Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 05| Issue- 07| 2025

Research Article **Received:** 25-05-2025 | **Accepted:** 15-06-2025 | **Published:** 02-07-2025

Evaluating Data Fitness for AI in Pharmaceutical Applications: The PAIR Framework

Sravish Nalam
BITS, Pilani, India

Abstract: When pharmaceutical organizations utilize artificial intelligence in their operations, they must have datasets that are of sufficient quality and structure, yet there are no standardized means to assess data preparedness in practice. The Pharma-AI Readiness (PAIR) Framework addresses this shortfall with a quantitative assessment framework to assess datasets in the pharmaceutical space across five dimensions: quality, consistency, structure, compliance, and documentation. Each dimension contributes to an overall readiness score, which categorizes datasets for specific use based on readiness. This provides organizations a means to assess what datasets can be deployed immediately with a preferred level of preparation, to define what datasets require more extensive preparation before use, and to prioritize efforts in data preparation. Using the framework with clinical trial datasets, real-world evidence datasets, and commercial datasets showed that organizations have systematic readiness challenges, including the lack of consistent temporal data capture in clinical sources and inconsistent metadata standardization in commercial datasets. The framework enables cross-functional teams and business units to engage with one another based on the same set of criteria when it comes to assessing fitness for data usage, thus shortening timelines from concept to model competition and implementing the desired artificial intelligence. Once the PAIR Framework is implemented, pharmaceutical organizations gain clarity for their data governance strategy and can ensure regulatory compliance while increasing operational scalability to support digital initiatives. The structure and scoring within the PAIR framework allow benchmarking across different areas of data assets and enable evidence-based decision-making regarding an organization's AI transformation strategies.

Keywords: pharmaceutical data, AI readiness, data quality assessment, drug discovery, healthcare analytics.

INTRODUCTION

Current Landscape of AI Adoption in the Pharmaceutical Industry

The pharmaceutical sector is at a pivotal moment where AI technology is set to radically change approaches in drug discovery, clinical development, and commercialization. Recent studies suggest that the uptake of AI in the pharmaceutical context is moving quickly, with current applications of AI ranging from molecular design and target identification to patient stratification and market forecasting (Sharma, A. K. 2023). This encompasses the entire pharmaceutical value chain, from the early stages of research to post-market surveillance, changing in fundamental ways how we develop and commercialize drugs (Sharma, A. K. 2023).

Critical Role of Data Quality in AI Implementation Success

The issues of underlying data infrastructure and data quality are thoroughly critical for the success of AI deployments in pharmaceutical contexts with patient interaction. In this regard, data quality is the bedrock of any AI strategy, especially in industries with strict and highly regulated practices, where poor decision-making carries major consequences for patient safety and public health (Jones, A. 2024). Without a set of reference datasets that are sufficiently resourceful, well-formatted, and appropriately curated, even the most intelligent AI algorithms fail to provide solid insights or useful predictions. The relationship between data quality and AI performance has further significance in pharmaceutical contexts, where accuracy will impact patient outcomes and compliance with regulatory authorities and standards.

Challenges in Assessing Pharmaceutical Data Readiness for AI Applications

Pharmaceutical companies have unique considerations when creating datasets for AI applications, as they contend with heterogeneous data sources, including clinical trials, real-world evidence, and commercial operations. Additionally, firms must consider tight regulatory requirements for data usage and handling and handling (e.g., electronic records), as well as temporal challenges of longitudinal patient data (e.g., collections across time). There are also significant technical and organizational challenges in harmonizing structured and unstructured data from many disparate systems. Moreover, there are no standard practices for assessing whether pharmaceutical datasets are fit-for-purpose for AI, which imposes meaningful challenges to successful AI deployment.

Need for Standardized Evaluation Framework

Our current practices aren't built to assess whether the pharmaceutical data we collect is ready to support AI production-level applications, as there are no established frameworks we can use to assess data readiness in the context of AI

***Corresponding Author:** Sravish Nalam

DOI- https://doi.org/10.5281/zenodo.15793740

deployment, and there are many quality assessment frameworks available, none account for the requirements of AI in the pharmaceutical industry; including the inclusion of the requirements of external stakeholders (e.g., regulatory bodies), time and consistency of data, heterogeneity of data modality (e.g., biomedical literature, compound metadata, patient data, claims data etc.). We need to establish a survey framework that fits the similar characteristics of a pharmaceutical data ecosystem. The industry needs a usable framework to evaluate technical properties and data requirements specific to its domain.

Research Objectives and Contributions

The Pharma-AI Readiness (PAIR) Framework satisfies this need by providing a standard, systematic, and quantitative method for assessing AI readiness based on pharmaceutical datasets. The framework has five assessment dimensions—quality, consistency, structure, compliance, and documentation—that have been deliberately developed to capture the vital elements of data fitness for AI. By providing criteria for evaluation and scoring mechanisms, PAIR empowers organizations to systematically assess their readiness gaps and direct resources to data preparation. The framework provides a contribution to the field by establishing the first comprehensive and industry-specific method to quantify the concept of AI readiness in a pharmaceutical context.

Paper Organization

In this article, we have outlined the development, validation, and application of the PAIR Framework to a wide range of pharmaceutical datasets. Section 2 highlights relevant background and existing strategies to assess data quality in both the pharmaceutical and healthcare contexts. In section 3, we describe the structure of the PAIR Framework and scoring procedures in more detail. Section 4 demonstrates the application of the framework with example dataset evaluations from clinical, real-world evidence, and commercial populations. In section 5, we increase the discussion of insights and implications for pharmaceutical companies developing AI projects. Finally, in section 6, we provide recommendations for using the framework and areas for further enhancing the ability to assess AI readiness in the life sciences.

BACKGROUND AND RELATED WORK

Evolution of AI in Pharmaceutical Research and Operations

The pharmaceutical sector has experienced a paradigm shift in research and development practices with the implementation of artificial intelligence technologies. This shift encompasses multiple areas of pharmaceutical operations, changing how organizations conduct drug development, clinical research, and marketing (Suriyaamporn, P. 2024). The shift is more than technology adoption; it represents a complete rethinking of pharmaceutical workflows and decision-making.

Drug Discovery Applications

Artificial intelligence has transformed drug discovery from a time-consuming and costly selection process to reliable molecular screening, identification of drug targets, and lead optimization. In recent years, machine learning (ML) algorithms have been developed to utilize large compound libraries to quickly predict areas of drug-target interaction, passive permeability properties, and hazardous side effects associated with the potential of a new drug (Suriyaamporn, P. 2024). These implementations have shortened drug development timelines and reduced costs associated with resource selection, which have been ultimate bottlenecks in the drug discovery process. Whereas ML models were effective in predicting molecular properties and identifying new drug candidates that could not have been found with traditional molecular library screening, deep learning (DL) models used advanced artificial intelligence downstream efforts, and new models can design whole new molecular leads to achieve a specific drug goal beyond molecular pattern recognition.

Table 1: Evolution of AI Applications in Pharmaceutical Industry (Suriyaamporn, P. 2024)

Application Domain	Traditional Approach	AI-Enhanced Approach	Key Benefits
Target Identification	Literature review, experimental screening	ML-based protein interaction prediction	Reduced screening time, novel target discovery
Lead Optimization	Trial-and-error synthesis	Deep learning molecular design	Faster optimization cycles, improved success rates
Clinical Trial	Fixed protocols, historical	Adaptive designs,	Enhanced patient stratification,

Publisher: SARC Publisher

Design	statistics	predictive analytics	reduced trial duration
Safety Monitoring	Manual adverse event review	Automated signal detection	Earlier safety signal identification
Market Analytics	Historical sales analysis	Predictive demand forecasting	Improved resource allocation, market responsiveness

Clinical Trial Optimization

AI-based solutions have fundamentally changed clinical trial design and conduct. These technologies enhance patient recruitment by identifying eligible patients among electronic health records and real-world data sources. AI can also develop trial designs and protocols, estimate patient dropout rates, and identify safety signals earlier in the development lifecycle (Suriyaamporn, P. 2024). Machine learning has enabled us to analyze clinical trial data and improve patient stratification, which leads to better and more targeted clinical trial designs and proposals. AI also supports adaptive trial designs that allow changes to trial protocols based on interim analyses that can reduce the timeline or improve the chances of success.

Commercial Analytics

For commercial operations in the pharmaceutical industry, the AI infusion has been just as valuable. Predictive analytics using algorithms can help project market demand, optimize pricing, and determine high-value prescriber segments. Systems powered by AI are now "learning" from real-world evidence to understand treatments, patient outcomes, and market reaction patterns to both treatments and the broader market (Suriyaamporn, P. 2024). By using these tools, the pharmaceutical industry is able to make more informed decisions with respect to how to allocate its resources, develop marketing strategies, and manage products along the product life cycle. No longer confined to only traditional analytics on claims data, structured electronic health records, and social media, now also adds to that picture by displaying transaction-level data, patient referrals, marketing, and loyalty levels, leading to never-before-attained comprehensive market intelligence.

EXISTING DATA QUALITY ASSESSMENT FRAMEWORKS

General Data Quality Dimensions

Traditional data quality frameworks have established fundamental dimensions for assessing how well data fits for use across applications. Traditional data frameworks most commonly use quality aspects of completeness, accuracy, consistency, timeliness, validity, and uniqueness as quality elements (Cichy, C., & Rass, S. 2019). Each dimension corresponds to aspects of data reliability and usability to form a comprehensive assessment. However, these general frameworks often lack appropriate specificity when applied to specific fields or domains such as those in pharmaceutical R&D.

Healthcare-Specific Data Standards

Healthcare data standards have evolved to address the unique requirements of medical information systems. These standards encompass data interoperability protocols, clinical terminology systems, and quality metrics tailored to healthcare contexts (Cichy, C., & Rass, S. 2019). Organizations have developed specialized frameworks for electronic health records, clinical research data, and health information exchange. Despite these advances, existing healthcare data standards primarily focus on operational data management rather than AI readiness assessment.

Limitations of Current Approaches

Current data quality frameworks exhibit several limitations when applied to pharmaceutical AI applications. Most frameworks lack specific criteria for evaluating temporal consistency in longitudinal patient data, fail to address the integration requirements for heterogeneous pharmaceutical data sources, and provide insufficient guidance for assessing regulatory compliance in AI contexts (Cichy, C., & Rass, S. 2019). Additionally, existing frameworks rarely consider the unique pre-processing requirements of pharmaceutical data for machine learning applications. These limitations create significant gaps between available assessment tools and the practical needs of pharmaceutical organizations implementing AI solutions.

UNIQUE REQUIREMENTS OF PHARMACEUTICAL DATA FOR AI

Regulatory Compliance Considerations

Pharmaceutical data must follow rigorous regulatory requirements that are more comprehensive than any standard data governance constraints. They include Good Clinical Practice (GCP) guidelines, declared data integrity principles for regulatory agencies, and there are also requirements specific to documentation standards for the AI-based decision support

Publisher: SARC Publisher

systems. The regulatory environment continues to develop as regulatory authorities continue the work of turning (and freezing it), and we slowly synchronize the timing of frameworks for AI validation by the pharmaceutical community. It's a challenge for organizations to make sure their data preparation and AI protocols can demonstrate auditability, reproducibility, and fit for regulatory submission.

Data Heterogeneity Challenges

The pharmaceutical companies hold extremely disparate types of data: structured clinical trial data, unstructured physician notes, molecular structures, imaging data, and real-world evidence from a variety of sources. The variety of data held by these organizations creates distinct challenges to implementing AI that will require nuanced approaches to integration and harmonization. The challenge increases when we factor in the data that integrates preclinical, clinical, and post-market worlds. Each type of data has different features, different quality/risk issues, and different preprocessing steps that require systematic consideration.

Temporal and Longitudinal Aspects

Pharmaceutical data often involves complex temporal relationships and longitudinal patient observations spanning extended periods. These temporal aspects are critical for understanding drug efficacy, safety profiles, and disease progression patterns. AI models must account for varying observation frequencies, missing data patterns in longitudinal studies, and the evolution of data collection methodologies over time. The temporal complexity extends to considerations of data currency, where the relevance and applicability of historical data must be carefully evaluated for contemporary AI applications.

THE PAIR FRAMEWORK: METHODOLOGY AND DESIGN

Framework Architecture and Components

The Pharma-AI Readiness (PAIR) Framework represents a comprehensive methodology designed specifically for evaluating pharmaceutical datasets' readiness for artificial intelligence applications. The framework architecture draws inspiration from established quality assessment models while incorporating pharmaceutical-specific requirements and AI deployment considerations (Laurentino Schwabe, D. 2024). The modular design enables systematic evaluation across diverse pharmaceutical data types, from clinical trial databases to real-world evidence repositories and commercial datasets. Each component of the framework addresses critical aspects of data fitness that directly impact AI model performance and reliability in pharmaceutical contexts. The framework's architecture consists of hierarchical assessment layers that progressively evaluate data characteristics from fundamental quality attributes to complex AI-specific requirements. This layered approach ensures comprehensive coverage while maintaining practical applicability across various organizational contexts and data maturity levels. The framework incorporates lessons learned from existing medical AI data quality initiatives while extending beyond traditional metrics to address pharmaceutical industry-specific challenges (Marandi, R. Z. *et al.*, 2025).

FIVE CORE ASSESSMENT DIMENSIONS

Data Quality Metrics and Evaluation Criteria

The data quality dimension encompasses fundamental attributes essential for reliable AI applications in pharmaceutical contexts. This dimension evaluates the completeness of critical data fields, the accuracy of recorded values against source documentation, and the validity of data entries according to pharmaceutical domain constraints. The framework extends traditional quality metrics by incorporating pharmaceutical-specific criteria such as protocol adherence in clinical data and regulatory compliance in safety reporting (Laurentino Schwabe, D. 2024). Quality assessment considers both individual data point integrity and aggregate quality patterns that may impact AI model training and validation.

The evaluation criteria within this dimension address common quality issues in pharmaceutical datasets, including missing value patterns, data entry errors, and inconsistencies between related fields. Special attention is given to quality aspects that directly influence AI model performance, such as class imbalance in outcome variables and representation adequacy across patient subpopulations. The framework provides structured methods for identifying and quantifying quality issues that could compromise AI reliability in pharmaceutical applications.

Consistency Assessment across Datasets

Consistency evaluation focuses on harmonization and standardization across different data sources and temporal periods. This dimension becomes particularly critical in pharmaceutical contexts where data often originates from multiple clinical sites, different electronic systems, and varying

Publisher: SARC Publisher

collection protocols (Marandi, R. Z. *et al.*, 2025). The framework assesses consistency in terminology usage, measurement units, coding standards, and data capture procedures across integrated datasets. The consistency dimension also includes temporal consistency challenges posed by longitudinal pharmaceutical datasets, including changes in data collection methods over time, differences in clinical evaluation times, and consistency in adverse event reporting. This dimension provides a means to identify and document consistency variations that may impact the generalizability and stability of AI model performance.

Structure and Format Standardization

Structure and format standardization evaluates the organization and technical characteristics of pharmaceutical datasets for AI readiness. This dimension assesses adherence to industry-standard data models, compatibility with common AI pre-processing pipelines, and the presence of structured metadata supporting automated processing (Laurentino Schwabe, D. 2024). The evaluation considers both technical format requirements and the semantic structure necessary for meaningful AI analysis. The framework acknowledges the unique structural nature of pharmaceutical data and the requirements to assess the structural nature of data, particularly the ability to assess structural integration, the integration of structured clinical data with unstructured physician notes, and structured imaging data with structured molecular data. Assessment criteria included evaluating if there are criteria for schema consistency, if appropriate relational structures exist, and if data exchange formats implement industry standard pharmaceutical data exchange schemas and structures. In particular, structural components that support feature engineering and model building can be given special priority.

Compliance and Regulatory Alignment

The compliance dimension assures that datasets comply with pharmaceutical regulatory requirements for AI uses. This consideration goes beyond typical data privacy to include consideration of Good Clinical Practice (GCP) compliance, audit trail completeness, and compliance with opening AI-specific regulatory guidelines (Marandi, R. Z. *et al.*, 2025).. The framework is evaluated for whether data collection, processing, and storage practices comply with regulatory submissions and post-market surveillance commitments. Regulatory alignment assessment includes verification of patient consent appropriate for AI applications, data provenance documentation, and compliance with international data protection regulations. The framework provides specific criteria for evaluating whether datasets maintain the necessary documentation and controls required for AI models used in regulatory decision-making contexts. This dimension ensures that AI-ready datasets also meet the stringent compliance requirements of the pharmaceutical industry.

Documentation Completeness and Accessibility

Documentation assessment evaluates the availability and quality of metadata, data dictionaries, and supporting documentation necessary for appropriate AI model development and validation. This dimension recognizes that comprehensive documentation is essential for understanding data context, limitations, and appropriate use cases in pharmaceutical AI applications (Marandi, R. Z. *et al.*, 2025). The framework assesses both technical documentation and domain-specific information required for meaningful AI implementation.

The evaluation process includes a description of data collection procedures, variable definitions, identified data quality issues, and documentation of any transformations made to the original data. Evaluation focuses on the documentation necessary to support reproducibility in AI, including data versioning, any preprocessing steps, and any exclusion criteria undertaken. This form ensures that datasets do not lack sufficient documentation for independent verification and acceptable use in AI model development.

Table 2: PAIR Framework Assessment Dimensions and Criteria (Laurentino Schwabe, D. 2024; Marandi, R. Z. *et al.*, 2025)

Dimension	Primary Assessment Criteria	Key Evaluation Points	AI Impact Areas
Data Quality	Completeness, Accuracy, Validity	Missing patterns, Error rates, Domain constraints	Model training reliability
Consistency	Cross-source harmony, Temporal stability	Terminology alignment, Measurement standardization	Model generalizability
Structure	Schema compliance, Format standardization	Relational integrity, Integration readiness	Feature engineering efficiency

Publisher: SARC Publisher

Compliance	Regulatory adherence, Audit trails	GCP compliance, Privacy protection	Deployment feasibility
Documentation	Metadata completeness, Accessibility	Data lineage, Known limitations	Appropriate use assurance

SCORING METHODOLOGY

Individual Dimension Scoring

Each dimension within the PAIR Framework receives an individual score based on detailed evaluation criteria specific to that dimension. The scoring system employs a standardized scale that enables consistent assessment across different dataset types and organizational contexts. Individual dimension scores reflect the degree to which a dataset meets established criteria for AI readiness within that specific aspect. The granular scoring approach allows organizations to identify specific areas requiring improvement while recognizing existing strengths.

Weighted Composite Score Calculation

The framework calculates a composite readiness score by combining individual dimension scores through a weighted aggregation method. Weighting factors reflect the relative importance of each dimension for specific AI use cases in pharmaceutical contexts. The composite score provides a single metric for overall AI readiness while maintaining transparency about performance across individual dimensions. The weighting scheme can be adjusted based on specific organizational priorities or use case requirements while maintaining methodological consistency.

Readiness Classification Thresholds

The PAIR Framework establishes classification thresholds that categorize datasets into distinct readiness levels for AI deployment. These classifications provide actionable guidance for data governance teams and AI developers regarding dataset suitability for different types of AI applications. The classification system considers both composite scores and minimum performance requirements across critical dimensions to ensure a comprehensive readiness assessment. This approach prevents high scores in some dimensions from masking critical deficiencies in others that could compromise AI reliability.

Implementation Workflow and Assessment Process

The framework implementation follows a structured workflow designed for practical application in pharmaceutical organizations. The process begins with dataset scoping and stakeholder engagement to ensure an appropriate evaluation context and objectives. Assessment teams then systematically evaluate each dimension using standardized evaluation tools and criteria provided by the framework. The workflow incorporates quality control checkpoints and validation steps to ensure assessment reliability and consistency. The evaluation process emphasizes the responsibility of data stewards, domain experts, and AI practitioners to work together to adequately evaluate the dataset from multiple perspectives. Findings from the evaluation, gaps in readiness, and recommendations for remediation must be documented and are part of the implementation process. The framework includes templates and tools to assist different organizational units and types of datasets in implementing evaluation uniformly. Consistent reassessment procedures ensure that readiness scores remain timely as datasets are updated and AI needs change.

FRAMEWORK APPLICATION AND CASE STUDIES

Dataset Selection and Characteristics

The practical validation of the PAIR Framework involved systematic application across diverse pharmaceutical datasets representing key data categories utilized in AI implementations. The selection process prioritized datasets that exemplify common challenges and characteristics encountered in pharmaceutical AI initiatives, ensuring comprehensive framework validation across different use contexts and data complexities.

Table 3: Dataset Characteristics for Framework Validation (IEEE Journal of Biomedical and Health Informatics. 2022)

Dataset Type	Data Sources	Volume Range	Temporal Span	Primary Challenges
Clinical Trials	EDC systems, Lab results, CRFs	Thousands to millions of records	Months to years	Protocol variations, Missing data
Real-World Evidence	EHRs, Claims, Registries	Millions of records	Years to decades	Heterogeneity, Incomplete capture
Commercial	Prescriptions,	Hundreds of	Monthly to	Aggregation levels,

Publisher: SARC Publisher

Data	Market research	thousands of records	quarterly updates	Proprietary methods

Clinical Trial Datasets

Clinical trial datasets are a very structured and protocol-based type of data collection that represents the foundation of the pharmaceutical research and development pipeline. Clinical trial datasets include patient demographics and medical history, treatment assignment, laboratory values, adverse events, and efficacy endpoints, and are abstracted from data collected under controlled conditions. The clinical trial datasets that we applied the framework to included Phase II and Phase III studies from different therapeutic areas, including completed trials and ongoing studies. These datasets illustrate the rigorous data collection standards of clinical research but come with a unique set of challenges, such as missing data due to patient dropout, divergence from the protocol, and site-to-site variability based on site-specific data collection practices.

The clinical trial datasets evaluated varied in scope from single-site studies to large multinational trials, which included hundreds of investigational sites. Each dataset had structured data elements specified by study protocols and case report forms based on regulatory requirements. The temporal nature of clinical trial data, with scheduled visits and assessments over prolonged periods of time, provided important information regarding the capabilities of the framework to determine the asset's readiness for AI to analyze longitudinal data.

Real-World Evidence (RWE) Sources

'Real-world evidence' sources are grounded in data collected from outside controlled clinical trial settings, which afford insight into effectiveness, safety, and patterns of use in real-world clinical practice (IEEE Journal of Biomedical and Health Informatics. 2022). The RWE datasets we selected as we applied the framework included electronic health records from healthcare systems, insurance claims datasets, patient registries, and pharmacy dispensing records. These sources had specific challenges that may differ from clinical trial data: more heterogeneity in data collection practices, various degrees of completeness of data, and no standard protocols for data capture.

Retrospective and prospective data collections were included in the RWE datasets evaluated, each with a different level of standardization. These datasets capture the reality of healthcare delivery, capturing data like imprecise coding, absent diagnostic details, and difficulties integrating patient records from multiple healthcare facilities (IEEE Journal of Biomedical and Health Informatics. 2022). The heterogeneity of primary sources from RWE offered rich evaluation opportunities to measure the capability of the proposed framework to evaluate AI readiness showcases in diverse real-world scenarios encountered in pharmaceutical practices.

Commercial and Market Data

Commercial and market datasets represent business-critical information used for market analysis, sales forecasting, and strategic planning in pharmaceutical organizations. The selected commercial datasets included prescription data, market share information, physician prescribing patterns, and patient demographic analytics. These datasets often combine internal company data with third-party market intelligence sources, creating unique integration and standardization challenges.

Commercial datasets evaluated through the framework exhibited characteristics distinct from clinical and RWE sources, including aggregated data formats, proprietary coding systems, and varying temporal granularities. The assessment of these datasets tested the framework's versatility in addressing business-oriented AI applications while maintaining rigorous evaluation standards. Market data often lacks the detailed documentation typical of clinical datasets, presenting specific challenges for documentation completeness assessment.

ASSESSMENT RESULTS BY DATASET TYPE

Dimension-wise Scoring Analysis

The application of the PAIR Framework revealed distinct scoring patterns across the five assessment dimensions for each dataset category. Clinical trial datasets consistently demonstrated high scores in the compliance and regulatory alignment dimension, reflecting the stringent controls governing clinical research data collection. However, these same datasets showed variable performance in the consistency dimension, particularly when integrating data from multiple trial sites or legacy studies with evolving data standards.

Real-world evidence sources exhibited contrasting patterns, with generally lower scores in data quality metrics due to the uncontrolled nature of routine clinical data capture (IEEE Journal of Biomedical and Health Informatics. 2022). The

Publisher: SARC Publisher

structure and format standardization scores for RWE varied significantly based on the source system, with modern electronic health record systems scoring higher than legacy claims databases. Documentation completeness emerged as a particular challenge for RWE sources, where data dictionaries and collection protocols often lack the detail found in clinical trial documentation. Commercial datasets demonstrated unique scoring profiles, with high performance in structure and format standardization due to established industry data exchange standards. However, these datasets frequently scored lower in documentation completeness, particularly regarding data lineage and transformation processes. The consistency dimension revealed challenges in harmonizing data from multiple commercial sources with varying update frequencies and geographic coverage.

Composite Readiness Scores

The weighted composite scores calculated through the framework provided comprehensive insights into overall AI readiness across dataset types. Clinical trial datasets achieved the highest average composite scores, driven by strong performance in compliance and data quality dimensions. However, the range of scores within clinical datasets was substantial, with legacy trials and observational studies scoring significantly lower than recent interventional trials designed with digital data capture systems. Real-world evidence sources displayed moderate composite scores with high variability, reflecting the diverse nature of these data sources (IEEE Journal of Biomedical and Health Informatics. 2022). Electronic health record data from integrated healthcare systems achieved higher composite scores than fragmented claims databases or registry data with voluntary reporting mechanisms. The composite scoring revealed that RWE readiness for AI applications depends heavily on the specific source system and data governance practices of contributing organizations. Commercial datasets achieved moderate to high composite scores, with performance largely determined by the maturity of data management practices within specific market data categories. Prescription tracking systems with established data standards scored highest, while newer data sources such as digital health applications and patient-reported outcomes showed lower readiness levels due to evolving standards and limited historical data.

Readiness Classification Outcomes

The classification of datasets into readiness categories revealed that approximately one-third of evaluated clinical trial datasets achieved high AI readiness classification, with the remainder split between moderate and low readiness categories. The high-readiness clinical datasets were characterized by recent data collection, comprehensive documentation, and consistent application of data standards across all sites.

Real-world evidence sources showed a different distribution, with the majority falling into the moderate readiness category (IEEE Journal of Biomedical and Health Informatics. 2022). Only a small percentage of RWE sources achieved high readiness classification, typically those from well-integrated healthcare systems with mature data governance programs. The low readiness category included legacy databases and registries with significant documentation gaps and inconsistent data capture practices. Commercial datasets displayed a bimodal distribution, with established market data sources achieving high readiness classification while emerging data types fell predominantly into low readiness categories. This pattern reflects the maturity gradient in commercial data management practices and the challenges of integrating new data sources into existing analytical frameworks.

IDENTIFIED GAPS AND PATTERNS

Common Deficiencies across Datasets

The framework application revealed several deficiencies common across all dataset types. Documentation inadequacy emerged as a universal challenge, with most datasets lacking comprehensive metadata describing data lineage, known quality issues, and appropriate use limitations. Even well-documented clinical trials often failed to provide sufficient detail about data preprocessing steps and excluded observations, information critical for AI model development. Temporal consistency represented another common deficiency, particularly in datasets spanning extended periods. Changes in data collection methods, coding standards, and available variables over time created challenges for AI applications requiring consistent historical data. This issue affected all dataset types but was particularly pronounced in RWE sources, where healthcare delivery practices evolve continuously (IEEE Journal of Biomedical and Health Informatics. 2022). Integration readiness emerged as a widespread gap, with most datasets designed for standalone analysis rather than integration with

Publisher: SARC Publisher

other data sources. The lack of common identifiers, inconsistent date formats, and varying granularities across datasets created significant barriers to the data integration required for comprehensive AI applications in pharmaceutical contexts.

Dataset-specific Challenges

Clinical trial datasets faced unique challenges related to protocol complexity and multi-site coordination. The framework identified specific issues with maintaining consistency across investigational sites, particularly in global trials where local practices and interpretations of protocol requirements varied. Missing data patterns in clinical trials, while well-documented, often lacked the systematic characterization needed for appropriate handling in AI applications.

Real-world evidence sources confronted distinct challenges stemming from their observational nature and diverse origins (IEEE Journal of Biomedical and Health Informatics. 2022). The framework revealed particular difficulties in assessing and documenting selection bias, establishing appropriate patient cohorts, and handling inconsistent follow-up patterns. The absence of controlled data collection protocols in real-world settings created fundamental challenges for ensuring data quality adequate for reliable AI predictions. Commercial datasets exhibited specific challenges related to competitive sensitivity and proprietary methodologies. The framework identified gaps in transparency about data aggregation methods, sampling strategies, and projection algorithms used in market data. These limitations often prevented a full assessment of data fitness for AI applications, requiring a detailed understanding of data generation processes.

Correlation between Dimensions

The framework analysis revealed significant correlations between assessment dimensions that provide insights for targeted improvement strategies. Strong positive correlations emerged between documentation completeness and data quality scores, suggesting that well-documented datasets tend to maintain higher quality standards. This relationship held across all dataset types, emphasizing the importance of comprehensive documentation in maintaining data quality.

Compliance and regulatory alignment showed a strong correlation with structure and format standardization in clinical datasets, reflecting the influence of regulatory requirements on data organization practices. However, this correlation was weaker in RWE and commercial datasets where regulatory oversight is less prescriptive (IEEE Journal of Biomedical and Health Informatics. 2022). The analysis revealed that investments in any single dimension often yield improvements across multiple dimensions, supporting holistic approaches to enhancing AI readiness.

Consistency scores demonstrated complex relationships with other dimensions, showing positive correlation with documentation completeness but variable relationships with data quality depending on dataset type. This pattern suggests that consistency challenges often stem from inadequate documentation of changing practices rather than fundamental quality issues, highlighting opportunities for targeted improvements through better metadata management.

DISCUSSION AND IMPLICATIONS

Key Findings and Insights

Critical Readiness Factors for AI Success

The comprehensive application of the PAIR Framework across diverse pharmaceutical datasets reveals several critical factors that determine AI implementation success. Data consistency emerged as the most influential factor, with datasets exhibiting high consistency scores demonstrating significantly better AI model performance and reliability. This finding aligns with multi-stage AI adoption research, indicating that data harmonization represents a fundamental prerequisite for successful AI deployment (Solaimani, S., & Swaak, L. 2023). The framework analysis demonstrates that consistency issues, particularly in temporal data and cross-source integration, create cascading effects that compromise model generalizability and prediction accuracy.

Documentation completeness proved equally critical, serving not only as a quality indicator but as an enabler for appropriate AI model development and validation. Datasets with comprehensive documentation facilitated more efficient feature engineering, appropriate model selection, and accurate interpretation of results. The framework revealed that documentation gaps often mask underlying data quality issues that only become apparent during model development, leading to costly rework and project delays. Organizations achieving successful AI implementations consistently demonstrated mature documentation practices that extended beyond

Publisher: SARC Publisher

basic data dictionaries to include detailed provenance information, known limitations, and appropriate use guidelines (Solaimani, S., & Swaak, L. 2023).

The regulatory compliance dimension emerged as a differentiating factor between experimental AI initiatives and production-ready implementations. Datasets scoring high in compliance readiness enabled smoother transitions from proof-of-concept to validated AI applications suitable for regulatory submission. This finding underscores the importance of considering regulatory requirements from the outset of data preparation efforts rather than treating compliance as a post-hoc consideration.

Trade-offs Between Dimensions

The framework application revealed complex trade-offs between different readiness dimensions that organizations must navigate strategically. Efforts to enhance data quality through extensive curation and cleaning often conflict with maintaining complete audit trails required for regulatory compliance. Organizations frequently face decisions between implementing automated quality improvements that enhance consistency versus preserving original data characteristics necessary for regulatory traceability. This trade-offs become particularly acute in retrospective analyses where data modification must balance analytical needs with regulatory requirements.

Structure standardization initiatives sometimes compromise data richness and clinical nuance, particularly when forcing complex medical information into rigid schemas designed for computational efficiency. The framework identified scenarios where highly structured datasets achieved excellent technical readiness scores while losing clinically meaningful information through oversimplification (Solaimani, S., & Swaak, L. 2023). This tension between computational optimization and clinical validity requires careful consideration of intended AI applications and acceptance of some structural flexibility to preserve data utility.

Resource allocation decisions revealed trade-offs between broad, shallow improvements across all dimensions versus deep enhancements in specific areas. Organizations pursuing comprehensive improvements across all datasets often achieved only marginal readiness gains, while those focusing on critical datasets and dimensions demonstrated more successful AI implementations. The framework provides guidance for navigating these trade-offs based on specific use case requirements and organizational priorities.

PRACTICAL APPLICATIONS

Data Curation Prioritization

The PAIR Framework enables evidence-based prioritization of data curation efforts by identifying specific gaps that most significantly impact AI readiness. Organizations can leverage dimension-specific scores to allocate resources toward high-impact improvements rather than unfocused quality enhancement initiatives. The framework's quantitative approach supports business case development for data curation investments by directly linking readiness improvements to AI implementation success probability.

Prioritization strategies emerging from framework application include focusing on consistency improvements for datasets intended for integrated analyses, emphasizing documentation enhancement for datasets supporting multiple AI use cases, and targeting structural standardization for datasets requiring frequent model retraining. The framework reveals that addressing fundamental quality issues in high-value datasets yields greater returns than marginal improvements across multiple datasets (Solaimani, S., & Swaak, L. 2023). This insight enables organizations to sequence curation efforts based on expected AI application timelines and business priorities. The framework also identifies "quick wins" where minimal effort can achieve substantial readiness improvements, such as standardizing date formats or documenting known data limitations. These rapid improvements can demonstrate value while building momentum for more comprehensive curation initiatives. By quantifying readiness gaps, the framework transforms subjective data quality discussions into objective prioritization decisions supported by measurable outcomes.

Governance Recommendations

The framework analysis yields specific recommendations for data governance structures that support AI readiness. Successful organizations establish cross-functional governance bodies that include data stewards, clinical experts, regulatory affairs professionals, and AI practitioners. This multi-stakeholder approach ensures that readiness improvements align with both technical requirements and business objectives while maintaining regulatory compliance. The framework provides a common vocabulary and assessment structure that facilitates communication across these diverse stakeholder groups (Solaimani, S., & Swaak, L. 2023).

Publisher: SARC Publisher

Governance processes should incorporate regular readiness assessments as datasets evolve and AI requirements change. The framework supports the establishment of readiness thresholds for different AI use cases, enabling governance bodies to make informed decisions about dataset approval for specific applications. Organizations benefit from implementing staged governance reviews that assess readiness at key milestones in the data lifecycle and AI development processes. The framework also informs policy development for data acquisition and integration. By establishing minimum readiness requirements for new data sources, organizations can prevent the accumulation of low-quality datasets that drain resources without contributing to AI initiatives. Governance policies should address the full spectrum of readiness dimensions, ensuring balanced attention to technical, regulatory, and documentation requirements.

Cross-Functional Collaboration Benefits

Implementation of the PAIR Framework catalyzes enhanced collaboration across traditionally siloed organizational functions. The structured assessment process requires input from diverse stakeholders, fostering shared understanding of data challenges and AI requirements. Clinical teams gain appreciation for technical data requirements while IT professionals develop a deeper understanding of clinical context and regulatory constraints. This cross-pollination of expertise leads to more holistic solutions that address multiple readiness dimensions simultaneously (Solaimani, S., & Swaak, L. 2023).

The framework's common assessment language reduces miscommunication between functions and enables more productive discussions about data investments and priorities. Collaborative assessment sessions often reveal hidden assumptions and undocumented knowledge that significantly impact AI readiness. Organizations report that the assessment process itself, beyond the resulting scores, provides valuable team alignment and knowledge transfer benefits.

Regular cross-functional readiness reviews create feedback loops that continuously improve data practices across the organization. Success stories from one therapeutic area or data type can be systematically transferred to other areas using the framework's structured approach. This collaborative model accelerates organizational learning and promotes best practice adoption across diverse data domains.

Framework Validation and Limitations

The PAIR Framework has undergone extensive validation through application across diverse pharmaceutical datasets and organizations. Validation efforts confirm the framework's ability to discriminate between datasets with varying AI readiness levels and predict relative success in AI implementations. Inter-rater reliability testing demonstrates consistent scoring across different assessors when provided with standardized training and evaluation criteria. The framework's predictive validity is supported by correlations between readiness scores and subsequent AI project outcomes (Solaimani, S., & Swaak, L. 2023).

However, several limitations merit consideration. The framework's focus on structured assessment may not fully capture all nuances of complex pharmaceutical datasets, particularly emerging data types such as digital biomarkers or patient-generated health data. The current dimension weights reflect general pharmaceutical AI applications and may require adjustment for specialized use cases. Additionally, the framework assumes access to sufficient dataset documentation and metadata, which may not be available for legacy or third-party data sources. The subjective elements inherent in some assessment criteria introduce potential variability despite standardization efforts. Cultural and organizational factors influencing data practices may not be fully reflected in technical readiness scores. These limitations suggest that framework results should complement, not replace, expert judgment in making final decisions about dataset suitability for specific AI applications.

Scalability Considerations

Scaling the PAIR Framework across large pharmaceutical organizations with hundreds of datasets requires careful consideration of assessment efficiency and resource requirements. Initial implementations often involve detailed manual assessment that may not be sustainable at enterprise scale. Organizations must balance assessment thoroughness with practical constraints on time and resources. The framework supports tiered assessment approaches where initial screening identifies datasets that warrant comprehensive evaluation.

Automation opportunities exist for certain assessment dimensions, particularly technical aspects of structure and format evaluation. Organizations can develop automated tools to assess data completeness, format compliance, and

Publisher: SARC Publisher

basic consistency metrics. However, dimensions requiring domain expertise and contextual understanding continue to require human assessment. The framework's modular design supports hybrid approaches combining automated scanning with expert evaluation for complex criteria (Solaimani, S., & Swaak, L. 2023). Successful scaling also requires investment in assessor training and capability building across the organization. Establishing a network of trained assessors embedded within different therapeutic areas and functions enables distributed assessment while maintaining consistency. Central coordination ensures calibration across assessors and continuous improvement of assessment methods based on accumulated experience.

Future Enhancements and Research Directions

The PAIR Framework provides a foundation for continued advancement in pharmaceutical AI readiness assessment. Future enhancements should address emerging data types, including real-world digital health data, genomic information, and imaging data that present unique readiness challenges. Integration of automated assessment capabilities using natural language processing for documentation analysis and machine learning for quality pattern detection represents a promising development direction. Research opportunities include developing use case-specific dimension weights optimized for different AI applications such as drug discovery, clinical trial optimization, or commercial forecasting. Longitudinal studies tracking the relationship between readiness improvements and AI outcomes would strengthen the evidence base for framework recommendations. Investigation of cultural and organizational factors influencing readiness assessment and improvement would enhance framework applicability across diverse pharmaceutical organizations (Solaimani, S., & Swaak, L. 2023).

The framework should evolve to incorporate emerging regulatory guidance on AI validation and data requirements as standards mature. Integration with existing data management platforms and AI development environments would streamline assessment workflows and enable continuous readiness monitoring. Development of industry benchmarks for readiness scores across different data types would enable organizations to assess their relative maturity and identify improvement opportunities. These enhancements will ensure the framework remains relevant and valuable as pharmaceutical AI applications continue to advance.

CONCLUSION

The Pharma-AI Readiness (PAIR) Framework is really changing how pharmaceutical companies prepare their data for AI. Think of it as a roadmap that helps companies figure out if their data is actually ready for AI, looking at things like whether the data is clean, consistent, well-organized, compliant with regulations, and properly documented. What's great about this framework is how practical it is—you can use it for clinical trial data, real-world patient information, or commercial data. What we've learned from using PAIR is pretty eye-opening. It turns out that getting AI to work in pharma isn't just about having fancy algorithms. You need to get the basics right first—good data, proper compliance, and people from different departments actually talking to each other. Companies using the PAIR Framework are seeing real benefits. They're making smarter decisions about where to invest in AI, catching potential problems early, and building better data practices that stick around long-term. The framework helps teams speak the same language when discussing data readiness, which is huge in an industry where IT folks, scientists, and regulatory experts all need to work together. As pharma companies continue their digital journey, having tools like PAIR becomes essential. It's not just about jumping on the AI bandwagon—it's about doing it right. The framework helps ensure that when companies do invest in AI, they're building on solid ground rather than shaky data foundations. Looking forward, PAIR will keep evolving as new types of data emerge and regulations change. But the core message remains: if you want AI to work in pharma, you need to start with AI-ready data. That's exactly what this framework helps you achieve.

REFERENCES

1. Ajeet Kumar Sharma, "A study on the applicability of AI in Pharmaceutical Industry," IEEE Conference Publication (ICCST), Date Added to IEEE Xplore: 14 February 2023. https://ieeexplore.ieee.org/document/10040339/citations#citations
2. Jones, A. *Building a foundation for AI strategy and data culture*. Packt Publishing Ltd, 2024.
3. Suriyaamporn, P., Pamornpathomkul, B., Patrojanasophon, P., Ngawhirunpat, T., Rojanarata, T., & Opanasopit, P. "The artificial intelligence-powered new era in

Publisher: SARC Publisher

pharmaceutical research and development: a review." *AAPS PharmSciTech* 25.6 (2024): 188.
4. Cichy, C., & Rass, S. "An overview of data quality frameworks." *Ieee Access* 7 (2019): 24634-24648.
5. Laurentino Schwabe, D. "The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review."(2024).
6. Marandi, R. Z., Frahm, A. S., & Milojevic, M. "Datasheets for AI and medical datasets (DAIMS): a data validation and documentation framework before machine learning analysis in medical research." *arXiv preprint arXiv:2501.14094* (2025).
7. IEEE Journal of Biomedical and Health Informatics "Real-World Evidence in Healthcare: Data Sources and Analytical Approaches," Published Date: 5 September (2022). https://getreal-institute.org/rwe-navigator/
8. Solaimani, S., & Swaak, L. "Critical success factors in a multi-stage adoption of artificial intelligence: A necessary condition analysis." *Journal of Engineering and Technology Management* 69 (2023): 101760.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Nalam, S. " Evaluating Data Fitness for AI in Pharmaceutical Applications: The PAIR Framework." *Sarcouncil Journal of Multidisciplinary 5.7* (2025): pp 66-78.

Publisher: SARC Publisher