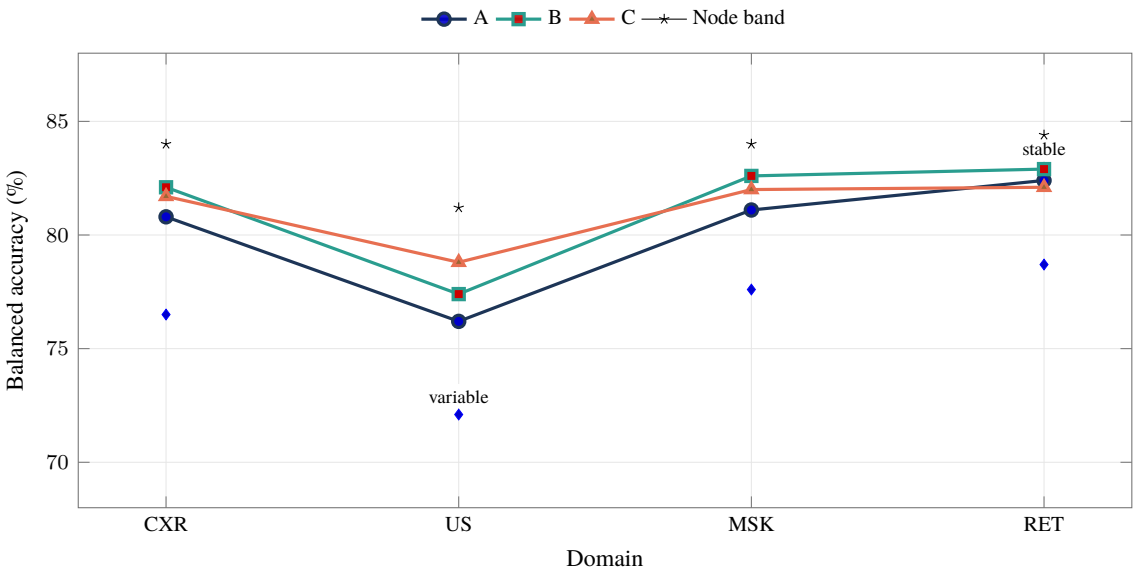

Figure 5: Domain-level traces show ultrasound as the most unstable modality and retinal imaging as comparatively stable, while configuration differences remain largest in radiography-heavy settings.

The audit identifies four recurring site-level risk patterns:

- nodes with high overall accuracy but weak rare-finding sensitivity;
- nodes with acceptable English performance but weaker Arabic or mixed-language performance;
- nodes with strong radiography results but unstable ultrasound results;
- nodes with low invalid-answer rates overall but elevated invalid responses in Arabic prompts.

These patterns are actionable. They indicate where further validation is needed before deployment. They also suggest that a model might be suitable for one domain within a hospital but not another.

The strongest example concerns Arabic-English mixed prompts. In several nodes, clinicians commonly preserve English abbreviations inside Arabic syntax [15]. The model handles some abbreviations well, such as CT and CXR, but is less stable with finding abbreviations and device terms. The issue is not simply Arabic translation. It is the combination of Arabic syntax, English shorthand, and local clinical habit. The federated protocol does not try to fix the language. It reports how much the model is affected by it.

Another important pattern concerns prevalence [16]. Some nodes have more positive examples for specific findings because they are referral centers. Other nodes have many screening or routine cases. A model with strong performance in a high-prevalence referral node may not have the same false-positive behavior in a low-prevalence screening node. Conversely, a model tuned to avoid false positives in a screening setting may miss positives in a referral setting. The audit report separates sensitivity and specificity by node to make this tradeoff visible.

The site-level confidence intervals also matter. Smaller nodes sometimes have wider intervals, especially for rare findings. A low point estimate in a small node should not be interpreted the same way as a low estimate in a large node. The protocol therefore combines point estimates with uncertainty and requires additional review when a node falls below a threshold with wide uncertainty. The goal is not to punish small sites but to avoid unjustified confidence.

The local error review also reveals that model invalid responses are not evenly distributed. Most invalid responses occur in Arabic prompts that include longer phrasing or compound questions. The answer parser marks these as invalid when the model gives an explanation without a clear yes or no [17]. Configuration C reduces invalid responses by using a stricter instruction. This partly explains its better site consistency. A model that answers fluently but not in the required format creates evaluation and workflow problems.

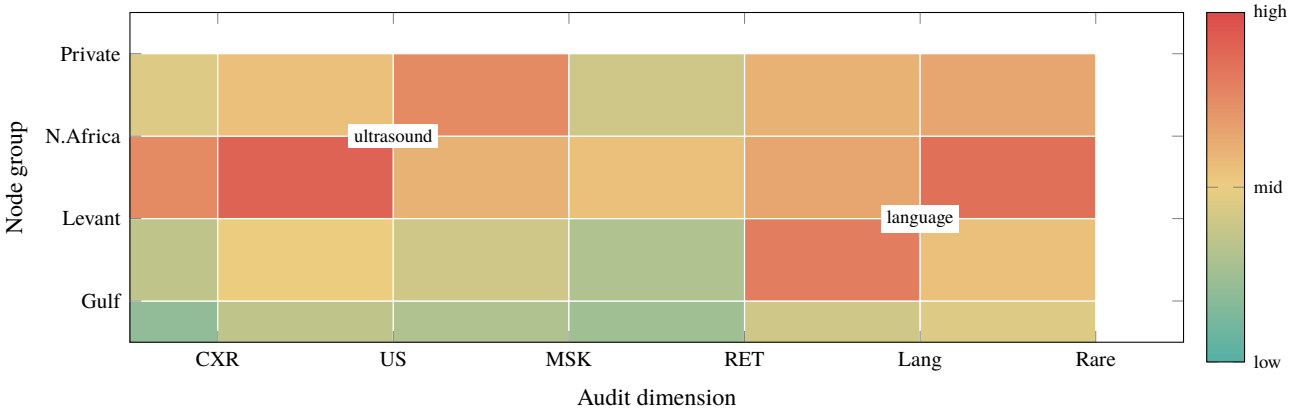

Figure 6: Grouped audit risks reveal different failure modes across regional hospital types, separating image-quality, language-form, and rare-finding weaknesses instead of compressing them into one pooled score.

6. Governance Implications

The main governance implication is that regional model approval should not be based only on pooled performance. A pooled score can be dominated by larger hospitals, easier cases, or better-resourced imaging environments. If a model is intended for use across a network, weaker nodes matter. A federated audit allows those nodes to contribute evidence without surrendering data custody.

The second implication is that data residency constraints do not prevent meaningful evaluation. They prevent certain forms of centralized analysis, but they do not prevent standardized local inference, aggregate reporting, or site-level governance. This distinction is important. Hospitals may decline participation if evaluation requires exporting images. They may participate if the evaluation package runs locally and produces auditable summaries. Federated evaluation can therefore expand the range of institutions included in model assessment.

The third implication is that language equity should be treated as a formal evaluation dimension. In Middle Eastern hospitals, English-only evaluation is insufficient. Arabic and mixed-language prompts should be tested directly. If a model performs well in English but less reliably in Arabic, the audit should show that. The solution might be prompt restriction, multilingual tuning, interface design, or local validation. The first step is measurement.

The fourth implication is that model selection should consider stability. A configuration with slightly lower average accuracy may be preferable if it avoids severe site-level failures. This is a common issue in clinical governance. Average performance is not the only value. Predictability across settings also matters. The proposed audit protocol makes that value explicit through site floors and subgroup margins.

The fifth implication is that hospitals should receive local reports, not only regional summaries. A regional coordinating body may need an overall decision, but each hospital needs to understand its own risk profile. A site with weak ultrasound performance may restrict use to radiography. A site with Arabic prompt instability may require English templates until further validation. A site with rare-finding weakness may require human review for those targets. Federated evaluation supports this local tailoring.

The sixth implication concerns accountability. If a model is evaluated only centrally, local hospitals may not know whether the evidence applies to them. If each hospital evaluates alone, the region loses the ability to compare behavior and identify systematic weaknesses. Federated audit creates shared accountability while preserving local control. It also creates a common language for discussing model limitations.

There are practical requirements. Hospitals need enough local cases, consistent labels, technical ability to run the evaluation package, and agreement on reporting definitions. These requirements are not trivial. Smaller hospitals may need support. Some sites may lack adequate labels for certain findings. The protocol should allow partial participation rather than forcing unreliable reporting. A transparent missingness report is better than low-quality metrics.

7. Limitations and Future Work

The first limitation is that the study is simulated rather than a live regional consortium. The hospital nodes are

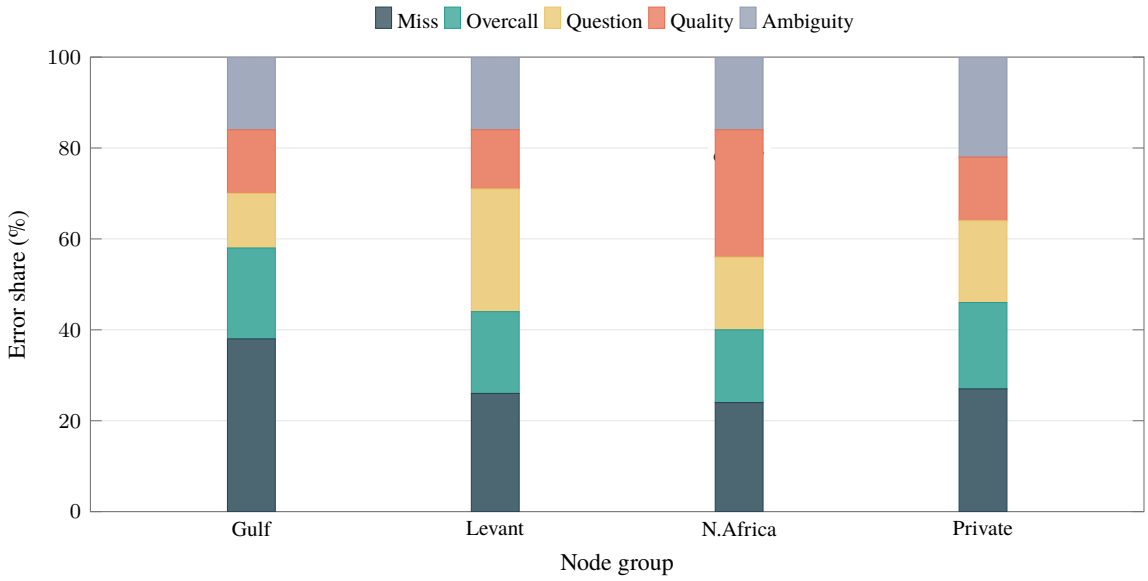

Figure 7: Local error review decomposes failure causes by node group, showing that remediation would differ across sites even when aggregate accuracy appears similar.

designed to represent plausible Middle Eastern variation, but real institutions would introduce additional complexity [18]. Data access, ethics review, information technology constraints, and local clinical priorities would affect participation. A prospective multi-hospital pilot is needed.

The second limitation is the binary question format. Binary VQA is useful for controlled evaluation, but clinical imaging decisions often require severity, location, comparison, and narrative explanation [19]. Federated evaluation should be extended to structured reports and multi-label outputs. That extension will require more complex metrics and stronger local review [20].

The third limitation is that the study evaluates a fixed model rather than local adaptation. In practice, a weak site might adapt the model or adjust prompts after audit. The present design intentionally separates evaluation from improvement. Future work should study federated evaluation followed by site-specific remediation and re-audit.

The fourth limitation is language coverage. The study includes English, Arabic, and mixed Arabic-English forms, but Middle Eastern hospitals are multilingual beyond these categories. Urdu, Hindi, Persian, Turkish, French, and other languages may appear in clinical communication depending on country and workforce [21]. Future evaluations should expand language coverage where relevant.

The fifth limitation is label heterogeneity. Local labels may differ because of reporting style, diagnostic thresholds, or available follow-up. The protocol includes error review and uncertainty intervals, but it cannot fully solve label inconsistency. Future federated audits should include label harmonization exercises and targeted adjudication for disputed findings [22].

The sixth limitation is that aggregate reporting can still hide some case-level failures. Federated evaluation protects data custody, but it restricts central analysis. One possible extension is privacy-preserving case review, where sites share de-identified error exemplars only after local approval. Another is secure enclave review or federated analytics with stronger privacy guarantees. These approaches require technical and legal development.

The seventh limitation is operational burden [23]. Running local evaluation packages and reviewing errors takes time. Hospitals with limited research infrastructure may find participation difficult. A regional evaluation network would need funding, technical support, and clear governance. The protocol is feasible, but feasibility is not automatic.

Future work should implement a real federated evaluation pilot across hospitals in at least three Middle Eastern countries. Such a pilot should test legal agreements, local execution, reporting templates, and governance interpretation [24]. It should also study how clinicians use the resulting reports. A technically sound audit is useful only if decision-makers understand and act on it [25].

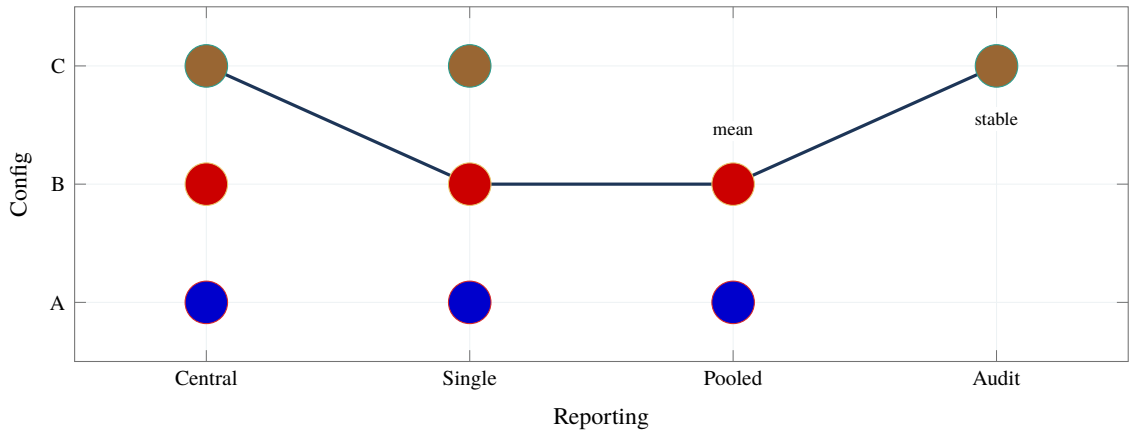

Figure 8: Model selection changes when site floors, rare-finding sensitivity, and language-equity margins are enforced, moving the final decision from pooled mean performance toward stable regional behavior.

Future work should also examine dynamic monitoring. A model that passes a federated audit at one time may drift as imaging equipment, patient populations, or clinical language changes. Periodic federated re-evaluation could identify when a model needs updating [26]. This is especially relevant in fast-growing health systems where hospital workflows change rapidly.

Another direction is patient-group analysis. The Middle East includes citizens, long-term residents, migrant workers, refugees, and medical tourists. Disease prevalence and imaging pathways may differ across groups. A responsible evaluation should examine whether model behavior differs across clinically relevant populations, while protecting privacy and avoiding stigmatizing reporting. Federated methods may help because sensitive subgroup analyses can remain local.

Finally, future work should connect federated evaluation with procurement. Hospitals and ministries increasingly need evidence before adopting AI tools. A federated audit report could become part of procurement requirements, ensuring that vendors demonstrate performance across local sites and languages rather than presenting only external benchmark results. This would make evaluation more relevant to actual use.

8. Conclusion

Our study presented a federated evaluation framework for medical vision-language binary question answering across Middle Eastern hospital data residency settings. The study focused on a practical constraint: many hospitals cannot centralize clinical images, yet regional governance still requires evidence about how a model behaves across institutions [27].

The empirical simulation showed that centralized and pooled reporting can obscure site-level weaknesses. A model configuration with strong average accuracy may still perform poorly in certain hospital nodes, rare findings, image-quality strata, or Arabic-English question forms. The proposed federated audit protocol exposes these differences through standardized local evaluation, aggregate reporting, site-level thresholds, and subgroup analysis.

The Middle Eastern setting highlights the importance of institutional and linguistic variation. English, Arabic, and mixed clinical language forms can produce different model behavior. Imaging practice and patient mix also vary across hospitals. Evaluation methods that ignore this variation risk approving models on evidence that does not generalize to all intended sites.

Federated evaluation is not a complete safety solution. It depends on local label quality, shared metric definitions, adequate participation, and meaningful governance thresholds. It cannot replace clinical review or prospective validation. Its value lies in making distributed evidence usable when raw data cannot be pooled.

The central conclusion is that medical vision-language model evaluation should match the governance reality of the health system. In Middle Eastern hospital networks, where data residency, multilingual practice, and institutional diversity are central, federated audit offers a practical path toward more transparent and locally accountable model assessment.

References

[1] Y. Song, M. Roy, M. Zhong, L. Chen, M. Lin, and R. Zhang, "Multimodal data fusion for whole-slide histopathology image classification.," *Journal of healthcare informatics research*, vol. 9, pp. 513–532, 9 2025.

[2] R. J. Peritz, "Three visions of managed competition, 1920-1950," *The Antitrust Bulletin*, vol. 39, pp. 273–287, 3 1994.

[3] P. A. Newman-Casey, L. M. Niziol, A. R. Elam, A. K. Bicket, O. Killeen, D. John, S. D. Wood, D. C. Musch, J. Zhang, L. Johnson, M. Kershaw, and M. A. Woodward, "Michigan screening and intervention for glaucoma and eye health through telemedicine program: First-year outcomes and implementation costs.," *American journal of ophthalmology*, vol. 251, pp. 43–51, 3 2023.

[4] M. B. Gillingham, D. Choi, A. Gregor, N. Wongchaisuwat, D. Black, H. L. Scanga, K. K. Nischal, J.-A. Sahel, G. Arnold, J. Vockley, C. O. Harding, and M. E. Pennesi, "Early diagnosis and treatment by newborn screening (nbs) or family history is associated with improved visual outcomes for long-chain 3-hydroxyacylcoa dehydrogenase deficiency (lchadd) chorioretinopathy.," *Journal of inherited metabolic disease*, vol. 47, pp. 746–756, 4 2024.

[5] B. Sadanandan and V. Behzadan, "Mechanistically guided lora improves paraphrase consistency in medical vision-language models," *arXiv preprint arXiv:2603.00148*, 2026.

[6] V. Lam, A. Parida, S. Dance, S. Tabaie, K. Cleary, and S. M. Anwar, "Analyzing pediatric forearm x-rays for fracture analysis using machine learning.," *International journal of computer assisted radiology and surgery*, vol. 20, pp. 1845–1850, 7 2025.

[7] I. Blundell, R. Brette, T. A. Cleland, T. G. Close, D. Coca, A. P. Davison, S. Diaz-Pier, C. F. Musoles, P. Gleeson, D. F. M. Goodman, M. L. Hines, M. Hopkins, P. Kumbhar, D. Lester, B. Marin, A. Morrison, E. Müller, T. Nowotny, A. Peyser, D. Plotnikov, P. Richmond, A. Rowley, B. Rumpe, M. Stimberg, A. B. Stokes, A. Tomkins, G. Trensch, M. M. Woodman, and J. M. Eppler, "Code generation in computational neuroscience: a review of tools and techniques," *Frontiers in neuroinformatics*, vol. 12, pp. 68–68, 11 2018.

[8] F. Ahmed, N. S. Naz, S. Khan, A. U. Rehman, W. M. Ismael, and M. A. Khan, "Explainable artificial intelligence (xai) in medical imaging: a systematic review of techniques, applications, and challenges.," *BMC medical imaging*, vol. 26, pp. 37–37, 1 2026.

[9] E. CIS, "Computer and information science, vol. 1, no. 4, november 2008, all in one file," *Computer and Information Science*, vol. 1, 10 2008.

[10] M. Groner, R. Groner, R. Müri, K. Koga, S. Raess, and P. Suri, "Abstracts of the 13th european conference on eye movements 2005," *Journal of Eye Movement Research*, vol. 1, pp. 135–, 8 2005.

[11] J. Gao, N. Fang, and Y. Xu, "Application of artificial intelligence in retinopathy of prematurity from 2010 to 2023: A bibliometric analysis.," *Health science reports*, vol. 8, pp. e70718–e70718, 4 2025.

[12] M. G. Crowson, D. Moukheiber, A. R. Arévalo, B. D. Lam, S. Mantena, A. Rana, D. Goss, D. W. Bates, and L. A. Celi, "A systematic review of federated learning applications for biomedical data.," *PLOS digital health*, vol. 1, pp. e0000033–e0000033, 5 2022.

[13] D. Haykal, F. Flament, M. Shadev, P. Mora, C. Puyat, B. Dréno, Q. Zheng, H. Cartier, M. Gold, and S. Cohen, "Advances in longevity: The intersection of regenerative medicine and cosmetic dermatology.," *Journal of cosmetic dermatology*, vol. 24, pp. e70356–e70356, 7 2025.

[14] S. Shah, X. Luo, S. Kanakasabai, R. Tuason, and G. Klopper, "Neural networks for mining the associations between diseases and symptoms in clinical notes.," *Health information science and systems*, vol. 7, pp. 1–9, 11 2018.

[15] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical image analysis*, vol. 24, pp. 205–219, 7 2015.

[16] E. McNeilly, "Redesigning an undergraduate russian language unit: Reflections on aligning departmental pedagogic objectives with the new teaching programme and assessment," *Education & Pedagogy Journal*, pp. 45–56, 12 2021.

[17] I. Wood, T. Bhojak, Y. Jia, E. Jacobsen, B. E. Snitz, C.-C. H. Chang, and M. Ganguli, "Predictors of driving cessation in older adults: A 12-year population-based study.," *Alzheimer disease and associated disorders*, vol. 37, pp. 13–19, 1 2023.

[18] N. Shah, A. Castellanos, Y. Chen, J. Piette, A. Bucher, and S. Murphy, "A scoping review on artificial intelligence-supported interventions for non-pharmacologic management of chronic rheumatic diseases.," *Arthritis care & research*, vol. 78, pp. 270–281, 11 2025.

[19] X. Pei and Z. Li, "Narrative review of comprehensive management strategies for diabetic retinopathy: interdisciplinary approaches and future perspectives.," *BMJ public health*, vol. 3, pp. e001353–e001353, 1 2025.

[20] L. E. Jones, G. Eliason, S. Gupta, E. G. Jeong, A. Besemer, D. A. Sterling, and J. M. Fagerstrom, "Primer on disability: Why accessibility is important for all medical physicists.," *Journal of applied clinical medical physics*, vol. 26, pp. e70003–e70003, 2 2025.

[21] J. Jalili, J. Huynh, E. Walker, B. G. Chuter, C. Bowd, A. Heinke, A. Belghith, M. H. Goldbaum, M. A. Fazio, C. A. Girkin, C. G. D. Moraes, J. M. Liebmann, S. L. Baxter, R. N. Weinreb, L. M. Zangwill, and M. Christopher, "Performance of general-purpose vision language models and ophthalmology foundation models in glaucoma detection and function prediction.," *Translational vision science & technology*, vol. 14, pp. 31–31, 11 2025.

[22] A. Karan, D. J. Campbell, and H. R. Mayer, "The effect of a visual aid on the comprehension of cataract surgery in a rural, indigent south indian population.," *Digital journal of ophthalmology : DJO*, vol. 17, pp. 16–22, 9 2011.

[23] J. Zhang, Y. Pan, H. Lin, Z. Sun, P. Wu, and J. Tu, "Infodemic: Challenges and solutions in topic discovery and data process.," *Archives of public health = Archives belges de sante publique*, vol. 81, pp. 166–166, 9 2023.

[24] A. Heidari, N. J. Navimipour, M. Unal, and S. Toumaj, "The covid-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions.," *Computers in biology and medicine*, vol. 141, pp. 105141–105141, 12 2021.

[25] L. K. Vora, A. D. Gholap, K. Jetha, R. R. S. Thakur, H. K. Solanki, and V. P. Chavda, "Artificial intelligence in pharmaceutical technology and drug delivery design.," *Pharmaceutics*, vol. 15, pp. 1916–1916, 7 2023.

[26] J. Egger, C. Gsaxner, G. Luijten, J. Chen, X. Chen, J. Bian, J. Kleesiek, and B. Puladi, "Is the apple vision pro the ultimate display? a first perspective and survey on entering the wonderland of precision medicine.," *JMIR serious games*, vol. 12, pp. e52785–e52785, 9 2024.

[27] K. Curyto, K. V. Haitsma, and G. L. Towsley, "Cognitive interviewing: Revising the preferences for everyday living inventory for use in the nursing home.," *Research in gerontological nursing*, vol. 9, pp. 24–34, 5 2015.