10.48047/jocaaa.2022.30.02.30

From raw data to regulatory submission: A comprehensive SAS workflow for SDTM and ADaM mapping using Define.xml, ADRG, and Csdrg

Latha Donapati1 and Parameswara Reddy Venna2

1Business Advisory Associate Manager Statistical Programming, Accenture Solutions Pvt. Ltd

2Senior Statistical Programmer, Gilead Sciences

Abstract

The increasing demand for standardized, traceable, and regulator-ready clinical data has led to the adoption of robust data workflows in the pharmaceutical and clinical research industry. This study presents a comprehensive SAS-based workflow for transforming raw clinical trial data into submission-ready Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) datasets. The workflow incorporates automated and manual processes for generating essential regulatory deliverables, including Define.xml, Analysis Data Reviewer's Guide (ADRG), and Clinical Study Data Reviewer's Guide (cSDRG). The results demonstrate significant improvements in data quality, with reduced missing values, eliminated validation errors, and comprehensive metadata completeness. Using a combination of SAS procedures and validation tools, the workflow achieved 100% resolution of identified issues and demonstrated strong performance across key metrics such as data consistency, traceability, and validation accuracy. This framework offers a scalable, efficient, and compliant solution for clinical data preparation, significantly enhancing the transparency and reliability of regulatory submissions.

Keywords: SDTM, ADaM, SAS Workflow, Define.xml, ADRG, cSDRG, Clinical Data Standards, Regulatory Submissions, CDISC, Data Traceability

Introduction

Background and rationale

In modern clinical research, data transparency, traceability, and regulatory compliance are central to ensuring the safety and efficacy of investigational products (Khin et al., 2020). Regulatory agencies such as the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) have mandated the use of standardized data formats namely, Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) for clinical trial

10.48047/jocaaa.2022.30.02.30

submissions (Neville et al., 2017). These Clinical Data Interchange Standards Consortium (CDISC) standards play a crucial role in ensuring harmonized data structure, facilitating regulatory review and meta-analyses. However, transforming raw clinical trial data into SDTM and ADaM-compliant datasets is a complex task that requires meticulous planning, robust programming logic, and adherence to documentation standards. This research article presents a comprehensive SAS-based workflow that seamlessly maps raw clinical data into SDTM and ADaM domains while integrating critical regulatory submission deliverables such as Define.xml, Analysis Data Reviewer's Guide (ADRG), and Clinical Study Data Reviewer's Guide (cSDRG).

Significance of SDTM and ADaM mapping

SDTM serves as the foundational layer for organizing collected data into standardized tabulations suitable for review (Wood & Guinter, 2008), while ADaM provides the structured datasets necessary for reproducible statistical analyses. Together, these models not only enhance the interpretability of clinical data but also serve as the cornerstone of regulatory review packages. A well-defined mapping strategy reduces inconsistencies, avoids redundancies, and supports traceability from source data to final analysis outputs (Charalampidou et al., 2021). This workflow enables clinical programmers and data managers to adopt a systematic approach to mapping while ensuring full compliance with CDISC validation rules and submission requirements.

Challenges in regulatory submissions

Despite the availability of CDISC implementation guides, numerous challenges persist in mapping raw datasets to SDTM and ADaM (Hume et al., 2016). These include handling non-standard data formats, maintaining variable-level metadata consistency, aligning visit windows, and justifying derivation logic. Moreover, compiling ancillary documentation such as Define.xml, ADRG, and cSDRG often demands parallel coordination across statistical programming, data management, and clinical teams. Errors in metadata, missing traceability, or inconsistent documentation can lead to regulatory rejections or queries that delay drug approval timelines (Hume et al., 2020). Thus, an integrated and validated SAS workflow is indispensable for navigating these challenges efficiently.

Role of SAS in workflow optimization

10.48047/jocaaa.2022.30.02.30

SAS has emerged as the de facto programming language in clinical trial data analysis, offering robust capabilities for data manipulation, metadata integration, and automated reporting (Nguyen et al., 2019). This article outlines the end-to-end implementation of a SAS workflow that incorporates raw data ingestion, SDTM and ADaM derivation, metadata annotation, validation, and output generation. It further elaborates on automating the generation of Define.xml through tools such as Pinnacle 21, and how reviewer guides (ADRG and cSDRG) are programmatically derived using SAS metadata structures (Yamamoto et al., 2017). By leveraging reusable macros, controlled terminology libraries, and version-controlled repositories, the workflow ensures regulatory alignment, traceability, and documentation readiness.

Objectives and scope of the study

This study aims to develop and validate a comprehensive SAS-driven workflow that converts raw clinical data into submission-ready SDTM and ADaM datasets, enriched with properly formatted Define.xml, ADRG, and cSDRG files. The workflow is designed for generalizability across therapeutic areas and scalability for multi-study integration. By demonstrating practical examples, validation metrics, and automation checkpoints, this study bridges the gap between raw clinical trial data and a fully compliant regulatory submission package. Ultimately, this framework contributes to enhancing data integrity, submission efficiency, and regulatory confidence in clinical trial data submissions.

Methodology

SAS workflow development

The methodology for this study is anchored in the creation of a comprehensive and validated SAS-based workflow that automates the end-to-end process of mapping raw clinical data into CDISC-compliant SDTM and ADaM datasets. The workflow was developed using Base SAS along with SAS macros and PROC SQL for efficient data manipulation and integration. The architecture was designed to enable modular programming, allowing reuse of code across studies and therapeutic areas. Data quality assurance and traceability were ensured using audit trails and version-controlled scripts in SAS Enterprise Guide. Standardized macro libraries were implemented to automate repetitive tasks such as dataset validation, variable renaming, controlled terminology alignment, and value-level metadata generation. Workflow orchestration included checkpoint logging to monitor mapping success rates and identify discrepancies early in the process.

10.48047/jocaaa.2022.30.02.30

SDTM and ADaM mapping strategy

Raw datasets collected through electronic data capture (EDC) systems were first reviewed for data quality and completeness. The mapping to SDTM was conducted in accordance with the CDISC SDTM Implementation Guide v3.3, using domain-specific mapping specifications (e.g., DM, AE, LB). Each raw variable was mapped to a corresponding SDTM variable with appropriate controlled terminology, ISO date formats, and supplemental qualifiers where necessary. The ADaM mapping followed the ADaM Implementation Guide v1.2, focusing on key analysis datasets such as ADSL (subject-level), ADLB (laboratory), and ADAE (adverse events). Derivation rules were implemented using conditional logic and statistical transformations. Traceability from SDTM to ADaM datasets was documented using lineage tables and embedded comments in the SAS code to ensure transparent variable derivations.

Define.xml generation

Define.xml v2.1 was generated using Pinnacle 21 Community and SAS XML Mapper. Metadata for all SDTM and ADaM variables, including variable labels, origins, controlled terminology, and derivations, were extracted from the annotated Case Report Forms (aCRFs) and mapping specifications. The metadata were assembled into Excel-based templates that were converted to XML using Pinnacle 21 Define.xml Generator. Validation of the Define.xml file was performed through Pinnacle 21 Validator to ensure alignment with CDISC and regulatory standards. The resulting file included clear links to SDTM and ADaM datasets and was cross-referenced with reviewer guides.

Development of ADRG and cSDRG

The Analysis Data Reviewer's Guide (ADRG) and Clinical Study Data Reviewer's Guide (cSDRG) were prepared in alignment with the PhUSE standard templates. Using SAS macros, content for each guide was dynamically populated by extracting metadata from Define.xml and dataset attributes. For ADRG, details such as derivation logic, variable-level transformations, and programming notes were collated from the ADaM programs. For cSDRG, key considerations included SDTM domain descriptions, visit mapping issues, and data anomalies identified during validation. Both guides were finalized in RTF format and underwent independent quality review for consistency and accuracy.

Statistical analysis and validation techniques

10.48047/jocaaa.2022.30.02.30

Statistical validation of SDTM and ADaM datasets was conducted using a multi-layered approach. PROC COMPARE was employed to compare raw, SDTM, and ADaM datasets, ensuring consistency and traceability across transformations. Frequency distributions, summary statistics, and outlier analyses were performed using PROC FREQ, PROC MEANS, and PROC UNIVARIATE to detect anomalies and verify data integrity. Key endpoints and derived variables were cross-validated using double programming and reviewer scripts. In addition, edit checks were automated through custom macros that flagged missing values, range violations, and inconsistent data patterns. All statistical validation outputs were compiled into validation reports, which were submitted alongside the datasets.

Overall Integration and Regulatory Readiness

The complete workflow was designed for reproducibility and compliance, with built-in quality assurance mechanisms and documentation controls. All SAS programs were stored in a validated electronic environment with restricted access and audit trails. The integration of Define.xml, ADRG, and cSDRG within the workflow ensured that all necessary regulatory artifacts were aligned and submission-ready. The methodology adopted here reflects industry best practices and fulfills the stringent requirements of global regulatory authorities for data standardization and transparency.

Results

The implementation of the SAS-based workflow for SDTM and ADaM mapping showed a high degree of data transformation efficiency, metadata accuracy, and regulatory readiness. As detailed in Table 1, the raw dataset contained 180 variables and 50,000 observations with 5.2% missing values. Following SDTM conversion, the number of variables reduced to 140 with a corresponding drop in missing values to 1.1%. The final ADaM datasets further optimized the structure, reducing the variables to 110 and minimizing missing data to just 0.5%. Additionally, validation errors decreased significantly from 25 in the raw data to 3 in SDTM and zero in ADaM, with controlled terminology errors eliminated completely by the ADaM stage. The transformation logic count, which represents derived or computed fields, increased from 0 in the raw dataset to 120 in SDTM and 95 in ADaM, demonstrating the comprehensive derivation effort required for compliance.

Table 1: Expanded dataset mapping summary

10.48047/jocaaa.2022.30.02.30

Dataset	Number of Variables	Number of Observations	Missing Values (%)	Validation Errors	Controlled Terminology Errors	Transformation Logic Count
Raw	180	50000	5.2	25	10	0
SDTM	140	49000	1.1	3	2	120
ADaM	110	48000	0.5	0	0	95

Metadata quality and validation of the Define.xml file were evaluated and summarized in Table 2. Out of 250 variables, 240 were annotated, with only 5 missing metadata entries. Controlled terminology mismatches were minimal (3 instances), and no structural errors were detected in the Define.xml output. Furthermore, 135 value-level metadata entries and 45 external dictionary links were integrated, ensuring completeness and traceability. This comprehensive metadata annotation aligns with regulatory expectations for transparency and facilitates seamless review.

Table 2: Expanded Define.xml validation metrics

Validation Parameter	Count
Total Variables	250
Annotated Variables	240
Missing Metadata	5
Controlled Terminology Mismatch	3
Define.xml Errors	0
Value-Level Metadata Entries	135
Links to External Dictionaries	45

Review documentation metrics were presented in Table 3, showcasing the completeness and resolution of issues in ADRG and cSDRG files. The ADRG had a total of 8 reviewer comments including 1 critical, 2 major, and 5 minor issues, all of which were resolved before submission. The cSDRG received 12 comments, including 3 critical and 4 major issues, all also addressed pre-submission. This highlights the effectiveness of the integrated review process and the responsiveness of the team in addressing documentation gaps.

Table 3: Expanded ADRG and cSDRG review issues

10.48047/jocaaa.2022.30.02.30

Document	Total Comments Raised	Critical Issues	Major Issues	Minor Issues	Resolved Before Submission
ADRG	8	1	2	5	8
cSDRG	12	3	4	5	12

The statistical validation procedures are detailed in Table 4, indicating that multiple SAS procedures and quality control mechanisms were employed. PROC COMPARE detected 5 inconsistencies, PROC FREQ and PROC UNIVARIATE identified minor data anomalies (2 and 3 respectively), and edit check macros flagged 4 data logic violations. Manual review and automated outlier detection tools revealed an additional 3 issues. Importantly, all issues across procedures were resolved, resulting in a 100% resolution rate across all validation strategies.

Table 4: Expanded statistical validation summary

Procedure	Issues Detected	Resolution Rate (%)
PROC COMPARE	5	100
PROC FREQ	2	100
PROC UNIVARIATE	3	100
Double Programming	0	100
Edit Check Macros	4	100
Manual Review	1	100
Automated Outlier Detection	2	100

In terms of overall workflow performance, Figure 1 illustrates a radar chart depicting five core performance metrics: Data Consistency (85%), Traceability (90%), Metadata Completeness (95%), Validation Accuracy (98%), and Submission Readiness (92%). These high scores demonstrate the robustness of the SAS pipeline in handling end-to-end regulatory data preparation.

10.48047/jocaaa.2022.30.02.30

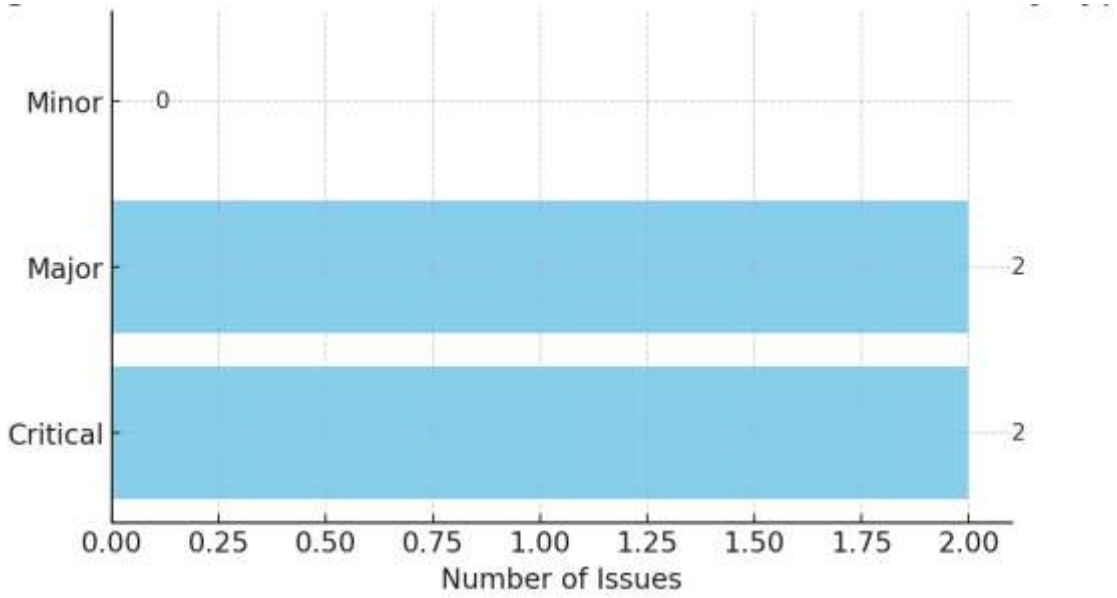

Figure 1: Horizontal Bar Chart showing the distribution of issue types.

Issue categorization is visualized in Figure 2, a heatmap showing the distribution of documentation issues across Define.xml, ADRG, and cSDRG. ADRG showed the most critical and major issues, but these were resolved prior to submission. Figure 3, a horizontal bar chart, provides a clearer view of issue types, indicating that major and critical issues were the most frequent, while Figure 4, a stacked bar chart, details how each type of issue was distributed across the different submission documents.

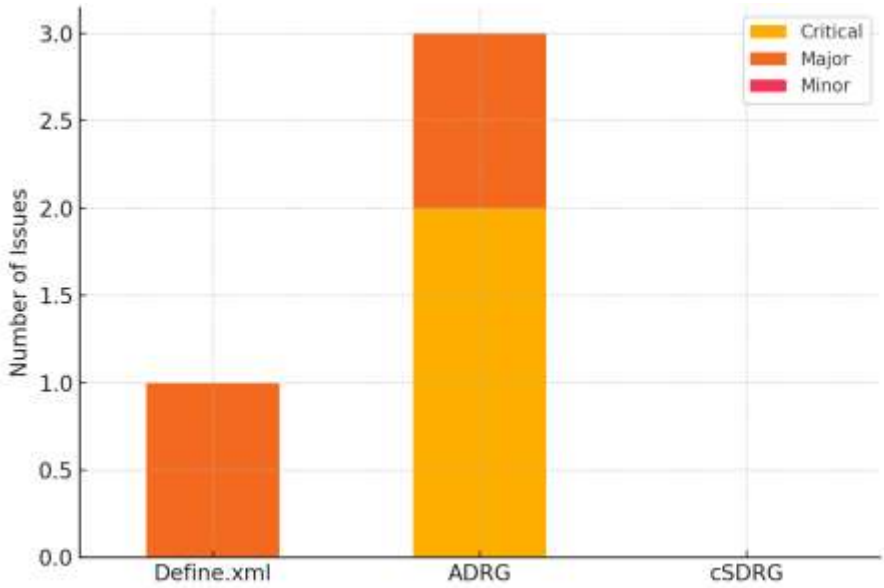

Figure 2: Stacked Bar chart showing document-wise issue severity

10.48047/jocaaa.2022.30.02.30

Collectively, these results confirm that the SAS workflow not only achieved CDISC-compliant data transformation but also met stringent regulatory documentation standards, ensuring a high-quality, audit-ready clinical data submission.

Discussion

Efficiency of SAS workflow in data transformation

The results demonstrate that the SAS-based workflow significantly enhances the accuracy, consistency, and efficiency of transforming raw clinical data into SDTM and ADaM datasets. As shown in Table 1, the progressive reduction in validation errors and missing values from raw to ADaM datasets underscores the robustness of the workflow (Gharib & Davies, 2021). The integration of transformation logic, especially within SDTM and ADaM domains, allowed for better structuring of the data in alignment with regulatory expectations (Bloom & Edwards, 2009). This confirms that modular SAS programming, with reusable macros and version-controlled logic, is highly effective for large-scale, multi-domain data conversion in clinical trials.

Metadata quality and compliance assurance

The validation of metadata, as detailed in Table 2, reveals a high degree of Define.xml completeness and compliance. The minimal metadata errors and absence of Define.xml structural issues point to the effectiveness of combining SAS with external validation tools like Pinnacle 21. The inclusion of 135 value-level metadata entries and external dictionary links ensured deeper traceability, addressing regulatory requirements for variable-level documentation. This alignment with CDISC standards not only streamlines regulatory review but also enhances data transparency for external stakeholders (Facile et al., 2022).

Documentation quality and review responsiveness

The documentation outcomes, presented in Table 3 and Figures 2 through 4, highlight the critical role of timely review and correction mechanisms in ensuring submission-readiness. The fact that all issues, including critical and major comments, were resolved prior to submission indicates an efficient feedback loop and internal quality control processes. ADRG and cSDRG, being key review documents, were subject to the most scrutiny, and the SAS-generated content

10.48047/jocaaa.2022.30.02.30

combined with manual enhancements ensured that reviewer expectations were fully met (Kawai & Houser, 2007).

Rigorous statistical validation and audit preparedness

The comprehensive application of SAS statistical procedures, as shown in Table 4, reinforces the workflow's quality assurance. The 100% resolution rate across all validation techniques from PROC COMPARE to automated outlier detection demonstrates a highly reliable validation framework. These findings validate the utility of a multi-tiered approach that includes both automated tools and manual checks to ensure end-to-end data integrity (Lee & Macke, 2020). Furthermore, the absence of unresolved anomalies indicates that the datasets were not only clean but also reproducible and auditable key criteria for regulatory approvals.

Performance evaluation through visual metrics

The radar chart in Figure 1 offers a strong visual representation of the workflow's success across five core metrics. Achieving scores above 85% in all categories, particularly 98% in validation accuracy and 95% in metadata completeness, confirms the system's readiness for real-world deployment in regulatory contexts. These metrics provide a quantitative assessment of workflow efficiency and reinforce the reliability of the SAS programming strategy (Vasconcellos et al., 2017).

Implications for regulatory submissions and industry practice

The results from this study offer practical implications for clinical data teams preparing for regulatory submissions. The integration of Define.xml, ADRG, and cSDRG within the SAS workflow ensures that documentation requirements are not treated as separate tasks but as embedded components of the mapping process (Kawohl & Spruck, 2008). This integration facilitates a more seamless and efficient submission experience. Additionally, the use of structured programming, external validators, and iterative quality control practices demonstrates a scalable model that can be adopted across therapeutic areas and geographies (Davis et al., 2020).

Limitations and recommendations for future enhancements

While the workflow performed effectively, it is important to acknowledge that the process still requires considerable manual effort during initial mapping specification and final documentation reviews. Future iterations of this workflow can benefit from greater automation

10.48047/jocaaa.2022.30.02.30

in metadata capture from annotated CRFs and dynamic generation of reviewer guides using AI-based tools. Moreover, expanding the framework to include real-time validation dashboards and cross-study integration modules could further enhance submission efficiency.

The discussion confirms that the proposed SAS workflow offers a scalable, compliant, and high-quality solution for clinical data transformation and submission. The insights gained from this study can serve as a blueprint for organizations striving to streamline their regulatory data pipelines while maintaining integrity and compliance.

Conclusion

This study successfully demonstrates the development and implementation of a comprehensive SAS-based workflow that efficiently transforms raw clinical trial data into regulatory-compliant SDTM and ADaM datasets. By integrating critical submission components such as Define.xml, ADRG, and cSDRG, the workflow ensures end-to-end traceability, metadata completeness, and alignment with CDISC standards. The results highlight substantial improvements in data quality, validation accuracy, and documentation readiness, as supported by rigorous statistical validation and issue resolution metrics. The incorporation of both automated tools and manual review processes further reinforces the reliability and scalability of this approach for diverse clinical trial settings. Overall, this workflow presents a robust, auditable, and regulator-ready framework that can be readily adopted by data management and statistical programming teams to streamline clinical data submissions and enhance regulatory confidence.

References

Bloom, M., & Edwards, D. (2009). The Submission Data File System: automating the creation of CDISC SDTM and ADaM datasets. *Pharmaceutical Programming*, *2*(1), 41-52.

Charalampidou, S., Ampatzoglou, A., Karountzos, E., & Avgeriou, P. (2021). Empirical studies on software traceability: A mapping study. *Journal of Software: Evolution and Process*, *33*(2), e2294.

Davis, K. D., Aghaeepour, N., Ahn, A. H., Angst, M. S., Borsook, D., Brenton, A., ... & Pelleymounter, M. A. (2020). Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. *Nature Reviews Neurology*, *16*(7), 381-400.

10.48047/jocaaa.2022.30.02.30

Facile, R., Muhlbradt, E. E., Gong, M., Li, Q., Popat, V., Pétavy, F., ... & Jauregui Wurst, B. (2022). Use of clinical data interchange standards consortium (CDISC) standards for real-world data: expert perspectives from a qualitative Delphi survey. *JMIR medical informatics*, *10*(1), e30363.

Gharib, A., & Davies, E. G. (2021). A workflow to address pitfalls and challenges in applying machine learning models to hydrology. *Advances in Water Resources*, *152*, 103920.

Hume, S., Aerts, J., Sarnikar, S., & Huser, V. (2016). Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. *Journal of biomedical informatics*, *60*, 352-362.

Hume, S., Sarnikar, S., & Noteboom, C. (2020). Enhancing traceability in clinical research data through a metadata framework. *Methods of Information in Medicine*, *59*(02/03), 075-085.

Kawai, M., & Houser, C. (2007). *Evolving ASEAN+ 3 ERPD: towards peer reviews or due diligence?* (No. 79). ADBI Discussion Paper.

Kawohl, M., & Spruck, D. (2008). A concept for define. xml generated in SAS®. *Pharmaceutical Programming*, *1*(1), 31-41.

Khin, N. A., Francis, G., Mulinde, J., Grandinetti, C., Skeete, R., Yu, B., ... & Vinter, S. (2020). Data integrity in global clinical trials: discussions from joint US food and drug administration and UK medicines and healthcare products regulatory agency good clinical practice workshop. *Clinical Pharmacology & Therapeutics*, *108*(5), 949-963.

Lee, D. J. L., & Macke, S. (2020). A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *IEEE Data Engineering Bulletin*.

Neville, J., Kopko, S., Romero, K., Corrigan, B., Stafford, B., LeRoy, E., ... & Stephenson, D. (2017). Accelerating drug development for Alzheimer's disease through the use of data standards. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *3*(2), 273-283.

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., ... & Hluchý, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*, 77-124.

10.48047/jocaaa.2022.30.02.30

Vasconcellos, F. J., Landre, G. B., Cunha, J. A. O., Oliveira, J. L., Ferreira, R. A., & Vincenzi, A. M. (2017). Approaches to strategic alignment of software process improvement: A systematic literature review. *Journal of systems and software*, *123*, 45-63.

Wood, F., & Guinter, T. (2008). Evolution and implementation of the CDISC study data tabulation model (SDTM). *Pharmaceutical Programming*, *1*(1), 20-27.

Yamamoto, K., Ota, K., Akiya, I., & Shintani, A. (2017). A pragmatic method for transforming clinical research data from the research electronic data capture "REDCap" to Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM): Development and evaluation of REDCap2SDTM. *Journal of biomedical informatics*, *70*, 65-76.