Viewpoint

A Pragmatic Framework for Federated Learning Risk and Governance in Academic Medical Centers

Daniel Bottomly[1], MS; Bridget Barnes[2], MBA, PhD; Kuli Mavuwa[2], JD; Nikki Lee[2], JD; Holger R Roth[3], PhD; Chester Chen[3], PhD; Shannon K McWeeney[1], PhD

[1]OHSU Knight Cancer Institute, Portland, OR, United States
[2]Office of Information, Privacy and Security, Information Technology Group, Oregon Health & Science University, Portland, OR, United States
[3]NVIDIA, Santa Clara, CA, United States

Corresponding Author:

Shannon K McWeeney, PhD
OHSU Knight Cancer Institute
3485 S Bond Ave
Portland, OR 97239
United States
Phone: 1 503-494-8311
Email: mcweeney@ohsu.edu

Abstract

With the rapid development of artificial intelligence (AI), particularly large language models, there is growing interest in adopting AI approaches within academic medical centers (AMCs). However, the vast amounts of data required for AI and the sensitive nature of medical information pose significant challenges to developing high-performing models at individual institutions. Furthermore, recent changes in government funding priorities may result in the decentralization of biomedical data repositories that risk creating significant barriers to effective data sharing and robust model development. This has generated significant interest in federated learning (FL), which enables collaborative model training without transferring data between institutions, thereby enhancing the protection of proprietary and sensitive information. While FL offers a crucial pathway to enable multi-institutional AI development while maintaining data privacy, it also exposes AMCs to novel governance, security, and operational risks that are not fully addressed by existing procedures. In response, this manuscript provides a perspective grounded in both leading international standards (NIST AI RMF [National Institute of Standards and Technology Artificial Intelligence Risk Management Framework], International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) 42001) and in the real-world governance experience of AMC leadership. We present a risk differentiation framework, an FL risk matrix, and a set of essential governance artifacts—each mapped to key institutional challenges and reviewed for alignment with core standards but offered as pragmatic, illustrative guides rather than prescriptive checklists. Together, these tools represent a novel resource to support AMC security, privacy, and governance leaders with standards-informed, context-sensitive tools for addressing the evolving risks of FL in biomedical research and clinical environments.

JMIR AI 2026;5:e80022; doi: 10.2196/80022

Keywords: federated learning; academic medical centers; governance; security and privacy; artificial intelligence

Introduction

Following the groundbreaking work in artificial intelligence (AI), especially with respect to large language models, there has been an exponential rise in AI model adoption across academic medical centers (AMCs) [1,2] . However, currently, AI governance across AMCs is highly variable [3,4] as is the maturity of information, privacy, and security oversight with respect to AI. Deploying AI effectively in AMCs often necessitates leveraging sensitive patient data for external vendor partnerships or collaborative foundational model training. However, the level of understanding about alternative approaches and perceived risk leads to “over-restrictive” policies that impact data sharing and collaboration. Federated learning (FL), a decentralized approach to training AI or other machine learning (ML) models, has been proposed as a solution for collaboration across AMCs and with industry partners [5]. While FL ensures that sensitive data remains within institutional boundaries and is not transferred between organizations, AMCs must still address significant security and privacy considerations. Even though raw data does not move, model updates exchanged during FL can potentially leak information or be vulnerable to adversarial attacks, and issues of trust and compliance remain critical. The motivation for this work is to provide AMCs with practical guidance and a structured framework for evaluating and mitigating these unique risks, supporting secure and privacy-preserving collaborative AI development in health care settings.

FL is an approach where multiple sites collaborate to jointly train a ML model [6]. Originally, FL was devised as a solution to data privacy issues by not requiring institutions or sites to share their data directly. Benchmark studies have shown its potential effectiveness to bolster the efficacy of model development in biomedical research as reviewed previously [7]. FL can generally be divided into 3 main types [8]: (1) vertical, where the sites share the same samples but have different features, (2) horizontal, where the sites share comparable features but on different samples, and (3) federated transfer learning, which is used when sites have both different samples and different features, allowing them to collaboratively improve models by leveraging transfer learning techniques, even in the absence of significant overlap in data. For the rest of the manuscript, we focus on horizontal FL as that is more likely to be the most common scenario for an AMC. For instance, a horizontal FL use case would be genomics, imaging, and electronic health record data collected from patients with similar diagnoses across different hospitals or catchment areas.

Although, in principle, allowing institutions or sites to control their own data improves privacy, the process of model training using FL has been shown to be susceptible to attacks affecting model performance (training or inference) as well as privacy attacks; for a recent survey, see Rodríguez-Barroso et al [9] and references within. It is important to note that the security and robustness of an FL system depends not only on the security of individual clients but also on the strength of the aggregation protocol and the defenses in place against adversarial behavior. While a compromised client can pose significant risks, properly designed FL systems can mitigate their impact. With regard to the risks if a client is compromised, they include not only the client's local data but also the data of other clients through model or gradient inversion attacks. A compromised client can also be used to influence the global model through the manipulation of the local data (data poisoning), interference with the local model training (model poisoning), or degradation of the overall training process through so-called Byzantine attacks [10,11]. In addition to the clients, the FL server itself must be protected and monitored, especially for high-risk data. Although no data resides on the server, access to the global model and subsequent updates provides an attacker the means to reconstruct or infer client data. For more information, these attacks and possible mitigations in an FL context have been thoroughly reviewed [12].

Within an AMC from a cybersecurity perspective, a "bad actor" would be an individual with malicious intent toward the data or model. However, we recognize that this is a low-probability scenario. The more likely concern would be at the organizational level when there is a federation with external partners.

Recent efforts to formalize AI governance, including National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (NIST AI RMF) and International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC; ISO/IEC) 42001, offer foundational structures for risk identification, evaluation, and organizational oversight of advanced AI systems in health care. However, while these standards address key technical and ethical dimensions, they do not fully capture the unique challenges AMCs face when implementing FL initiatives. Specifically, FL introduces cross-institutional accountability gaps, complex role-based risks, new vectors for privacy leakage, and ambiguity in operationalizing risk stratification for projects involving sensitive data or novel model architectures.

For AMCs, these governance gaps manifest in several areas: (1) lack of shared accountability and decision-making models across partnered institutions, (2) difficulties integrating local policy and national or international standards, especially when risk is jointly determined by data sensitivity, model complexity, and evolving clinical context, (3) ambiguity in managing platform-specific risks (eg, role privilege and authorization configuration) and documenting privacy or security practices that span organizations, and (4) insufficient practical guidance for balancing innovation with patient safety, ethical considerations, and regulatory compliance. Our framework aims to address these gaps by combining standards-informed principles with lessons learned from direct AMC experience, offering practical, context-sensitive tools to support cross-functional governance for FL.

As part of this manuscript, we lay out the information, privacy, and security risks for an AMC associated with the use of FL. There are currently a diverse array of FL platforms available (Table S1 in Multimedia Appendix 1). We will illustrate the framework using the NVIDIA FLARE (NVIDIA Federated Learning Application Runtime Environment) [13] platform, as this was the solution used for the first year of the National Cancer Institute FL network prototype [14]. NVIDIA FLARE is NVIDIA's full-featured open-source FL software development kit [13]. It can support a variety of ML models, including neural networks (colloquially referred to as AI), and is platform-agnostic so that models can be easily migrated to the federated setting. Importantly, it supports the implementation of privacy and security methods like differential privacy (DP) and homomorphic encryption (HE). Although we discuss specific risks related to roles and artifacts for NVIDIA FLARE, the topics and general recommendations can be applied to any FL platform. This information is expected to be relevant to leaders involved in security, privacy, and IT (eg, Chief Information Security Officer [CISO], Chief Privacy Officer, and Chief Information Officer), as well as those involved in data and AI governance at AMCs. Drawing on our experience with the National Cancer Institute FL network initial efforts, we bridge the gap between existing standards and their real-world implementation by providing a practical mapping that has so far been missing.

The development of the proposed FL risk matrix in this manuscript was explicitly guided by international standards and best practices in AI risk management and governance. We combined the flexible, risk-based approach of the NIST AI RMF [15] with the structured, process-driven requirements of ISO/IEC 42001 [16] to create practical tools for oversight in AMC FL contexts.

Proposed FL Risk Matrix

Our goal was to develop a systematic, transparent approach for managing the complex risks associated with FL in AMCs that leverages established AMC procedures for risk determination—evaluating both data and model risk based on defined criteria (Table 1)—as the critical first step in the review process. This table summarizes critical factors influencing risk: data sensitivity (including regulatory scope), model complexity, and operational context. This was based on both "Map" and "Measure" functions of the NIST AI RMF and the risk assessment procedures in ISO/IEC 42001 Clause 6.1. These standards emphasize identifying, categorizing, and proportionally responding to diverse risk dimensions, including legal and operational impact, which are especially salient in AMC settings. We note that risk determination (using Table 1) is conducted collaboratively by the Artificial Intelligence Governance Committee (AIGC) and the CISO, integrating technical, ethical, and operational expertise to ensure robust, context-sensitive evaluation from the outset. Once risk levels are assigned, the FL matrix (Table 2) directly guides the appropriate level of artificial intelligence governance review (AIGR) and security review (SecR): projects classified as high risk undergo comprehensive, multilevel oversight, while those identified as low risk are eligible for expedited, streamlined review. This design follows the "Manage" function of the NIST AI RMF and ISO/IEC 42001 Clauses 8 and 9, which require that risk controls and performance evaluations are matched proportionately to risk profiles determined in earlier steps. Through this, the matrix ensures that high-risk FL initiatives undergo the strictest oversight, in line with international guidance. To facilitate this review process, we have also identified key artifacts such as essential documentation, authorization files, and privacy configuration files, which are pivotal for transparency, audit readiness, and reliable governance (Textbox 1). These requirements map to the NIST AI RMF's recommendations for rigorous documentation and continuous oversight and ISO/IEC 42001 Clause 7.5, which details retention of documented information as evidence of compliance and governance efficiency. Together, these tables embody leading principles from NIST AI RMF and ISO/IEC 42001, adapted to the practical realities of AMC FL. They are intended as conceptual and illustrative guides, framing rigorous governance conversations, rather than serving as prescriptive or validated assessment instruments. By aligning review pathways with the specific risk profile of each project, the matrix ensures that institutional resources are efficiently allocated, enabling responsible innovation while maintaining rigorous compliance and data protection standards.

Building on the risk framework, the following sections delve into the specific governance and security mechanisms that ensure context-based, responsible, and trustworthy implementation of FL within AMCs, as well as two illustrative examples.

Table 1. Key differentiators between high and low risks for data and models.

Category and factor	High risk	Low risk
Data		
Data sensitivity	PHI[a], genomic, rare conditions, regulatory considerations (HIPAA[b], GDPR[c], EO[d] 14117)	Aggregated, synthetic, logs, general research data protection(s)
Model		
Model complexity	LLMs[e], GANs[f], high-capacity CNNs[g]	Linear models, decision trees
Operational context	Direct impact on patient care, diagnostics, or clinical decision support; real-time or near–real-time use; high stakes for errors, regulatory requirements (FDA SAMD[h], EU AI[i] Act)	Administrative, research, or quality improvement tasks; retrospective analysis; minimal patient safety impact

[a] PHI: protected health information.

[b] HIPAA: Health Insurance Portability and Accountability Act.

[c] GDPR: General Data Protection Regulation.

[d] EO: executive order.

[e] LLMs: large language models.

[f] GANs: generative adversarial networks.

[g] CNNs: convolutional neural networks.

[h] FDA SAMD: Food and Drug Administration Software as a Medical Device.

[i] EU AI: European Union's Artificial Intelligence Act.

Table 2. Proposed federated learning risk matrix.

Data risk	Model risk	AIGR[a,b]	SecR[c]
High	High	• Suitability • Generalizability • Bias/fairness • Memorization • Robustness to attack	• ProjectAdmin and OrgAdmin must be highly trusted and reviewed • RBAC[d] • Mandatory HE[e] and/or DP[f]
High	Low	• Memorization • Robustness to attack	• ProjectAdmin and OrgAdmin must be highly trusted and reviewed • RBAC
Low	High	• Suitability • Generalizability • Bias/fairness • Robustness to attack	• RBAC with additional scrutiny for Leads and OrgAdmins
Low	Low	• Expedited	• RBAC

[a] AIGR: artificial intelligence governance review.

[b] We are separating into AIGR and SecR sections, but in practice, as mentioned in the paper, these activities would be handled jointly.

[c] SecR: security review.

[d] RBAC: role-based access control following the least privilege principle.

[e] HE: homomorphic encryption.

[f] DP: differential privacy.

Textbox 1. Relevant artifacts for artificial intelligence (AI) governance review and security review.

The following are the artifacts for AI governance review and security review:

- Suitability
  - Computational notebooks
  - Model cards
  - Prior publications
- Pretraining evaluation and certification
  - Computational notebooks
- Roles and authorization
  - Role qualification document
  - Authorization.json
- Privacy and security of model updates
  - privacy.json
  - config_fed_server.conf
  - config_fed_client.conf
  - Computational notebooks
    - Provide data for choice of *eps* values if differential privacy is used

AI Governance Review

Following risk determination for both the models and data, the project advances to the AIGR. Recognizing the critical importance of FL, our approach assumes that robust AI governance structures are in place at AMCs [3,4]. Based on our FL risk matrix, we propose that for any project involving high-risk data and/or high-risk models, AIGR consists of 2 primary steps: (1) assessing model suitability for its intended use and (2) conducting pretraining evaluation and certification for FL. For suitability, comprehensive model documentation (such as Model Cards [17]) is required or, at minimum, evidence that the model architecture and training pipeline have been validated on publicly available or simulated data. It is essential to confirm that the model's intended use matches its validated context and that security, privacy, and ethical, legal, and social implications are explicitly addressed.

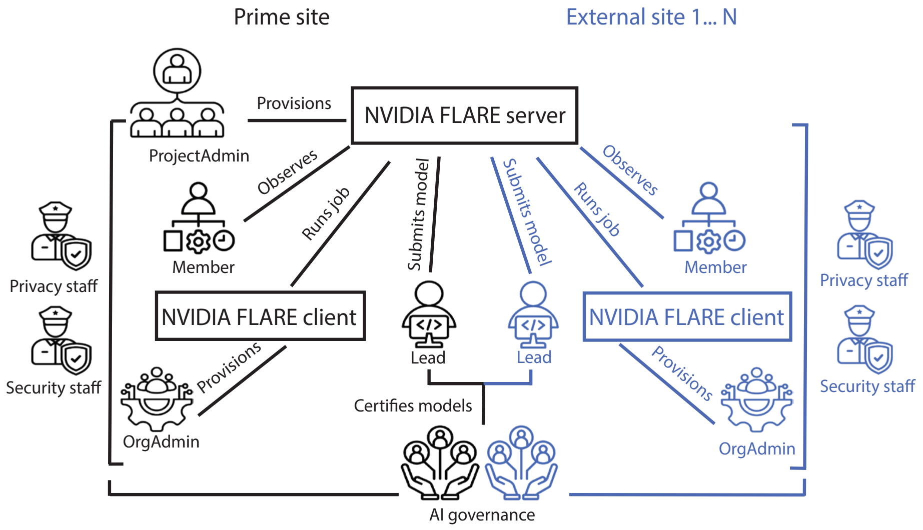
Pretraining evaluation and certification consists of 4 main components: generalizability, fairness and bias, memorization risk, and susceptibility to attack. With regard to generalizability, does the training pipeline maintain consistent performance across all datasets from participating sites? Are there any detectable issues with fairness or bias against certain groups, which can be quantified using several toolkits [18-20]? Note, this is separate from suitability above, which is focused on the concept of the predictive task—here we are addressing specifically the data to be used for this application. How likely is the model to memorize training data, especially rare instances? Memorization or the ability of an AI or ML approach (therefore not unique to FL) to recall specific training examples [21] can be measured after training using, for instance, the exposure metric for natural language data [22] or the *M* score as proposed in the context of medical imaging [23] as well as other metrics now being proposed [24]. We note that memorization can be mitigated using components in FL such as DP [25-28] (as discussed later). Finally, the most challenging assessment would be to determine the susceptibility of a model training pipeline to the wide range of possible attacks. This could be performed via red-teaming endeavors using NVIDIA FLARE's simulation software.

Once a model is certified and trained using FL, it undergoes posttraining assessments. These would be like the pretraining assessment but would be applied to the final trained model. A model passing these assessments could then be deployed. Monitoring of a deployed model would then proceed as part of normal ML operations [29] best practices. The collaborative involvement of the CISO and AIGC in this phase ensures that both technical and ethical standards are rigorously maintained and that the institution remains compliant with evolving regulatory and industry best practices.

Security Review

The SecR is completed in parallel with the AIGR and focuses on (1) user roles, (2) authorizations, and (3) the privacy and security of model updates. Within the NVIDIA FLARE platform, users can have at least one of 4 potential roles [30]—Project Admin, OrgAdmin, Lead, and Member—each with distinct responsibilities and associated risks (description in Table S2 in Multimedia Appendix 1, relationships between the roles in Figure 1). For high-risk data or models, individuals who will be assigned to critical roles must be highly trusted and reviewed (Table 2, impacts and recommendations for vetting described in Table S2 in Multimedia Appendix 1).

Figure 1. Relationships and responsibilities of entities and roles involved in federated learning with NVIDIA FLARE. In the creation of a federated learning (FL) network, there will be a "prime site" who provisions the network and provides guidance for other external sites (blue). For this discussion based on the National Cancer Institute F network, we assume the Prime site is also the site that is responsible for the federated model. Outside of the Project Admin role and the Artificial Intelligence Governance Committee (AIGC), each site will have representatives assuming one or more of the NVIDIA FLARE roles. The AIGC should have representatives from the sites and oversee all of the modeling work that occurs as part of the FL network as well as approving the FL framework itself. Each site will have security and privacy officers who will be in charge of ensuring the security of any high-risk data and/or models and compliance with privacy laws, with input from the AIGC. Shown as icons are the different roles in addition to the individuals involved in governance or oversight. Boxes indicate the computational resources such as the NVIDIA FLARE server and clients. Lines indicate the main relationships with accompanying text. AI: artificial intelligence; NVIDIA FLARE: NVIDIA Federated Learning Application Runtime Environment.



Complicating the relationships between the roles is governmental oversight. The Department of Justice's newly created data security program that is based on an executive order to prevent the sharing of US bulk data with Countries of Concern (CoCs) [31] was launched recently [32] (28 CFR Part 202). This recently escalated to barring CoCs from access to National Institutes of Health databases [33]. Human genomic data, a likely data type to be included in datasets used in FL by AMCs, is covered under "human omic data and associated biospecimens," one of the 6 categories of sensitive personal data. With FL, the data does not move. However, there is a potential vulnerability if the lead role is a "bad actor" or just careless and implements a high-capacity large language model that memorizes the data [34]. This could allow the recovery of the training data via targeted prompting [35]. While this is a clear privacy concern, there are additional ramifications, as the executive order impacts not just entities associated with CoCs but also individuals who reside in a CoC or are employed by entities within a given CoC. This highlights the importance of having a formal mechanism for reviewing the individuals granted each role, especially for higher risk data as well as the model itself.

NVIDIA FLARE provides a mechanism to fine-tune privileges within the roles through authorization policies, which allows each institution (site) to enforce their own access requirements. Authorization policies are carefully reviewed by the CISO, who works closely with the AIGC to ensure alignment between security and operational requirements. The general recommendation for projects of all risks is to follow the design principle of "least privilege" [36], that is, providing the minimal amount of access required to perform a given role with respect to both the model training process and the data. Restricted role authorization would be seen as necessary for any high-risk data (this would be unaffected by model risk), but application to both low- and high-risk datasets would not place an undue burden on the sites.

One area that is likely to be of contention for model training is support for BYOC or "bring-your-own-code." This is a good example of the difficulty balancing the responsibility of the Prime Site Lead (lead at site that developed model) to protect the model in terms of both performance (including fairness) as well as security and the responsibility of the external OrgAdmin(s) to protect the data or system of their site. Specifically, custom code from the Prime Site Lead can be perceived as a potential threat to both the external sites and hosted data by the OrgAdmins. When the Prime Site Lead needs to make an update, and BYOC is not enabled, they are dependent upon the OrgAdmin or other system administrators at the external sites to implement those changes. Importantly, a compromised client at the external site would have access to the code. This has several implications: first, any confidential code could be accessed, jeopardizing relevant intellectual property; additionally, the integrity of the code could be compromised. For instance, the code could be modified to attack the model, allowing data extraction or performance degradation.

Privacy policies implement additional protections and can be implemented separately for each site [37]. They are designed to thwart so-called feature-inference or reconstruction attacks [9], which attempt to recover the client's training data, given the updates obtained by the server. In particular, this is where privacy-preserving filters such as DP are implemented [38]. DP is used in FL to lower the probability that patient-level information is revealed through model updates, most notably in sensitive domains like genomics [39] and EHR analyses. While DP offers mathematical privacy guarantees, practical implementation requires careful selection of privacy budgets and recognition of real tradeoffs in model utility. DP must be governance-reviewed and considered alongside other protections, given its clear limitations in the federated paradigm. One DP implementation available in NVIDIA FLARE involves the specification of three main parameters: *eps1*, *eps3,* and fraction of the model to upload [38]. The main challenge is the choice of *eps* values as they can be dataset or site specific [40] and impact performance. This was highlighted in an evaluation that used the electronic intensive care unit Collaborative Research Database [41]. The authors found it difficult to achieve good performance for DP (both for hyperparameters that were fixed globally or chosen at each site). This well-known trade-off between security and performance [42] is also observed in benchmarking using this DP implementation [38].

To make the determination of the *eps* values tractable, we suggest that the DP approach be parameterized by first conducting a smaller-scale study with lower-risk data to evaluate the DP parameters with respect to the model performance measures, for instance, using similar publicly available data from Genomic Data Commons [43] or one of the other Cancer Research Data Commons resources [44]. The lowest viable *eps* values would then be chosen from the smaller-scale study. If this is not the case using acceptably low values of DP, then the Prime Site Lead can reevaluate the model, the need for access to sensitive data or FL as a viable strategy. Also, as a guide, federated DP has been evaluated in settings relevant to AMCs such as genomics [45-48].

In addition to DP, NVIDIA FLARE provides an implementation of another potentially complementary tool for reducing data risk to AMCs, called HE [49]. HE works in addition to standard security provided by Secure Sockets Layer and allows each client to further specially encrypt each model update such that the server can still aggregate them [50]. Importantly, the server never has access to the unencrypted updates. The client, however, can decrypt the received model and can continue training on its local data. In this manner, HE increases data security with minimal impacts on model performance and can protect against attempts to recover client data from model updates. However, the extra encryption steps do result in increased time needed for an FL training run. For instance, in benchmarks from NVIDIA, a 20% increase in time was noted along with a 15× increase in message size [49]. This addresses the risk of insider attacks on server infrastructure hosting FL but comes with substantial computational and resource costs. Its application in AMC settings is highly specialized, suitable mainly for high-sensitivity projects and should be governance-reviewed rather than adopted as standard practice. Note that HE does not protect against private information being memorized in the final model as part of the training process and therefore could also be considered in conjunction with a privacy-preserving approach such as DP [12]. Due to the complexity of DP and HE and related approaches that are under development, it is vital that staff members who are versed in their use and performance tradeoffs are consulted when implementing methods to enhance security and privacy.

Throughout the entire process, effective collaboration between the CISO and AIGC teams is essential. The CISO provides leadership on technical security, compliance, and risk mitigation, while the AIGC brings expertise in ethical, operational, and clinical considerations. Together, they create a comprehensive, enforceable governance structure that balances innovation with safety, privacy, and regulatory compliance. This partnership is crucial for ensuring that FL initiatives at AMCs are both secure and aligned with the institution's mission and values.

Operationalizing the Framework: Illustrative Examples

In order to demonstrate how this framework could be operationalized, we will highlight 2 examples. In our first example, there is a collaboration among multiple AMCs to develop an extreme gradient boosting model using deidentified clinical data for research use only purposes. As the data are deidentified, they would be considered low risk; additionally, since the model has limited capacity and would be used only in a research context, it would also be considered low risk. Only an expedited AIGR would be needed in this scenario with standard role-based access control. On the other hand, if genomic data were also used, the data could be considered high risk depending on the institutional (and national) policies while the model would continue to be considered low risk. In this scenario, an AIGR and SecR would be necessary to ensure that the data were protected even if used with a low-risk model. The necessary artifacts here would be literature documentation and computational notebooks with analyses, indicating lack of memorization and robustness to attack for the extreme gradient boosting model. Similarly, role qualification documentation and the "authorization.json" config file for NVIDIA FLARE would be needed to ensure appropriate access to the data and model.

As a second example, a group of AMCs is collaborating to develop a large language model to be used as part of clinical decision support using electronic health record data—including clinical genomics. This would be a scenario where both the data and model would be considered high risk, especially given its potential use operationally in a clinical setting. In this case, a thorough AIGR and SecR would be needed, and the implementation of additional privacy and security measures such as HE and DP would be required. The full range of artifacts presented in Textbox 1 would need to be collected and reviewed by both AIGC and the office of the CISO.

Discussion

FL holds tremendous promise for high-risk data sharing and model development in AMCs. Our proposed risk framework for AMCs assists CISOs and AIGC as well as other leadership by providing 4 main categories based on data and model risk, ensuring they are clearly communicated and transparent. This is critical given the rapidly evolving US regulatory developments around AI.

In addition, there is significant movement internationally toward proportionate governance of AI in health care. The European Union's Artificial Intelligence Act, which came into force in August 2024, classifies health care AI applications into divergent risk categories (eg, high risk: clinical decision support, and low risk: AI chatbots providing advice on well-being) [51]. High-risk AI in health care must meet strict requirements for risk management, data governance, human oversight (eg, clinicians must be able to contest AI outputs), transparency, and bias auditing, with integrated conformity assessments for both safety and data protection. In parallel, Canada's Artificial Intelligence and Data Act applies a risk-based framework to "high-impact" AI systems, focusing on protecting health, safety, and fundamental rights. The Artificial Intelligence and Data Act emphasizes robust risk assessment and mitigation, data management integrity, and continuous monitoring to ensure fairness and prevent harm or bias in health care applications. These initiatives demonstrate an emerging global consensus: international policy is converging on standards-informed and risk-proportionate oversight for AI in health. FL leads to shared legal responsibilities that need to be clearly understood based on data and model risks [52]. The work presented here is timely and can also support these efforts by allowing us to move from policy and regulation to implementation.

This work can also help guide operational strategies. For example, the use of privacy and security technologies such as DP and HE should be limited to high-risk data due to their costs, whether in terms of lower performance for DP or increased resource utilization, such as HE. In addition, there are hardware-based alternatives that can be considered such as Trusted Execution Environments as well as other approaches to confidential computing. NVIDIA FLARE can leverage Trusted Execution Environments as part of a confidential computing strategy to provide stronger security and privacy guarantees for FL. NVIDIA FLARE is adding support for virtual machine-based confidential computing technologies for both central processing units and graphics processing units, enabling further protection against compromised clients [53,54].

AI governance should play a key role in mitigating model risk by requiring certification of models before they are to be deployed on an FL network. This would help ensure that the models being developed are effective, fair, and preserve privacy. However, current levels of heterogeneity in AI maturity models and governance have implications for FL partnerships if the prime site does not have the appropriate infrastructure in place to support the required activities (including pretraining certification, etc). For consortiums, the best approach may be a centralized AI governance process with the most mature site leading the evaluation. The establishment of consensus benchmarking metrics for security and privacy review (as proposed in the former executive order 14110 [55]) would also help facilitate this process tremendously. Security and risk management in FL, as with most other complex systems, is a continually evolving landscape of newly identified risks and countermeasures. For instance, even with the implementation of DP and HE, there are still cases where information can be extracted from an FL system [56]. Importantly, attacks such as this require insider knowledge of the system and problem domain. We acknowledge that FL is an evolving domain, and as it becomes increasingly adopted by AMCs, guidance will need to be updated due to improved methodologies and to address new challenges.

Acknowledgments

The authors would like to extend their sincere thanks to Shannon Silkensen and Umit Topaloglu at National Cancer Institute for their outstanding leadership and visionary guidance in launching the inaugural federated learning supplement. While not directly related to this security and governance focus of this paper, Dirk Petersen was instrumental in the initial set of the National Cancer Institute NVIDIA Federated Learning Application Runtime Environment server and in streamlining the onboarding for other sites. They also acknowledge Roland Niedner, Greg Flanigan, and Linmin Pei for their coordination and support at National Cancer Institute.

Funding

This work was supported by a National Cancer Institute supplement to the OHSU Knight Cancer Institute Cancer Center Support grant (3P30CA069533-25S1).

Authors' Contributions

DB: project conception, manuscript preparation, and editing. BB, KM, NL: assistance in developing security risk matrix and manuscript editing. HRR, CC: assistance in developing security risk matrix and implications and best practices for federated learning implementation, manuscript preparation, and editing. SKM: project conception, project oversight, manuscript preparation, and editing. All authors reviewed the final manuscript.

Conflicts of Interest

HRR and CC are employees of NVIDIA, which developed the open-source framework NVIDIA Federated Learning Application Runtime Environment.

Multimedia Appendix 1

Listing of federated learning platforms, which provides an overview and security considerations for the NVIDIA Federated Learning Application Runtime Environment user roles.
[XLSX File (Microsoft Excel File), 18 KB-Multimedia Appendix 1]

References

1. Nong P, Adler-Milstein J, Apathy NC, Holmgren AJ, Everson J. Current use and evaluation of artificial intelligence and predictive models in US hospitals. Health Aff (Millwood). Jan 2025;44(1):90-98. [doi: 10.1377/hlthaff.2024.00842] [Medline: 39761454]
2. Lenharo M. Medicine's rapid adoption of AI has researchers concerned. Nature New Biol. Jun 9, 2025. [doi: 10.1038/d41586-025-01748-y] [Medline: 40490519]
3. Nong P, Hamasha R, Singh K, Adler-Milstein J, Platt J. How academic medical centers govern AI prediction tools in the context of uncertainty and evolving regulation. NEJM AI. Feb 22, 2024;1(3). [doi: 10.1056/AIp2300048]
4. Lyons PG, Dorr DA, Melton GB, Singh K, Payne PRO. Meeting the artificial intelligence needs of U.S. health systems. Ann Intern Med. Oct 2024;177(10):1428-1430. [doi: 10.7326/ANNALS-24-00396] [Medline: 39186786]
5. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119. [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]
6. Luzón MV, Rodríguez-Barroso N, Argente-Garrido A, et al. A tutorial on federated learning from theory to practice: foundations, software frameworks, exemplary use cases, and selected trends. IEEE/CAA J Autom Sinica. Apr 2024;11(4):824-850. [doi: 10.1109/JAS.2024.124215]
7. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health. May 2022;1(5):e0000033. [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
8. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol. Mar 31, 2019;10(2):1-19. [doi: 10.1145/3298981]
9. Rodríguez-Barroso N, Jiménez-López D, Luzón MV, Herrera F, Martínez-Cámara E. Survey on federated learning threats: concepts, taxonomy on attacks and defences, experimental study and challenges. Information Fusion. Feb 2023;90:148-173. [doi: 10.1016/j.inffus.2022.09.011]
10. Mhamdi EME, Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in byzantium. Presented at: Proceedings of the 35 th International Conference on Machine Learning; Jul 10-15, 2018; Stockholm, Sweden. 2018.URL: https://proceedings.mlr.press/v80/mhamdi18a/mhamdi18a.pdf [Accessed 2026-01-16]
11. Lamport L, International SRI, Shostak R, Pease M, SRI International. The byzantine generals problem. In: Concurrency: The Works of Leslie Lamport. Association for Computing Machinery; 2019:203-226. [doi: 10.1145/3335772.3335936] ISBN: 9781450372701
12. Zhao J, Bagchi S, Avestimehr S, et al. The federation strikes back: a survey of federated learning privacy attacks, defenses, applications, and policy landscape. ACM Comput Surv. Sep 30, 2025;57(9):1-37. [doi: 10.1145/3724113]
13. Roth HR, Cheng Y, Wen Y, et al. NVIDIA FLARE: federated learning from simulation to real-world. Presented at: International Workshop on Federated Learning, NeurIPS 2022; Nov 28 to Dec 9, 2022; New Orleans, USA. 2022.URL: https://www.catalyzex.com/paper/nvidia-flare-federated-learning-from [Accessed 2026-01-16]

14. Administrative supplements for P30 cancer centers support grants (CCSG) to establish proof-of concept federated learning frameworks that will run multimodal artificial intelligence models. National Cancer Institute; 2023. URL: https://cdn.bcm.edu/sites/default/files/admin-suppl-2023-rwd-final.pdf [Accessed 2026-01-16]
15. Tabassi E. Artificial intelligence risk management framework (AI RMF 10). National Institute of Standards and Technology; 2023. URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf [Accessed 2026-01-16]
16. ISO/IEC 42001:2023(en), information technology — artificial intelligence — management system. International Organization for Standardization; 2023. URL: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en [Accessed 2026-01-16]
17. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019:220-229; Atlanta, GA, USA. Jan 29, 2019.[doi: 10.1145/3287560.3287596]
18. Bellamy RKE, Dey K, Hind M, et al. AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev. 2019;63(4/5):4. [doi: 10.1147/JRD.2019.2942287]
19. Fairlearn. Github. URL: https://github.com/fairlearn/fairlearn [Accessed 2026-01-16]
20. Saleiro P, Kuester B, Hinkson L, et al. Aequitas: a bias and fairness audit toolkit. arXiv. Preprint posted online on Nov 4, 2018. URL: http://arxiv.org/abs/1811.05577 [Accessed 2026-01-16]
21. Carlini N, Tramèr F, Wallace E, et al. Extracting training data from large language models. Presented at: Proceedings of the 30th USENIX Security Symposium; Aug 11-13, 2021:2633-2650; Boston, MA, USA. 2021.URL: https://www.usenix.org/system/files/sec21-carlini-extracting.pdf [Accessed 2026-01-16]
22. Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. Presented at: Proceedings of the 28th USENIX Conference on Security Symposium; Aug 14-16, 2019; Santa Clara, CA, USA. 2019.[doi: 10.5555/3361338.3361358]
23. Hartley J, Sanchez PP, Haider F, Tsaftaris SA. Neural networks memorise personal information from one sample. Sci Rep. Dec 4, 2023;13(1):21366. [doi: 10.1038/s41598-023-48034-3] [Medline: 38049432]
24. Schwarzschild A, Feng Z, Maini P, Lipton ZC, Kolter JZ. Rethinking LLM memorization through the lens of adversarial compression. Presented at: Proceedings of the 38th International Conference on Neural Information Processing Systems; Dec 10-15, 2024; Vancouver, BC, Canada. 2024.[doi: 10.5555/3737916.3739706]
25. Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. Presented at: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; Oct 24-28, 2016; Vienna Austria. Oct 24, 2016. [doi: 10.1145/2976749.2978318]
26. Anil R, Ghazi B, Gupta V, Kumar R, Manurangsi P. Large-scale differentially private BERT. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2022; Dec 7-11, 2022; Abu Dhabi, United Arab Emirates. 2022. [doi: 10.18653/v1/2022.findings-emnlp.484]
27. Tramèr F, Kamath G, Carlini N. Position: considerations for differentially private learning with large-scale public pretraining. Presented at: Proceedings of the 41st International Conference on Machine Learning; Jul 21-27, 2024; Vienna, Austria. 2024.[doi: 10.5555/3692070.3694051]
28. Li X, Tramèr F, Liang P, Hashimoto T. Large language models can be strong differentially private learners. Presented at: Proceedings of the 10th International Conference on Learning Representations (ICLR 2022); Apr 25-29, 2022. 2022.URL: https://openreview.net/pdf?id=bVuP3ltATMz [Accessed 2026-01-16]
29. John MM, Olsson HH, Bosch J. Towards mlops: a framework and maturity model. Presented at: Proceedings of the 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA); Sep 1-3, 2021; Palermo, Italy. 2021.[doi: 10.1109/SEAA53835.2021.00050]
30. Terminologies and roles. NVIDIA FLARE. URL: https://nvflare.readthedocs.io/en/main/user_guide/security/terminologies_and_roles.html [Accessed 2026-01-16]
31. Executive order on preventing access to americans' bulk sensitive personal data and United States government-related data by countries of concern. The White House. 2024. URL: https://www.federalregister.gov/documents/2025/01/08/2024-31486/preventing-access-to-us-sensitive-personal-data-and-government-related-data-by-countries-of-concern [Accessed 2026-1-23]
32. Justice department implements critical national security program to protect americans' sensitive data from foreign adversaries. Office of Public Affairs - US Department of Justice. 2025. URL: https://www.justice.gov/opa/pr/justice-department-implements-critical-national-security-program-protect-americans-sensitive [Accessed 2026-01-16]
33. Stone R. Researchers from china and five other 'countries of concern' barred from NIH databases. American Association for the Advancement of Science. 2025. URL: https://www.science.org/content/article/researchers-china-and-five-other-countries-concern-barred-nih-databases [Accessed 2026-01-16]

34. Thakkar OD, Ramaswamy S, Mathews R, Beaufays F. Understanding unintended memorization in language models under federated learning. Presented at: Proceedings of the Third Workshop on Privacy in Natural Language Processing; Jun 11, 2021:1-10; 2021.[doi: 10.18653/v1/2021.privatenlp-1.1]
35. Bossy T, Vignoud J, Rabbani T, Pastoriza JRT, Jaggi M. Mitigating unintended memorization with lora in federated learning for llms. arXiv. Preprint posted online on Feb 7, 2025. [doi: 10.48550/arXiv.2502.05087]
36. Saltzer JH, Schroeder MD. The protection of information in computer systems. Proc IEEE. 1975;63(9):1278-1308. [doi: 10.1109/PROC.1975.9939]
37. Site policy management. NVIDIA FLARE. URL: https://nvflare.readthedocs.io/en/2.4/user_guide/security/site_policy_management.html [Accessed 2026-01-16]
38. Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation. Presented at: Proceedings of the 10th International Workshop on Machine Learning in Medical Imaging (MLMI 2019), in conjunction with MICCAI 2019; Oct 13-17, 2019; Shenzhen, China. 2019.[doi: 10.1007/978-3-030-32692-0_16]
39. Pathade C, Patil S. Securing genomic data against inference attacks in federated learning environments. arXiv. Preprint posted online on May 12, 2025. [doi: 10.48550/arXiv.2505.07188]
40. Lee J, Clifton C. How much is enough? Choosing ε for differential privacy. Presented at: Proceedings of the 14th International Conference on Information Security; Oct 26-29, 2011:325-340; Xi'an, China. 2011.[doi: 10.1007/978-3-642-24861-0_22]
41. Pfohl SR, Dai AM. Federated and differentially private learning for electronic health records. arXiv. Preprint posted online on Nov 13, 2019. [doi: 10.48550/arXiv.1911.05861]
42. Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy. Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019; Vancouver, Canada. 2019.[doi: 10.5555/3454287.3455674]
43. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. Sep 22, 2016;375(12):1109-1112. [doi: 10.1056/NEJMp1607591] [Medline: 27653561]
44. Hinkson IV, Davidsen TM, Klemm JD, Chandramouliswaran I, Kerlavage AR, Kibbe WA. Corrigendum: a comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. Front Cell Dev Biol. 2017;5:108. [doi: 10.3389/fcell.2017.00108] [Medline: 29243742]
45. Aziz MMA, Anjum MM, Mohammed N, Jiang X. Generalized genomic data sharing for differentially private federated learning. J Biomed Inform. Aug 2022;132:104113. [doi: 10.1016/j.jbi.2022.104113] [Medline: 35690350]
46. Wen G, Li L. Federated transfer learning with differential privacy for multi-omics survival analysis. Brief Bioinform. Mar 4, 2025;26(2):bbaf166. [doi: 10.1093/bib/bbaf166] [Medline: 40230038]
47. Li W, Kim M, Zhang K, Chen H, Jiang X, Harmanci A. COLLAGENE enables privacy-aware federated and collaborative genomic data analysis. Genome Biol. Sep 11, 2023;24(1):204. [doi: 10.1186/s13059-023-03039-z] [Medline: 37697426]
48. Zhou J, Chen S, Wu Y, et al. PPML-omics: a privacy-preserving federated machine learning method protects patients' privacy in omic data. Sci Adv. Feb 2, 2024;10(5):eadh8601. [doi: 10.1126/sciadv.adh8601] [Medline: 38295178]
49. Roth H, ZephyrM, Harouni A. Federated learning with homomorphic encryption. NVIDIA DEVELOPER. 2021. URL: https://developer.nvidia.com/blog/federated-learning-with-homomorphic-encryption/ [Accessed 2026-01-16]
50. Cheon JH, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. Presented at: Proceedings of the 23rd Annual International Conference on Theory and Application of Cryptology and Information Security, ASIACRYPT 2017; Dec 3-7, 2017:409-437; Hong Kong. 2017.[doi: 10.1007/978-3-319-70694-8_15]
51. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy. Nov 2024;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
52. Woisetschläger H, Mertel S, Krönke C, Mayer R, Jacobsen HA. Federated learning and AI regulation in the European Union: who is responsible?: An interdisciplinary analysis. arXiv. Preprint posted online on Jul 11, 2024. [doi: 10.48550/arXiv.2407.08105]
53. Decentralized collaborative AI with federated learning in trustworthy environments. NVIDIA On-Demand. 2024. URL: https://www.nvidia.com/en-us/on-demand/session/gtc24-s62149/ [Accessed 2026-01-16]
54. Verma S, Arunagiri S, Chen C. Building llms with NVIDIA nemo and securing them with confidential computing. Confidential Computing Summit. 2025. URL: https://www.confidentialcomputingsummit.com/e/ccs25/page/agenda [Accessed 2026-01-16]
55. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023. URL: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence [Accessed 2026-01-16]

56. Yuan X, Ma X, Zhang L, Fang Y, Wu D. Beyond class-level privacy leakage: breaking record-level privacy in federated learning. IEEE Internet Things J. Feb 2022;9(4):2555-2565. [doi: 10.1109/JIOT.2021.3089713]

Abbreviations

AI: artificial intelligence
AIGC: Artificial Intelligence Governance Committee
AIGR: artificial intelligence governance review
AMC: academic medical center
CISO: Chief Information Security Officer
CoC: Countries of Concern
DP: differential privacy
FL: federated learning
HE: homomorphic encryption
IEC: International Electrotechnical Commission
ISO: International Organization for Standardization
ML: machine learning
NIST AI RMF: National Institute of Standards and Technology Artificial Intelligence Risk Management Framework
NVIDIA FLARE: NVIDIA Federated Learning Application Runtime Environment
SecR: security review

Edited by Gang Luo; peer-reviewed by Karthigeyan Kuppan, Luk Arbuckle; submitted 02.Jul.2025; final revised version received 20.Nov.2025; accepted 11.Dec.2025; published 27.Feb.2026

Please cite as:
Bottomly D, Barnes B, Mavuwa K, Lee N, Roth HR, Chen C, McWeeney SK
A Pragmatic Framework for Federated Learning Risk and Governance in Academic Medical Centers
JMIR AI 2026;5:e80022
URL: https://ai.jmir.org/2026/1/e80022
doi: 10.2196/80022