The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 299-306
DOI 10.37547/tajet/Volume07Issue08-26

Check for updates

Optimizing Data Quality and Compliance Through Integrated Validation Strategies for Clinical Systems

Ratna Jyothi Kommaraju
Data Manager, R&D Data strategy and governance, Sanofi (Contractor via Vivid Soft Global Inc) Old Tappan, New Jersey, USA

OPEN ACCESS

SUBMITTED 31 July 2025
ACCEPTED 16 August 2025
PUBLISHED 29 August 2025
VOLUME Vol.07 Issue 08 2025

CITATION
Ratna Jyothi Kommaraju. (2025). Optimizing Data Quality and Compliance Through Integrated Validation Strategies for Clinical Systems. The American Journal of Engineering and Technology, 7(8), 299–306. https://doi.org/10.37547/tajet/Volume07Issue08-26

Abstract: This article presents a comprehensive analysis of integrated data validation strategies in clinical systems, aimed at enhancing their quality and regulatory compliance. The study employs an interdisciplinary approach that combines risk-based quality management, data model standardization, and multi-level assessment procedures, including DQA and SSDQA, with a focus on their reproducibility and scalability. Particular attention is given to a comparative analysis of the operational efficiency of direct data entry and automated transfer from medical information systems, with a detailed evaluation of their impact on data preparation speed, metric reproducibility, reduction of transcription errors, and monitoring workload. Key factors determining validation effectiveness have been identified, including trial portfolio size, maturity of digital infrastructure, personnel readiness, and regional implementation specifics. Quantitative indicators of RBQM adoption and related tools, as well as data fitness-for-use metrics obtained from multicenter projects with varying levels of quality control maturity, are presented. The optimal validation architecture is defined as incorporating unified standards, continuous control during the execution stage, and adaptation of tools to local conditions to minimize risks. The article will be useful for clinical research professionals, data quality management system developers, regulatory experts, healthcare IT architects, and researchers in the field of digital transformation of medical technologies.

Keywords: integrated data validation, clinical systems,

quality management, risk-based monitoring, eSource technologies, data standardization.

Introduction

Clinical systems are undergoing a shift toward integrating risk-based quality management (RBQM), quality tolerance limits and central statistical monitoring (QTL/CSM), direct data capture (DDC), and automated EHR-to-EDC flows (EHR2EDC) on top of multi-level data quality assessment procedures (DQA/SSDQA) to simultaneously improve validity and regulatory compliance. Remote oversight plays the leading role in this transition. Risk-based central monitoring is positioned as a key mechanism for ensuring data integrity and participant safety under constraints on on-site supervision [1].

Large-scale empirical evaluations of central statistical monitoring across thousands of observations confirm its value as an instrument for early detection of systematic anomalies and inter-site variability [5]. At the quality-policy level, the formal QTL framework describes the selection of critical parameters, justification of thresholds, and action plans for excursions, thereby embedding risk management into the study life cycle [2]. Confidence in real-world evidence (RWE) depends directly on standardized quality assessment and source transparency, which requires reproducible procedures and explicit acceptability criteria [3].

The empirical base for integration is heterogeneous and reveals maturity gaps. RBQM adoption remains fragmented. The consolidated implementation level across components is about 57%, with considerable variation by stage and study portfolios [6]. At the same time, data-model standardization shows institutional maturity. Transformation to OMOP-CDM covers roughly 12% of electronic medical records, 453 databases, and more than 928 million unique patients in 41 countries, creating a foundation for unified governance and distributed analytics anchored in FAIR/CARE and the Five Safes [7].

Statistical implementation of QTL includes SPC control charts, beta-binomial, and Bayesian hierarchical approaches. Practice permits one-sided limits and recommends initiating monitoring after about 30 participants have been enrolled to reduce false alarms [8]. The EHR DQA landscape is characterized by the predominance of completeness checks and a substantial share of correctness, consistency, timeliness, and plausibility assessments. Additional dimensions—conformance and bias—are distinguished, indicating a shift toward structural-conformance checks and identification of systematic distortions [9].

The aim is to analyze integrated validation strategies for clinical systems by synthesizing standardized, multi-level DQA/SSDQA procedures with risk-oriented mechanisms and operational approaches based on eSource technologies, drawing on published qualitative and quantitative indicators to identify reproducible elements of a quality architecture and zones of maximum implementation effect.

Materials and Methods

The methodological foundation of this study is formed at the intersection of risk-based quality management, clinical data standardization and verification procedures, and electronic modes of primary data capture and transfer. As the basic organizational anchor, we used a body of evidence on the actual level of RBQM adoption and key sponsor- and site-side barriers, which defined the contours of integration and expectations around managerial risks, as shown by Dirks [6]. Data-handling rules were organized around a unified model of clinical information and agreed principles of access, protection, and reuse. As reviewed by Hallinan [7], applying the OMOP model in combination with the principles of findability, accessibility, interoperability, and reusability, alongside collective-responsibility guidelines and the Five Safes, provides a unified interpretation of entities, local record storage, and exchange of de-identified aggregates only.

To design acceptable quality limits and a mechanism for early detection of systematic failures, a suite of statistical procedures was adopted: one-sided thresholds; observed–expected and observed-to-expected control graphs; cumulative proportions; Bayesian schemes ranging from beta-binomial to hierarchical models. Serial control is advisable once approximately thirty observations have accrued to reduce the likelihood of early-phase false alarms, as systematized by Kilaru [8]. Constructive descriptions of fields, expected values, and actions for limit excursions were based on the applied framework proposed by Bhagat [2].

The basic data quality verification layer was structured along five traditional dimensions—completeness, correctness, consistency, plausibility, and timeliness—

with the addition of model conformance and systematic bias. For each dimension we used element matching, presence checks, comparisons with external sources, distributional analyses, and audit-log review of entry practices, as summarized by Lewis [9]. Above this layer sits the "fitness for purpose" contour, where verification is tied to a study's target variables, implemented in two rounds—first on aggregated results, then on row-level data—and accompanied by prioritization of detected issues and targeted feedback to sites, as shown by Razzaghi [11].

The observation contour is complemented by centralized statistical monitoring to flag sites and periods with atypical profiles and to route signals into the risk-management cycle, as examined by de Viron [5]. Under constraints on on-site activity, remote forms of centralized monitoring were used with re-focusing as the risk profile changed, as shown by Afroz [1]. Finally, requirements for data provenance transparency and reporting reproducibility were integrated into the methodology as a necessary condition for trust in observational findings, as emphasized by Blacketer [3].

Results

To fix the "starting conditions" for integrated validation, current utilization levels of RBQM components and related tools were compared across life-cycle stages, portfolio scale, and regions. Table 1 presents consolidated indicators derived from a multi-center survey of sponsor companies and contract research organizations.

Table 1 – RBQM adoption (components/tools) by subgroup (Compiled by the author based on source [6])

Metric	Planning & Design	Execution	Documentation	Total (components)	Tools (overall)
All companies	56	52	60	57	46
Annual trial volume <25	47	41	57	48	31
25–100	62	54	62	59	43
≥100	59	63	62	63	50
Europe	64	59	69	64	43
North America	53	50	57	54	41
Rest of the world	42	43	52	45	32

Across all companies, stage imbalance is evident: RBQM component use is higher at the documentation stage (60%) compared with planning (56%) and especially execution (52%), while overall tool use is 46%. This configuration indicates a tilt toward post hoc procedures and substantiates the need to strengthen risk control during active data collection and monitoring, when initial deviations in quality-critical parameters emerge. Portfolio size is a systemic differentiator. With fewer than 25 studies, total RBQM component use is 48% versus 59% for 25–100 and 63% for ≥100; the tools gap reaches 19 percentage points between the extremes (31% versus 50%). The sharpest lag among small organizations is at execution (41% versus 63% in the largest group), pointing to limited maturity of operational control contours and a deficit of tooling at the point where rapid risk response is required [6].

Regional differences are consistent across stages. European samples show higher RBQM component adoption—64% overall versus 54% in North America and

45% elsewhere; similar advantages appear in planning (64% vs 53% and 42%), execution (59% vs 50% and 43%), and documentation (69% vs 57% and 52%). Tool use is also higher in Europe (43%) than in North America (41%) and the rest of the world (32%), which provides a benchmark for adapting practices in jurisdictions with lower values [6].

In sum, the slices presented support three directions relevant to integrated validation. Effort must pivot from the "back-end" documentation contour to continuous support during execution, where risk concentration is maximal. Strategy should account for resource constraints in small portfolios and provide for phased tooling ramp-up without loss of methodological integrity. When transferring solutions across regions, explicit tuning to infrastructural and regulatory contexts is required, as differences are structural rather than incidental.

Evaluation of digital data sources in clinical systems showed clear differences between direct entry into the electronic case report form and automated transfer from the clinical information system. As shown by Yaegashi [12], direct entry shortens the time to final status by several days and reduces the time burden of on-site monitoring due to simplified source-data reconciliation. In addition, Mueller [10] reports that automated transfer from the medical system to the data collection system enables regular synchronization and partially closes the gap between routine practice and research variable requirements. Table 2 summarizes metrics for both approaches, reflecting the share of fields, speed to final status, monitoring burden, and the volume of automatically transferable variables.

Table 2 – Key eSource metrics (DDC and EHR2EDC) (Compiled by the author based on sources [10,12])

Metric	Value
Share of DDC fields across sites	61.9–84.5%
Same-day entry	76% DDC vs 72% non-DDC (median = 0 days in both)
Event → finalization (median)	24 DDC vs 28 non-DDC days
Entry → finalization (median)	22 DDC vs 27 non-DDC days
CRA visit duration	43 vs 52 min/visit (–9 min)
Time saving per subject	≈8.6 hours (57 visits × 9 min)
Effort break-even threshold	2–13 subjects/site
Auto-transferable variables (EHR2EDC)	67 of 274 (24% of all; 36% of eligible)

These values show that direct case-form entry confers an advantage specifically on the "entry–finalization" segment: the five-day median reduction indicates fewer review iterations and elimination of manual transcription, whereas the moment of initial entry remains synchronized to the clinical event in both groups. The nine-minute reduction in monitor-visit duration accumulates across repeated visits; at 57 visits, the saving reaches roughly 8.6 hours per participant, directly affecting on-site workload and monitoring resource planning [12]. The effort break-even threshold ranges from two to thirteen participants per site and depends on the share of fields entered directly into the case form, which requires design-time tuning of forms and documentation roles during site preparation [3].

Automated transfer from the clinical system to the data

collection system exhibits a different efficiency profile: a strictly bounded yet stable pool of automatically transferable variables (67 of 274; 24% of all fields and 36% of those deemed eligible) with daily synchronization establishes a repeatable data flow without manual copying and ensures traceability via predefined mapping and security rules [10]. This contour requires regulated interfaces and dictionaries and relies on access-governance procedures and role alignment within clinical practice.

A comparison of approaches shows that direct entry scales gains by shortening the review–correction cycle—especially when a high share of fields is assigned to the coordinator—whereas automated transfer strengthens repeatability for a tightly defined subset of variables and reduces transcription risk in operational systems [4]. The sustainability of both approaches requires uniform schemas and governance policies, as confirmed by findings on the importance of a unified data model and responsible-access principles in integrated clinical data management [7]. At the quality-control level, adding a data quality layer with fixed dimensions—completeness, correctness, consistency, plausibility, timeliness, structural conformance, and systematic bias—enables early identification of vulnerabilities in direct-entry and auto-transfer flows [9]. In addition, fitness-for-purpose procedures identify and document anomalies affecting endpoint calculations and support prioritization of fixes by analytical impact [11]. Site readiness for digital-source deployment and practices for initial infrastructure setup remain necessary conditions for achieving these metrics and call for formal change management at start-up.

Discussion

The starting point for integration is unifying structures and access rules. According to Hallinan [7], broad implementation of a common OMOP-CDM (about 12% of electronic records, 453 databases, over 928 million patients in 41 countries), coupled with FAIR/CARE and the Five Safes, creates a reproducible basis for quality checks and secure distributed analytics. This foundation enables subsequent layered validation, where unified vocabularies and structures lower the risk of discrepancies during data transfer and reconciliation.

The next layer is quality assessment. Lewis [9] shows that publications most often test completeness (74% of studies), with growing attention to structural conformance and systematic bias; this shift reflects the need to control format/structural errors and identify non-random missingness distortions. At the study level, extended SSDQA can surface and close analysis-critical gaps before statistical computations. Razzaghi [11] demonstrates two sequential verification rounds with prioritized issue tracking and traceable quality improvements at sites. The integrated scheme then incorporates quality limits and central statistical control. As shown by Kilaru [8], quality limits are tuned using statistical control charts and beta-binomial and hierarchical Bayesian models. One-sided bounds and initiation of operational monitoring after roughly thirty participants are justified to enhance sensitivity to systematic shifts at an acceptable false-alarm rate. De Viron [5] reviews scaling of central statistical monitoring to large collections of sites, confirming applicability in multi-site programs. Afroz [1] shows that remote centralized monitoring practices were successfully employed under constraints, supporting the case for a persistent remote-control contour.

Completing the chain is the operational layer of digital sources. Yaegashi [12] shows that direct entry into eCRFs accelerates time to final status and reduces on-site monitoring burden, with effects accumulating with the number of visits. Mueller [10] shows that automated clinical-to-EDC transfer eliminates double entry for standardizable variables and ensures regular synchronization via defined mappings, complementing direct entry and enhancing reproducibility for a fixed subset of measures. Finally, Dirks [6] shows that RBQM implementation remains uneven by region and portfolio scale, setting the “starting conditions” for integrating the components above and requiring stepwise practice alignment. Table 3 presents the outcomes of two-round data fitness checks in a multi-center study, illustrating how the SSDQA layer operationalizes the results of prior standardization.

Table 3 – Two-round SSDQA results in PRESERVE (Compiled by the author based on source [11])

Parameter	DQ1 (aggregated, distributed)	DQ2 (row-level, centralized)

Number of checks	79	65
Issues identified	115	157
Priority distribution	Urgent 9%; High 23%; Medium 23%; Low 45%	Urgent 5%; High 46%; Medium 47%; Low 2%
Resolved / improved	50% issues (ETL corrections, additional data)	34% improvements at sites; substantial share due to source-data specifics (~38%)
Typical themes	Eligibility criteria, missing key variables, and code-use variability	Heterogeneity of clinical values/trajectories, event-sequence anomalies, and geodata

These results show that distributed, aggregate-level checks rapidly capture critical breaks in eligibility and key variables, while subsequent row-level work reveals subtler anomalies in values and event sequences—thereby influencing refinements to outcome and covariate definitions. Such "stepped" problem revelation justifies introducing quality limits and central statistical checks and indicates where digital operational flows—direct entry and automated transfer—will yield the greatest gains once structures and access rules have been unified.

Interpreting the results requires accounting for source applicability and context differences. The works by Afroz [1], Bhagat [2], Blacketer [3], Cramer [4], and de Viron [5] serve as conceptual anchors for centralized control, quality tolerance limits, trust in real-world data, and eSource start-up practices, but lack quantitative detail, limiting direct cross-approach comparisons.

eSource portability is heterogeneous because examples come from different jurisdictions and regulatory contours. In the German auto-transfer case, daily synchronization and 67 automatically transferable variables (24% of all fields and 36% of those eligible) are shown [10]. In the Japanese direct-entry study, time to final status is accelerated by four–five days, monitor visit duration is reduced by nine minutes, and labor break-even thresholds are estimated at two–thirteen subjects given a 61.9–84.5% direct-entry field share [12].

Some discrepancies originate at the source and are not resolved by extraction/transformation adjustments alone. Changes to the analysis plan and clarification of variable definitions are required, as shown in the two-stage data-fitness review in the multi-center project [11]. Meanwhile, the DQA review highlights the absence of a universally accepted standard amid ongoing automation and expansion of measurement dimensions, including structural conformance and systematic bias [9]. Unification based on a shared data model and responsible information principles reduces variability but does not eliminate gaps [7].

Methodological boundaries are also evident in tuning limit controls and central statistical surveillance. Limit methods differ in effectiveness, and initiating operational monitoring is advisable after about thirty participants [8]. The multi-level heterogeneity of RBQM adoption by region and portfolio scale sets different starting positions for integrating standardization, quality checks, threshold limits, and digital sources.

Conclusion

This study confirms the critical role of integrated validation strategies as the basis for sustainable clinical data quality management. The greatest effect is achieved when a standardized data model, multi-level quality assessment procedures, and risk-oriented operational controls are connected sequentially, minimizing the likelihood of systematic and random distortions early in the study life cycle.

The optimal validation configuration shifts emphasis from post hoc documentation to continuous support during execution, when risk concentration is highest. In this context, digital sources—direct entry and automated transfer—provide complementary effects:

the former shortens the review–correction cycle; the latter enhances reproducibility and lowers transcription burden. The effectiveness of each approach is determined by the share of covered fields, infrastructure maturity, and role distribution among process participants.

Adapting tools to portfolio scale and regional contexts is especially important. Differences in RBQM maturity and digital infrastructure availability require tailoring the validation architecture to specific regulatory and operational conditions. Phased implementation, beginning with priority elements, enables managed efficiency gains without loss of methodological integrity.

Accordingly, sustainable development of integrated validation strategies requires a unified methodological base, a combination of operational flexibility and technological standardization, and institutional readiness for inter-site and inter-regional harmonization. Future research should focus on forming universal effectiveness metrics, refining mechanisms for automating data-fitness assessment, and integrating validation procedures into end-to-end risk management systems for clinical research.

References

1. Afroz, M. A., Schwarber, G., & Bhuiyan, M. A. N. (2021). Risk-based centralized data monitoring of clinical trials during the time of COVID-19 pandemic. Contemporary Clinical Trials, 104, 106368. https://doi.org/10.1016/j.cct.2021.106368
2. Bhagat, R., Bojarski, L., Chevalier, S., et al. (2021). Quality tolerance limits: Framework for successful implementation in clinical development. Therapeutic Innovation & Regulatory Science, 55, 251–261. https://doi.org/10.1007/s43441-020-00209-0
3. Blacketer, C., Defalco, F. J., Ryan, P. B., & Rijnbeek, P. R. (2021). Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association, 28(10), 2251–2257. https://doi.org/10.1093/jamia/ocab132
4. Cramer, A. E., King, L. S., Buckley, M. T., Casteleyn, P., Ennis, C., Hamidi, M., Rodrigues, G. M. C., Snyder, D. C., Vattikola, A., & Eisenstein, E. L. (2024). Defining methods to improve eSource site start-up practices. Contemporary Clinical Trials Communications, 42, 101391. https://doi.org/10.1016/j.conctc.2024.101391
5. de Viron, S., Trotta, L., Steijn, W., Young, S., & Buyse, M. (2024). Does central statistical monitoring improve data quality? An analysis of 1,111 sites in 159 clinical trials. Therapeutic Innovation & Regulatory Science, 58(3), 483–494. https://doi.org/10.1007/s43441-024-00613-w
6. Dirks, A., Florez, M., Torche, F., Young, S., Slizgi, B., & Getz, K. (2024). Comprehensive assessment of risk-based quality management adoption in clinical trials. Therapeutic Innovation & Regulatory Science, 58(3), 520–527. https://doi.org/10.1007/s43441-024-00618-5
7. Hallinan, C. M., Ward, R., Hart, G. K., Sullivan, C., Pratt, N., Ng, A. P., Capurro, D., Van Der Vegt, A., Liaw, S. T., Daly, O., Luxan, B. G., Bunker, D., & Boyle, D. (2024). Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM. BMJ Health & Care Informatics, 31(1), e100953. https://doi.org/10.1136/bmjhci-2023-100953
8. Kilaru, R., Amodio, S., Li, Y., et al. (2024). An overview of current statistical methods for implementing quality tolerance limits. Therapeutic Innovation & Regulatory Science, 58, 273–284. https://doi.org/10.1007/s43441-023-00598-y
9. Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R. O., & Gupta, A. (2023). Electronic health record data quality assessment and tools: A systematic review. Journal of the American Medical Informatics Association, 30(10), 1730–1740. https://doi.org/10.1093/jamia/ocad120
10. Mueller, C., Herrmann, P., Cichos, S., Remes, B., Junker, E., Hastenteufel, T., & Mundhenke, M. (2023). Automated electronic health record to electronic data capture transfer in clinical studies in the German health care system: Feasibility study and gap analysis. Journal of Medical Internet Research, 25, e47958. https://doi.org/10.2196/47958
11. Razzaghi, H., Goodwin Davies, A., Boss, S., Bunnell, H. T., Chen, Y., Chrischilles, E. A., Dickinson, K., Hanauer, D., Huang, Y., Ilunga, K. T. S., Katsoufis, C., Lehmann, H., Lemas, D. J., … Bailey, L. C. (2024). Systematic data quality assessment of electronic health record data to evaluate study-specific fitness:

Report from the PRESERVE research study. PLOS Digital Health, 3(6), e0000527. https://doi.org/10.1371/journal.pdig.0000527

12. Yaegashi, H., Hayashi, Y., Takeda, M., et al. (2024). Efficiency of eSource direct data capture in investigator-initiated clinical trials in oncology. Therapeutic Innovation & Regulatory Science, 58, 1031–1041. https://doi.org/10.1007/s43441-024-00671-0