Federated Quality Control of Clinical TLF Outputs: Extending TLFQC with Privacy-Preserving Cross-Site Validation for Multi-Center Trials

Zhouhua Chen
Department of Computer Science, George Mason University, Fairfax, VA, USA.
zhouhua.chen618@gmu.edu

Quentin Gregory
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
helloquentin@colostate.edu

Abstract

The quality control of Tables, Listings, and Figures (TLFs) in clinical trials remains a critical yet labor-intensive step, particularly in multi-center settings where data privacy regulations restrict the sharing of patient-level information. Existing automated QC platforms, such as TLFQC, offer substantial efficiency gains for single-site operations but lack native support for cross-site validation without compromising data confidentiality. This paper proposes a federated quality control framework that extends the capabilities of TLFQC to enable privacy-preserving cross-site validation for multi-center trials. The architecture leverages a combination of federated learning principles, differential privacy noise injection, and secure aggregation protocols to allow participating sites to collaboratively assess the quality and consistency of TLF outputs without exposing raw clinical data. We discuss the structural trade-offs involved in designing such a system, including the tension between privacy guarantees and validation accuracy, the overhead of cryptographic operations, and the need for harmonized metadata standards across sites. The governance implications of trustless coordination, auditability, and compliance with regulatory frameworks such as HIPAA and GDPR are examined in depth. Deployment considerations, including infrastructure requirements for site-level computing resources and network bandwidth, are addressed through a tiered architecture that accommodates heterogeneous site capabilities. Robustness against adversarial inputs, fairness in quality metrics across diverse patient populations, and long-term sustainability of the federated system are analyzed through cross-domain comparisons with federated learning in medical imaging and electronic health records. A case illustration from an oncology trial demonstrates how the framework can detect systematic discrepancies in adverse event reporting across sites while preserving privacy. The paper concludes with policy recommendations for adopting federated QC in future multi-center trials and outlines directions for extending the approach to real-time monitoring and adaptive quality thresholds.

Keywords

federated quality control, clinical TLF output validation, privacy-preserving cross-site validation, multi-center clinical trials, differential privacy, secure aggregation, TLFQC, clinical data governance.

1. Introduction

The generation and validation of Tables, Listings, and Figures (TLFs) constitute a fundamental component of clinical trial reporting, supporting regulatory submissions, safety monitoring, and efficacy analyses. As clinical trials increasingly involve multiple geographically dispersed centers, the need for consistent and high-quality TLF outputs across sites has intensified. However, traditional quality control (QC) processes remain predominantly manual, site-specific, and vulnerable to inconsistencies in interpretation, formatting, and statistical methodology. Automated tools such as TLFQC have emerged to streamline TLF generation and validation within single-site environments [5], yet they are not designed to handle the cross-site coordination required in multi-center trials where patient-level data cannot be shared due to privacy regulations and institutional policies. This paper presents a federated quality control framework that extends the principles of federated learning to the domain of TLF validation, enabling sites to collaboratively verify the correctness and comparability of their outputs without centralizing sensitive data. We argue that such a system not only improves QC efficiency but also strengthens the statistical integrity of multi-center trials by detecting site-level drifts and systematic errors early in the data pipeline.

The remainder of this paper is structured as follows. Section 2 provides background on TLF QC challenges and the current state of automated platforms. Section 3 describes the proposed federated architecture, detailing the roles of nodes, orchestrators, and privacy-preserving mechanisms. Section 4 focuses on the cross-site validation algorithms and their privacy guarantees. Section 5 discusses governance and policy implications. Section 6 addresses deployment and infrastructure. Section 7 examines robustness, fairness, and sustainability. Section 8 presents a case illustration in oncology. Section 9 concludes with future directions.

2. Background and Problem Context

Clinical trial data analysis produces a large volume of TLFs that summarize demographics, adverse events, laboratory results, and efficacy endpoints. The QC of these outputs is a multi-step process involving format checks, logic consistency verifications, and cross-referencing against source data. In a multi-center trial, each site generates its own TLF outputs based on its subset of patient data. The overall trial QC must ensure that TLFs from all sites are internally consistent, follow a common template, and adhere to the statistical analysis plan. Currently, this is achieved by either pooling de-identified data at a central coordinating center or by exchanging aggregated summary statistics. Both approaches have limitations. Pooling de-identified data still carries re-identification risks, and sharing summary statistics may not reveal subtle errors such as misformatted tables or incorrect denominators. Furthermore, centralizing QC creates a single point of failure and increases administrative burden.

Automated QC platforms like TLFQC [5] have addressed the single-site scenario by providing a codeless environment in R Shiny that automates TLF generation and validation against predefined rules. This platform significantly reduces human effort and error, but it remains confined to the data of one site. Extending such a platform to a multi-site context requires mechanisms for cross-site validation that do not violate data privacy. The concept of federated learning, originally developed for training machine learning models across decentralized data [1], offers a promising blueprint. In federated learning, model parameters are exchanged rather than raw data, and privacy is further enhanced through techniques like differential privacy [2] and secure multi-party computation [3]. Applying similar ideas to TLF QC presents unique challenges because TLF validation often involves comparing counts, distributions, and formatting rules that are not straightforwardly represented as model

gradients. Nonetheless, the core principle of aggregating site-level quality metrics without revealing underlying patient records is directly applicable.

3. Architecture of Federated QC for TLF Outputs

The proposed federated quality control framework, which we term FedTLFQC, builds upon the existing TLFQC platform [5] by adding a privacy-preserving cross-site validation layer. The architecture comprises three tiers: the site level, the federation orchestrator, and the regulatory auditor interface. At each site, an instance of TLFQC runs locally, generating and validating TLF outputs according to site-specific data. These local QC processes produce a set of quality indicators, such as the number of tables generated, the proportion of cells with missing values, the consistency of formatting with the predefined template, and statistical checks like McNemar tests for paired comparisons. Instead of transmitting these indicators in plain text, the system applies a privacy-preserving transformation before sending them to the orchestrator.

The orchestrator is a central server, but it never receives raw data or unprotected quality indicators. Site-level outputs are first perturbed using differential privacy via the Laplace mechanism, ensuring that any single patient's contribution is indistinguishable within a bounded privacy budget. The orchestrator then aggregates these noisy indicators using secure aggregation protocols [3] to prevent the server from learning individual site contributions. The aggregated results, after subtracting the added noise aggregated over sites, yield an estimate of the true cross-site quality statistics. These statistics are used to detect anomalies: for instance, if the aggregated proportion of a certain adverse event category across sites deviates significantly from the expected range, the orchestrator flags the trial for further investigation. Sites can then request a targeted re-validation of specific TLFs without exposing additional data.

A key architectural decision is the trade-off between privacy and accuracy. Increasing the privacy budget reduces noise but increases re-identification risk. Conversely, a lower privacy budget may obscure true quality issues. The framework allows trial sponsors to set a global privacy budget that balances these concerns based on the sensitivity of the trial. Additionally, the architecture supports differential privacy with adaptive composition [2] to account for repeated queries during the trial. Another trade-off involves computational overhead. Secure aggregation requires pairwise masking and unmasking, which adds latency and bandwidth consumption. To mitigate this, the architecture leverages a hierarchical aggregation scheme where sites within the same regional network are first aggregated locally before sending results to the global orchestrator.

4. Privacy-Preserving Cross-Site Validation Mechanisms

The core validation mechanisms in FedTLFQC are designed to compare TLF outputs across sites without revealing patient-level details. We distinguish between two types of validation: structural validation and statistical validation. Structural validation checks that all TLFs follow the same template, including column headers, row order, and font styles. These checks are performed locally, and each site produces a binary or categorical conformity score. To compare these scores across sites without leaking which site deviates, the system uses a secure comparison protocol based on garbled circuits [4]. Each site submits an encrypted version of its conformity score; the orchestrator runs a garbled circuit that computes the set of sites where conformity fails, but only reveals an aggregated count rather than individual identities.

Statistical validation involves verifying consistency of aggregated statistics such as incidences, means, and proportions across sites. For count data, the system employs a federated Chi-square test. Each site computes its observed and expected counts within its own data, but instead of sharing the raw counts, it shares a noisy version of the Chi-square statistic, masked with differential privacy. The orchestrator sums the noisy statistics across sites and adjusts for the added noise. The resulting aggregated test statistic is compared against a threshold to detect systematic differences. This approach is analogous to the federated statistics methods used in multi-site EHR studies [6]. A limitation is that differential privacy noise can lead to false alarms or missed detections when site counts are small. To address this, the framework includes a calibration step that increases noise only for small sample sizes and uses a relaxed privacy budget for summary-level queries.

Another mechanism is the federated comparison of TLF formatting rules. Many TLFs include conditional formatting such as bolding of significant p-values or highlighting of out-of-range values. Each site maintains a local rule engine that checks whether the conditional formatting is applied correctly. Rather than transmitting the actual conditional flags, sites share a hash of the rule application pattern, which is further encrypted using a secret sharing scheme. The orchestrator can then verify that all sites have applied the same rules without learning the actual pattern. This hash-based approach, combined with secret sharing, provides computational efficiency while preserving privacy, though it is vulnerable to collision attacks if the hash space is too small. Therefore, the framework recommends a collision-resistant hash function and periodic salt refresh.

5. Governance and Policy Considerations

Implementing federated QC in multi-center trials raises significant governance questions regarding data stewardship, trust, and regulatory acceptance. Traditional quality control relies on a central authority that has full access to data; in a federated setting, no single party holds all information. This shifts the locus of trust from a central entity to a distributed trust model enforced by cryptographic protocols. Trial sponsors must negotiate data use agreements that specify the privacy budget, the types of queries allowed, and the procedures for escalating anomalies. The framework must also comply with HIPAA in the United States and GDPR in Europe. Under GDPR, the concept of data minimization is central: federated QC inherently minimizes data transfer, but the differential privacy parameters must be documented to demonstrate compliance with the principle of data protection by design.

Another governance challenge is auditability. Regulatory agencies such as the FDA require the ability to reconstruct the QC process for inspection. In a federated system, the orchestrator logs all aggregated queries and the privacy budget consumed, but individual site contributions remain encrypted. To satisfy auditability, the framework incorporates a verifiable audit trail using blockchain-inspired hashing of the aggregated outputs. Each site cryptographically signs its noisy contributions, and the orchestrator appends these signatures to an immutable ledger. During an audit, the agency can verify that the aggregated statistics were computed correctly without needing to decrypt site-level data. This approach has been proposed for federated learning in healthcare [7] and adapts naturally to QC.

Policy implications extend to liability and error attribution. If a QC discrepancy is discovered, the framework must enable the identification of the responsible site while preserving privacy for others. This is achieved through a selective disclosure protocol: the orchestrator, with the consent of the trial sponsor, can request that a specific site reveal its raw quality indicators after the fact, but only if the site’s deviation is statistically significant beyond a high threshold.

This threshold is agreed upon in the data use agreement. Such a mechanism balances individual site autonomy with the need for accountability.

6. Deployment and Infrastructure Challenges

Deploying FedTLFQC across diverse clinical trial sites requires careful consideration of local computing resources, network connectivity, and staff expertise. Many clinical sites have limited IT infrastructure and rely on air-gapped systems for data security. The federated architecture must accommodate these constraints without compromising functionality. We propose a tiered deployment model. Tier 1 sites have dedicated servers and high-speed internet; they run the full TLFQC instance with local validation and differential privacy mechanisms. Tier 2 sites have limited connectivity and may need to batch-process their validations and transmit them asynchronously. For Tier 2, the framework employs a store-and-forward mechanism with encrypted archives that are uploaded when connectivity is restored. Tier 3 sites (e.g., rural clinics) may only have a basic infrastructure; for these, the system can run a lightweight version that only computes a minimal set of quality indicators and transmits them via short message-based protocols.

Network bandwidth is a critical constraint. Secure aggregation and garbled circuits introduce significant communication overhead. For a trial with 20 sites each contributing 50 quality indicators per TLF type (e.g., demographics, adverse events, lab data), the cumulative data transmitted could reach several gigabytes over the trial duration. To reduce overhead, the framework uses compression and differential privacy with a relaxed budget for less sensitive indicators. Additionally, the orchestrator can schedule aggregation rounds at off-peak hours and cache intermediate results.

Another infrastructure challenge is the harmonization of data dictionaries and TLF templates across sites. Differences in local coding systems (e.g., MedDRA version, unit conversions) can lead to misinterpretations of quality indicators. FedTLFQC requires all sites to map their local codes to a common ontology before validation. This mapping can be done locally using a shared LOINC or SNOMED mapping service, but the mapping itself may leak information about patient populations. To address this, the mapping process is performed on a placeholder dataset that does not contain real patient data, and the mapping rules are agreed upon before the trial begins. This preprocessing step is analogous to the data harmonization required in multi-site observational studies [8].

7. Robustness, Fairness, and Sustainability

A federated QC system must be robust to various failure modes, including site dropout, malicious submissions, and network partitions. Site dropout can be handled by using robust aggregation methods that exclude contributions from sites that do not respond within a timeout. However, if a site is compromised and submits fabricated quality indicators, the aggregated results may be skewed. To counter this, the framework incorporates a Byzantine fault-tolerant aggregation protocol [9], which uses median or trimmed mean instead of the average. These robust statistics are less sensitive to outliers at the cost of some statistical power. The choice between robustness and efficiency should be calibrated based on the risk profile of the trial.

Fairness considerations arise when the patient populations across sites are heterogeneous. For example, a site enrolling predominantly elderly patients may have different adverse event patterns than a site enrolling younger patients. Federated QC should not penalize such sites for legitimate demographic differences. To avoid bias, the system normalizes quality

indicators by stratified rates based on age, sex, and comorbidities, using local summary statistics that are themselves privacy-preserving. The normalized indicators are then compared across sites, reducing the risk of flagging true population differences as errors. This fairness mechanism aligns with the broader call for equitable AI in healthcare [10].

Sustainability of the federated framework depends on the ability to update validation rules as the trial progresses without incurring excessive privacy costs. Adaptive composition of differential privacy [2] allows the system to manage the cumulative privacy budget across multiple queries. However, if the number of queries grows large, the noise level may become unacceptable. One solution is to limit the total number of validation queries per site per patient cohort, a common practice in federated analytics. Another is to use local differential privacy with randomization at each query, which trades off accuracy for unlimited query support. The sustainability also relies on the willingness of sites to invest in the required infrastructure. Providing open-source toolkits and training programs can lower adoption barriers, as demonstrated by the successful uptake of federated learning platforms in oncology imaging [11].

8. Case Illustration: Oncology Trial Adverse Event Monitoring

To illustrate the utility of FedTLFQC, consider a Phase III multi-center randomized trial for a new oncology drug, with 15 participating sites across the United States and Europe. The primary endpoint is progression-free survival; secondary endpoints include the incidence of grade 3 or higher adverse events (AEs). Each site produces TLFs summarizing AE counts by system organ class and severity grade. The trial sponsor is concerned about site-level differences in AE reporting due to varying local diagnostic practices. Using the proposed federated framework, each site first runs TLFQC locally [5] to generate formatted tables and list all AEs. The QC engine performs structural checks (e.g., table column alignment, page numbering) and statistical checks (e.g., comparing observed AE frequencies to expected frequencies from published literature). The structural checks yield a binary conformity flag; the statistical checks produce a normalized deviation score.

Each site perturbs its conformity flag and deviation score using differential privacy with epsilon = 1.0. The orchestrator receives aggregated noisy data from all 15 sites, applies secure aggregation, and subtracts the aggregated noise. The resulting cross-site average deviation score is 0.12, which is within the normal range of 0 to 0.2, so no site is flagged. However, the structural conformity flag shows that 3 out of 15 sites reported 0 (non-conformant) while 12 reported 1. The federated system reveals that the aggregated non-conformant count is 3, but does not identify which sites. The sponsor then triggers a selective disclosure protocol: the orchestrator requests all sites to submit a second-round stricter conformity check only for the TLF type that failed. The second round uses a higher privacy budget (epsilon = 0.1 per site) and a secure comparison that identifies the specific sites without revealing their other data. The result shows that the three non-conformant sites all had an extra column inadvertently added to their AE table. A conference call resolves the issue by updating the template. Throughout this process, no patient-level data left any site, and the sponsor only learned the minimal information needed to address the problem.

This case demonstrates how federated QC can detect systematic errors while respecting privacy. It also highlights the trade-off of using differential privacy: the initial flag of three non-conformant sites could have been masked if the noise was too high. Therefore, the trial's privacy budget must be chosen carefully based on the expected number of errors.

9. Conclusion

The extension of TLFQC with a federated privacy-preserving cross-site validation layer offers a viable path toward scalable and secure quality control in multi-center clinical trials. This paper has presented the architectural design, validation mechanisms, governance frameworks, and deployment strategies for such a system. The approach balances the competing demands of privacy, accuracy, robustness, and auditability. While the system is currently conceptual, its building blocks—federated learning, differential privacy, secure aggregation, and garbled circuits—are mature enough to be integrated into production environments with careful engineering. Future work should focus on implementing a prototype FedTLFQC as an open-source add-on to existing QC platforms, conducting benchmark studies with real clinical trial data, and extending the framework to support real-time streaming QC during active trial enrollment. Additionally, the integration of fairness metrics and adaptive privacy budgeting will be essential for adoption across diverse global trial networks. The regulatory acceptance of such systems will depend on transparent documentation of the privacy guarantees and validation of the accuracy trade-offs. As multi-center trials become the norm, federated QC represents a necessary evolution towards trustworthy and efficient clinical data management.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 54, 1273–1282.

2. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 308–318.

3. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS), 1175–1191.

4. Yao, A. C. (1986). How to generate and exchange secrets. In 27th Annual Symposium on Foundations of Computer Science (FOCS), 162–167.

5. Ling, C., & Wang, Y. (2025). TLFQC: A High-compatible R Shiny based Platform for Automated and Codeless TLFs Generation and Validation. In PharmaSUG 2025 conference proceedings.

6. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

7. Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119.

8. Hripcsak, G., Duke, J. D., Shah, N. H., et al. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Studies in Health Technology and Informatics, 216, 574–578.

9. Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In Proceedings of the 35th International Conference on Machine Learning (ICML), 80, 5650–5659.

10. Rajkomar, A., Hardt, M., Howell, M. D., et al. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

11. Sheller, M. J., Edwards, B., Reina, G. A., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, 12598.

12. Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

13. Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

14. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

15. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570.

16. Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML), 807–814.

17. Ziller, A., Trask, A., Lopardo, A., et al. (2021). Privacy-preserving machine learning for healthcare. In Machine Learning for Healthcare Conference (MLHC), 2021.

18. Papernot, N., Abadi, M., Erlingsson, U., et al. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In Proceedings of the 5th International Conference on Learning Representations (ICLR).

19. Truex, S., Baracaldo, N., Anwar, A., et al. (2019). A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec), 1–11.

20. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), 1310–1321.