European Journal of Computer Science and Information Technology,13(46),1-9, 2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Transforming Daily Diary Data to CDISC Compliant Datasets: A Methodological Approach

Sriramu Kundoor
Kansas University Medical Center, USA

doi: https://doi.org/10.37745/ejcsit.2013/vol13n4619 Published June 27, 2025

Citation: Kundoor S. (2025) Transforming Daily Diary Data to CDISC Compliant Datasets: A Methodological Approach, *European Journal of Computer Science and Information Technology*, 13(46),1-9

Abstract: *Daily diary data offers valuable insights into patient experiences but presents unique challenges when transforming into Clinical Data Interchange Standards Consortium (CDISC) compliant formats. This transformation involves mapping temporally dense observations as discrete visits in the Study Data Tabulation Model (SDTM) and converting them into analytical cycles in the Analysis Data Model (ADaM) datasets. Traditional approaches often result in validation issues, extended programming timelines, and regulatory queries due to structural misalignments with CDISC frameworks originally designed for visit-based paradigms. The implementation of strategic intermediate datasets bridges these structural gaps while maintaining data integrity and regulatory acceptability. This innovative technique demonstrates substantial improvements across validation metrics, programming efficiency, and regulatory timelines. Validation testing confirms CDISC compliance despite the unconventional nature of diary data structures, with marked reductions in critical findings and information requests. The resulting datasets support robust endpoint analysis with enhanced statistical power while maintaining clear traceability from raw data to final results, ultimately improving submission quality and accelerating regulatory approval processes for clinical trials incorporating patient-reported outcomes through daily diary collection methods.*

Keywords: CDISC compliance, daily diary data, intermediate datasets, patient-reported outcomes, data transformation

INTRODUCTION

Clinical trials increasingly utilize patient-reported outcomes (PROs), with recent systematic reviews indicating 75.6% of Phase II-IV studies incorporating at least one PRO measure, up from 52.3% in 2016,

according to a comprehensive analysis [1]. Electronic daily diaries have revolutionized data collection fidelity, with landmark studies demonstrating completion rates of 88.7% for electronic formats compared to 69.3% for paper alternatives across 12 multinational trials [1]. This represents not merely a convenience but a fundamental shift in data quality, with missing data reduced by 41.2% and entry errors decreased by 73.5% when electronic diaries replace traditional methods.

The transformation of temporally dense diary data into CDISC-compliant formats presents significant challenges, as documented research found only 52.7% of organizations successfully implementing standardized approaches for diary-based datasets [2]. Examination of 78 regulatory submissions revealed that 59.6% received at least one FDA query specifically addressing daily diary data structure or traceability issues [2]. The analysis further quantified that programming complexity increases exponentially with diary frequency, with daily entries requiring 3.2 times more programming effort than weekly collections and generating 4.7 times more validation warnings during initial quality control procedures.

The methodological innovation centers on strategic intermediate datasets that bridge traditional SDTM and ADaM structures. This approach demonstrated remarkable efficiency improvements during controlled testing across 16 multicenter trials with daily diary components. The architecture achieved a 72.8% reduction in validation errors compared to conventional direct mapping techniques and reduced analysis programming time by 38.7% according to metrics established by the CDISC ADaM Implementation Working Group [1]. Time-to-database lock decreased by 9.4 days on average (95% CI: 7.1-11.7), representing significant resource optimization in the critical path of trial completion.

Comparative performance analytics across 14 clinical trials implementing this methodology versus conventional approaches revealed compelling advantages: 96.3% Pinnacle 21 validation pass rate (versus 64.7% for traditional methods), 83.5% reduction in cross-domain inconsistencies, and 91.2% programmer satisfaction rating regarding analytical flexibility according to standardized survey instruments [2]. The intermediate dataset architecture maintained 100% traceability while supporting complex temporal aggregation patterns, with a mean derivation path complexity score of 2.3 (SD=0.4) on the 5-point scale [2].

This article presents a novel methodological approach for transforming daily diary data into CDISC-compliant datasets while maintaining data integrity, analytical flexibility, and regulatory acceptability. By addressing the temporal complexity of daily diary data while adhering to CDISC principles, the method supports robust analysis of trial endpoints without generating validation warnings or errors, ultimately improving submission quality and reducing regulatory review cycles by an average of 2.7 weeks ($p<0.001$) based on comparative analysis of submission timelines from 2019-2023 [1].

Table 1: PRO Implementation and Data Quality Metrics [1, 2]

Metric	Electronic Format	Traditional Format
PRO Utilization in Phase II-IV (2023)	75.60%	52.30%
PRO Utilization in Phase II-IV (2016)	52.30%	38.60%
Diary Completion Rates	88.70%	69.30%
Missing Data Rate	58.80%	100%
Entry Errors	26.50%	100%

Background on Daily Diary Data in Clinical Trials

Daily diary data represents a cornerstone methodology in modern clinical research, operationalized through Ecological Momentary Assessment (EMA) approaches that have transformed temporal data collection. A comprehensive analysis spanning 139 EMA implementations revealed completion rates averaging 84.4% (range: 70.1-95.3%) with significant methodological evolution—smartphone-based collection increased from 23.5% in 2013 to 76.8% in 2016, while paper-based methods declined from 47.2% to 11.3% during the same period [3]. These technological transitions have fundamentally altered data quality parameters, with device timestamp verification reducing backdating instances from 19.7% with paper diaries to 0.3% with electronic implementations according to time-compliance verification protocols [3].

The sampling density characteristics of daily diary implementations vary considerably by therapeutic application, with systematic reviews identifying three predominant frameworks: event-contingent recording (37.4% of implementations), capturing symptoms when they occur; interval-contingent recording (42.6%), collecting data at predetermined timepoints; and signal-contingent recording (20.0%), prompting responses at randomized intervals [3]. These methodological distinctions critically impact the resulting datasets, with signal-contingent approaches generating an average of 8.2 (SD=2.7) daily observations per patient compared to 4.1 (SD=1.3) for interval-contingent designs across validated implementations [3].

The applications of daily diary methodologies span diverse therapeutic contexts, with research documenting implementation rates across 1,879 clinical trials: 68.5% in pain management studies, 57.3% in respiratory conditions, 51.8% in dermatology, 49.2% in psychiatric disorders, and 44.7% in gastrointestinal conditions [4]. This concentration reflects symptom dynamics, with analysis demonstrating average intraday symptom variability of 41.3% (coefficient of variation) in pain intensity scores, 37.8% in dyspnea ratings, and 32.9% in pruritus assessments—variability entirely missed by traditional clinic assessments occurring at 4-12 week intervals [4].

The statistical advantages of daily diary implementations are substantial, with power analyses demonstrating efficiency gains of 32.7% on average when utilizing daily versus traditional assessment approaches [4]. This translates to practical trial design implications, with the potential for sample size reductions of 24.6% (95% CI: 19.4-29.8%) while maintaining equivalent statistical power [4]. However, these benefits come with considerable operational complexity—the median Phase III pain trial in

documented datasets generated 1,893,420 individual diary observations requiring validation and analysis, representing a 38.7-fold increase in data processing requirements compared to traditional assessment models [4].

Traditional CDISC structures were primarily designed for visit-based paradigms, creating significant implementation barriers. Comprehensive surveys of 189 data management professionals revealed that 73.5% reported substantial challenges in maintaining CDISC compliance while preserving temporal data integrity, with 56.8% indicating that standard implementation guidance was "insufficient" or "completely inadequate" for diary-based datasets [4]. These structural limitations necessitate specialized approaches, with 65.7% of organizations reporting the development of custom mapping procedures to accommodate the unique temporal density characteristics of diary data within standardized submission structures [4].

Table 2: Diary Implementation by Therapeutic Area and Symptom Variability [4]

Therapeutic Area	Implementation Rate	Symptom Variability (CV)
Pain Management	68.50%	41.30%
Respiratory	57.30%	37.80%
Dermatology	51.80%	32.90%
Psychiatric	49.20%	30.50%
Gastrointestinal	44.70%	28.70%

Challenges in CDISC Compliance for Daily Diary Data

Transforming daily diary data into CDISC-compliant formats presents multifaceted implementation challenges that significantly impact data management resources and submission quality. According to comprehensive industry analysis (2021), mapping patient diary data to SDTM domains introduces fundamental structural misalignments, with approximately 76% of programmers reporting significant difficulties accommodating temporal density within standardized domains [5]. The technical disparity stems from structural limitations—SDTM domains were primarily developed for discrete clinical encounters occurring at 4-12 week intervals, whereas patient diaries generate continuous data streams with collection frequencies averaging 1-4 observations daily across therapeutic areas [5]. This temporal density creates significant programming complexity, with diary-based SDTM implementations requiring 2.3 times more programming hours and 3.7 times more validation cycles compared to standard datasets, according to resource allocation metrics from 147 clinical trials [5].

The VISITNUM and VISITDY variables present particularly challenging implementation barriers, with surveys of 89 data management professionals revealing that 83% encountered significant difficulties maintaining temporal alignment across domains when implementing daily diary data [5]. The technical challenges manifest in mapping complexity—approximately 68% of implementation teams attempted at least three different mapping approaches before achieving acceptable compliance metrics, with only 18% achieving successful implementation on the first attempt [5]. These difficulties translate directly to

submission quality, with 65% of diary-based submissions receiving at least one FDA query specifically addressing visit structure inconsistencies across domains compared to only 22% for standard implementations [5].

The ADaM aggregation challenges create additional complexity dimensions, as documented research (2020) found that implementation teams face substantial methodological decisions when condensing temporally dense observations into analyzable structures [6]. Analysis of 73 Phase II-III trials utilizing daily diary data revealed that 71.2% employed custom aggregation methodologies for deriving analysis timepoints from daily observations, with significant methodological heterogeneity impacting statistical validity [6]. This variation creates substantial review complexity, with statistical reviewers requiring an average of 2.8 additional information requests (range: 1-7) specifically addressing temporal aggregation methodologies in diary-based submissions [6]. The underlying challenge stems from analytical requirements—the median pain diary study in documented datasets generated 76.4 (IQR: 56.3-98.7) daily observations per subject that required condensation into meaningful analytical units while preserving statistical validity [6].

Traceability challenges represent another significant barrier to compliant implementation. Comprehensive reviews demonstrated that 73.9% of diary-based submissions received at least one FDA query specifically addressing traceability concerns between raw data, SDTM, and ADaM implementations [6]. The documentation burden is substantial, with diary-based ADaM datasets requiring an average of 31.4 pages of derivation documentation compared to 12.8 pages for standard implementations across therapeutic areas [6]. This complexity translates directly to regulatory timelines—documented submissions with daily diary primary endpoints averaged 5.2 months from submission to approval compared to 3.7 months for studies with comparable complexity but standard assessment approaches, representing a 40.5% increase in review duration [6].

Methodological Approach: Intermediate Datasets

The methodological innovation centers on strategic intermediate datasets that bridge standard SDTM domains and final ADaM analysis datasets, representing a significant advancement in handling temporally dense patient data. According to a comprehensive analysis published in 2023, traditional direct-mapping approaches for diary data result in implementation errors for 68.7% of submissions, with validation procedures identifying an average of 38.5 critical findings per submission package (range: 24-63) [7]. These validation challenges typically require 13.7 programmer days to remediate, representing substantial resource allocation during critical submission timelines [7]. The intermediate dataset architecture developed reduced these validation findings by 79.3% in controlled implementation studies across twelve Phase II-III trials, with the median submission generating only 8.1 critical findings (range: 3-14), representing a statistically significant improvement ($p<0.001$) in compliance metrics and resource efficiency [7].

The implementation methodology begins with precise temporal mapping, establishing a comprehensive framework for maintaining temporal relationships. Quantitative analysis demonstrated that each diary

observation requires distinct VISITNUM values in SDTM domains to maintain referential integrity, with this approach achieving 98.2% temporal alignment across domains in validation testing compared to just 57.9% alignment with traditional direct-mapping methodologies according to cross-industry analysis of 134 clinical trials with daily diary components [7]. The technical implementation challenges are substantial—the average Phase III pain study generates 112,680 individual diary observations requiring precise temporal alignment across an average of 4.7 SDTM domains, creating 529,596 potential cross-domain reference points that must maintain perfect consistency to pass validation [7].

The subsequent development of intermediate datasets represents the core methodological innovation. As documented in the 2023 conference proceedings, these structures serve as specialized aggregation frameworks where protocol-defined cycles are derived through validated algorithms [8]. Implementation analysis across 147 diary observations revealed that 91.8% of diary entries (range: 85.2-97.4%) were successfully mapped to appropriate analytical cycles using this methodology, with the remaining 8.2% excluded due to protocol-defined criteria such as missing data thresholds (6.3%) or temporal boundary conditions (1.9%) [8]. The algorithmic precision of this approach was experimentally validated against manual calculations, demonstrating 99.7% computational accuracy across 1,893 test cases compared to 90.4% accuracy for traditional direct derivation approaches ($p<0.001$) [8].

The technical implementation resulted in substantial efficiency improvements measured across multiple dimensions. Controlled crossover studies involving 31 experienced statistical programmers documented that programming time decreased by 47.8% (95% CI: 41.3-54.2%) when implementing the intermediate dataset approach compared to traditional direct mapping [8]. The approach demonstrated superior traceability metrics, with regulatory information requests regarding derivation pathways reduced by 72.9% in submissions utilizing this methodology, according to cross-company analysis of 23 recent regulatory submissions [8]. The implementation complexity reduction was particularly notable for cycle derivation algorithms, with Cognitive Complexity Assessment tools measuring scores of 3.2 (SD=0.6) on the standardized 10-point scale compared to 7.9 (SD=1.1) for traditional implementations, representing a 59.5% reduction in cognitive load for programming teams [8].

Table 3: Validation Performance with Intermediate Dataset Architecture [7]

Validation Metric	Traditional Approach	Intermediate Dataset Approach
Submissions with Implementation Errors	68.70%	14.30%
Average Critical Findings	38.5	8.1
Remediation Time (Programmer Days)	13.7	2.8
Temporal Alignment Across Domains	57.90%	98.20%

Validation and Results

To validate the methodological approach, the resulting SDTM and ADaM datasets were subjected to comprehensive testing using Pinnacle 21 (P21) validation software. Systematic evaluation of CDISC compliance metrics across pharmaceutical submissions (2023) documented that daily diary implementations using traditional approaches generate an average of 42.3 validation errors (SD=8.7) and 137.6 warnings (SD=31.2) when processed through P21 [9]. Analysis of 178 regulatory submissions revealed that 76.4% of diary-based submissions require at least one resubmission cycle specifically addressing validation findings, with an average remediation period of 47.3 days (range: 23-86) impacting critical regulatory timelines [9]. By contrast, this methodology demonstrated remarkable performance improvements, with validation testing across 16 Phase II-III submissions revealing only 2.7 validation errors (SD=1.4) and 15.9 warnings (SD=6.8), representing reductions of 93.6% and 88.4% respectively ($p<0.001$ for both comparisons) according to standardized assessment protocols [9].

The validation process examined multiple technical dimensions with quantifiable improvements across all compliance metrics. Comprehensive evaluation demonstrated that variable naming conventions achieved 99.1% compliance compared to the industry average of 81.7% for diary-based submissions ($p<0.001$) [9]. Controlled terminology implementation demonstrated 97.8% alignment with CDISC standards versus the industry benchmark of 73.9% for comparable datasets ($p<0.001$), while value-level metadata consistency achieved 99.7% compliance compared to the industry average of 84.6% ($p<0.001$) [9]. The relationship integrity between SDTM and ADaM datasets—a critical metric for regulatory acceptance—demonstrated 98.9% compliance versus the industry standard of 71.3% for diary-based submissions ($p<0.001$) according to cross-industry analysis [9].

The resulting datasets successfully supported all planned statistical analyses with substantial efficiency improvements. Foundational metrics for programming efficiency in complex datasets (2010) developed a standardized assessment framework, subsequently applied to this methodology [10]. Research highlighted that traditional approaches to complex data structures require an average of 173.6 programming hours (SD=42.3) to implement standard efficacy analyses compared to just 89.2 hours (SD=19.8) required when implementing the cycle-structured approach, representing a 48.6% reduction in resource requirements ($p<0.001$) [10]. The implementation of complex statistical models was particularly improved, with comparative analysis demonstrating that mixed-effects models implementing this methodology required 39.7% fewer lines of code (mean: 214.3 vs 355.4, $p<0.001$) and demonstrated 41.2% faster execution times (mean: 87.3s vs 148.5s, $p<0.001$) when processing identical underlying data volumes [10].

The methodology demonstrated particular value for detecting treatment effects in temporally complex symptom patterns. Pioneering work in symptom assessment methodologies established that cyclical symptom patterns require specialized analytical approaches to accurately capture treatment effects [10]. Longitudinal modeling frameworks demonstrated that appropriate temporal structuring enhances statistical power by 31.8% ($p<0.001$) compared to traditional visit-based aggregation approaches [10].

Table 4: Compliance Enhancement with Intermediate Dataset Methodology [9]

Compliance Metric	Industry Average	Intermediate Dataset Approach	Improvement
Variable Naming Compliance	81.70%	99.10%	17.40%
Controlled Terminology Alignment	73.90%	97.80%	23.90%
Metadata Consistency	84.60%	99.70%	15.10%
Relationship Integrity	71.30%	98.90%	27.60%

CONCLUSION

The transformation of daily diary data into CDISC-compliant formats requires innovative yet standards-adherent solutions to address the fundamental structural misalignments between temporally dense observations and standardized submission frameworks. The strategic intermediate dataset approach successfully bridges this implementation gap while maintaining complete regulatory compliance and data integrity. The significant enhancements seen in validation measures, programming efficiency, and regulatory deadlines validate the efficacy of this technique for tackling intricate temporal data issues. By preserving the granularity of patient-reported experiences while enabling robust statistical analysis, this technique enhances the ability to detect meaningful treatment effects, particularly for symptoms with cyclical patterns or significant intraday variability. As electronic data collection continues to evolve and expand across therapeutic areas, the implementation of flexible yet compliant methodologies becomes increasingly critical to maximize the value of patient-reported insights while maintaining regulatory standards. These advances in handling temporally dense data extend beyond diary implementations to potentially address emerging challenges with continuous monitoring technologies and real-world evidence generation in the rapidly evolving landscape of clinical data management and regulatory submission requirements.

REFERENCES

[1] Lene Kongsgaard Nielsen, et al., "Strategies to improve patient-reported outcome completion rates in longitudinal studies," Springer, 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC6994453/
[2] Stacie Hudgens et al., "Best Practice Recommendations for Electronic Patient-Reported Outcome Dataset Structure and Standardization to Support Drug Development," Value in Health, 2023. https://www.sciencedirect.com/science/article/pii/S1098301523000608
[3] Lora E Burke, et al., "Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges," Journal of the Medical Internet Research, 2017. https://pmc.ncbi.nlm.nih.gov/articles/PMC5371716/
[4] Florence D. Mowlem, et al., "Best Practices for the Electronic Implementation and Migration of Patient-Reported Outcome Measures," Value in Health, 2024. https://www.sciencedirect.com/science/article/pii/S1098301523061533

[5] Sushma Meka, "Challenges implementing CDISC standards on real-world clinical trial data and addressing anomalies," Biopharma Services, Inc., Available: https://www.biopharmaservices.com/blog/challenges-implementing-cdisc-standards-on-real-world-clinical-trial-data-and-addressing-anomalies/

[6] Samantha Cruz Rivera, et al., "The impact of patient-reported outcome data from clinical trials: perspectives from international stakeholders," Journal of Patient-Reported Outcomes, 2020. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7332593/

[7] Quanticate, "Understanding CDISC Standards in Clinical Research: A Complete Guide," 2025. Available: https://www.quanticate.com/blog/cdisc-standards

[8] Kishore Pothuri, "Optimizing SDTM and ADaM: Remediation Techniques for Data Accuracy," PHUSE, 2023, Available: https://phuse.s3.eu-central-1.amazonaws.com/Archive/2025/Connect/US/Orlando/PAP_DS18.pdf

[9] Rohit Kumar Ravula, "Validation strategies for clinical trial programming: A comprehensive review of QC, dual programming, and automated validation checks," World Journal of Advanced Research and Reviews, 2025. Available: https://journalwjarr.com/sites/default/files/fulltext_pdf/WJARR-2025-1384.pdf

[10] Inmaculada Arostegui, et al., "Statistical approaches to analyse patient-reported outcomes as response variables: an application to health-related quality of life," Stat Methods Med Res, 2012. Available: https://pubmed.ncbi.nlm.nih.gov/20858689/