Paper DH07

Enhancing Data Accuracy and Collaboration through Efficient Data Reconciliation using SAS in Clinical Trials

Piyush Makkapati, Vita Global Sciences, Waltham, MA
Swathi Nattam, Vita Global Sciences, Waltham, MA

ABSTRACT

Reconciling data between Electronic Data Capture (EDC) systems and external vendor databases is vital for ensuring the accuracy and integrity of clinical trial data. Effective, timely reconciliation mitigates discrepancies that could compromise study outcomes. While reconciliation is important throughout the trial, its significance is heightened as database locks (DBLs) approach, where preventing costly delays and data integrity issues post-DBL is critical.

This paper introduces a streamlined reconciliation approach, utilizing key variables within SAS to integrate data from multiple sources. Tailored discrepancy reports and specifications, developed in accordance with Vendor Data Transfer Agreements (DTAs), facilitate collaboration between Data Management (DM) and programming teams.

By addressing discrepancies proactively—particularly near DBLs—this method not only enhances data quality but also reduces the manual workload for DM teams and fosters stronger collaboration, ultimately contributing to the success of clinical trials.

INTRODUCTION

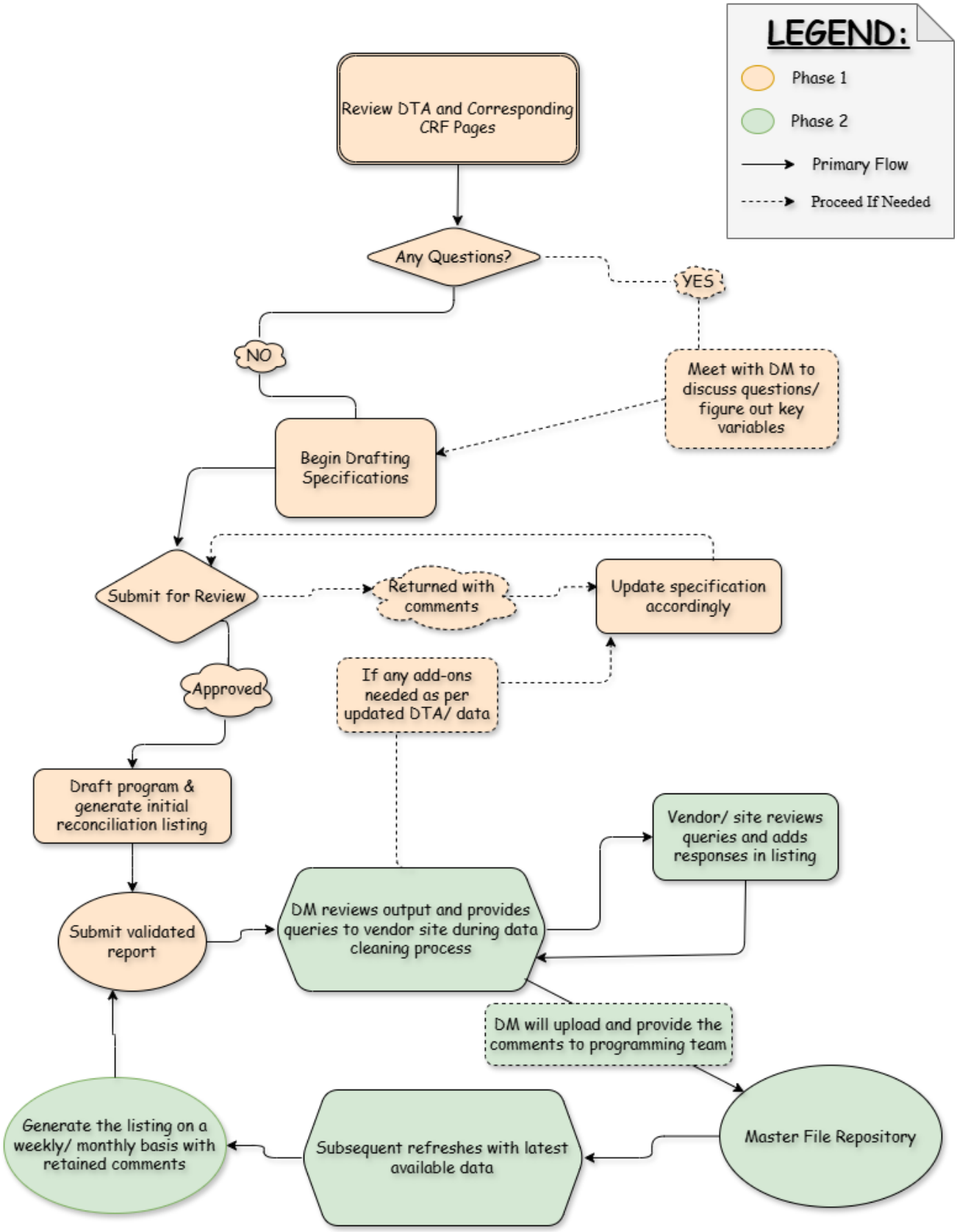
Imagine a pivotal moment in a clinical trial: the database lock (DBL) is only days away, and clinical teams are racing to reconcile data from multiple sources. One minor discrepancy can unravel months of progress by triggering a surge of data queries and frantic last-minute fixes. In an environment where timing, accuracy, and trust are paramount, every misalignment carries the risk of stalled timelines, escalating costs, and shaken confidence in final study results.

Data reconciliation is a critical, but often an underestimated, component to prevent such scenarios. When mismatches between EDC systems and external vendor databases remain undetected, they can easily linger until the worst possible moment. The objective of this paper is to highlight a proactive, SAS-driven strategy for closing these gaps early on before any harm is done to the clinical trial. By carefully building on provided Vendor Data Transfer Agreements (DTAs) and fostering close collaboration between DM and programming teams, we leverage clear-cut, drafted specifications to generate robust discrepancy reports, while strengthening data integrity and conserving valuable time and resources.

BACKGROUND

Data reconciliation is the cornerstone of any successfully managed clinical trial, ensuring data accuracy and consistency across various databases. In practice, it involves merging and comparing data using key variables from external vendor databases with the EDC case report form (CRF) data to identify and resolve inconsistencies. By establishing a robust reconciliation process, clinical teams not only uphold data accuracy but also mitigate the risk of delayed database locks and costly, last-minute corrections. The following sections outline the key components of an effective strategy: Vendor Data Transfer Agreements (DTAs), detailed specifications, strategic use of key variables within SAS programs, and collaboration between data management and programming teams.

Figure 1: SAS-Based Reconciliation Process Flow



THE ROLE OF VENDOR DTAs, CRFs, AND SPECIFICATIONS

An essential step in effective automated data reconciliation is drafting well-defined specifications that draw on both the vendor DTAs and the corresponding CRF pages. DTAs detail the types of data expected from the vendor, such as lab results or serious adverse events, file formats, and frequency of transfer schedules, which may occur weekly, monthly, quarterly, or on an ad hoc basis. The frequency of transfers will directly correlate to how often the reconciliation will be performed. They also showcase the mapping of study visits and other variables in relation to the EDC. CRF pages, on the other hand, clarify how the same data elements are collected and stored in the EDC database. By combining these two references, teams can create documented specifications on how to match and compare information between the two sources. These specifications are commonly compiled in an excel workbook that consists of three core sheets: the Template Sheet, the Programming Instructions, and the Modification History. Within the template sheet, the columns required for the final output are further divided into three sections as seen below in **Figure 2**.

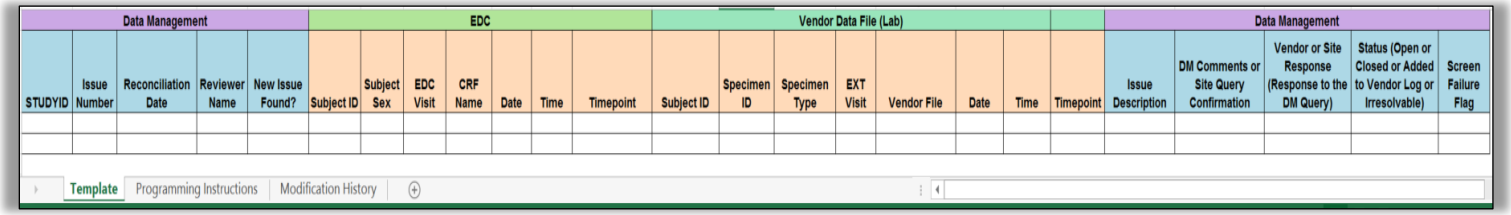


Figure 2: Specification Template Tab

The first is for Data Management entries, allowing DM teams to record the reviewer's name and log queries under "DM comments," to which the site or vendor can respond. The second and third focus on displaying study-specific variables such as specimen type, lab category, or subject demographics of both vendor and EDC data respectively. Once the issue is confirmed as resolved or remains as pending, the DM team can update the status accordingly.

Programming Instructions and Key Variables

In addition to the Template Sheet, the specifications incorporate a Programming Instructions tab that specifically describes how each EDC CRF page aligns with the vendor data file. A crucial element of this tab is the listing of key variables, including subject ID, visit, date, time, and time points which are required for accurately matching records between the two databases. These fields serve as the anchors for comparing data, ensuring that when a subject's information is pulled from the vendor file, it can be aligned to its corresponding record in the EDC system. The programming instructions also outline the primary reconciliation checks to be performed, including standard mismatches in dates or times and special conditions such as screen failures that must be flagged or excluded. Once the team confirms that these key variables are consistently captured across both data sources, they are incorporated into the specifications so that programmers can systematically check for and address any discrepancies. This proactive clarity expedites subsequent steps and helps maintain a smooth workflow when integrating and reviewing the data.

Examples of Common Discrepancies

Discrepancies must be identified and addressed appropriately before critical timelines. They may include records present in the vendor file but missing from the EDC (or vice versa), misaligned dates and times that could interfere with visit windows or dosing schedules, and timepoint-level inconsistencies where incorrect time points are populated for visits. Commonly, these discrepancies can involve visit-level mismatches as well, where a sample marked as "Day 30" in EDC might appear as "Day 60" on the vendor side.

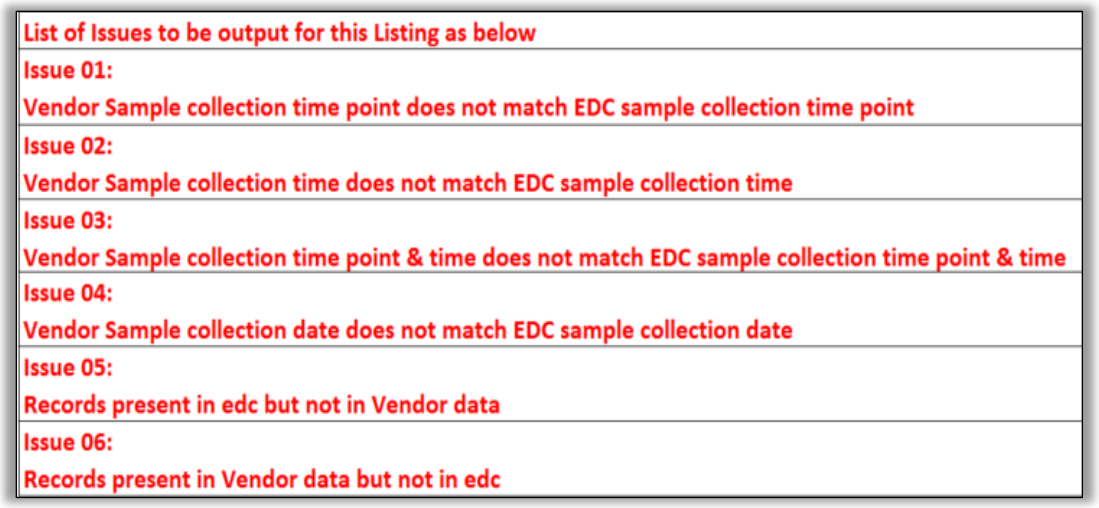


Figure 3: Sample Lab Issue Descriptions

Our SAS-based approach generates automated reconciliation listings that systematically detect and flag mismatches, like those in **Figure 3**, ensuring timely resolution. This approach allows data management and vendor teams to investigate and resolve each issue as efficiently as possible, ultimately maintaining the overall quality and consistency of the study data.

Modification History and Final Steps before Programming

To accommodate for inevitable changes in evolving study requirements or DM needs, the specifications include a modification history sheet that tracks all updates over the course of the trial. Once the reconciliation specifications are fully drafted, the DMs perform a final review. After receiving DM approval, programmers can implement reconciliation programs based on these specifications, generating listings that mirror the established specs as closely as possible. These listings, which lay out each flagged record and relevant data points, are then returned to the DMs for review and resolution. By following this structured workflow, beginning with DTAs, drafting dynamic specifications, and concluding with iterative reviews and modifications, clinical teams can maintain high data quality and minimize delays that could drastically affect study timelines.

IN-HOUSE SAS-BASED RECONCILIATION APPROACH

Following the finalization of specifications, the internally developed programming process produces custom-tailored outputs for Data Management. Designed to accommodate all types of vendor data, the process unfolds in two distinct phases, as illustrated in **Figure 1**. Each phase follows a structured workflow focused on aligning data, identifying discrepancies, and generating reconciliation listings that support timely and efficient resolution.

Phase I - Initial Data Alignment and Listing

In the first phase, the process begins by importing both the EDC and vendor data files into SAS and standardizing key variables such as subject IDs, visits, dates, times, and time points. Maintaining consistent formatting across these fields is crucial, especially when combining multiple CRF pages to align with external vendor files. Inconsistent formatting can lead to incorrect data merging, causing mismatch errors that affect the accuracy of checks. Once the data is cleaned and formatted properly, the two sources are merged (**Figure 4**) using the key variables mentioned in the specifications, and the programmed mismatch checks identify and flag any discrepancies with a corresponding issue description.

```
/*---Merge EDC & EXT and check match/un-match records---*/
data Edc_ext Edc_only Ext_only;
    merge Edc(in=a) Ext(in=b);
    by Subjid Visit Date Time Timepoint;
    if a and b then output Edc_ext;
    if a and ^b then output Edc_only;
    if ^a and b then output Ext_only;
run;
```

Figure 4: EDC and EXT Data Merging Step

The final output is a single excel sheet that contains all identified reconciliation issues, with the sheet name reflecting the most up to date EDC and vendor data files used for that run. Next, it undergoes thorough validation before being submitted to Data Management for review, serving as the foundation for subsequent listing runs in Phase II.

Phase II - Iterative Refreshes and Master Files

After the reviewed listing is returned from Data Management with feedback, the previously submitted listing gets used as a master file. When a subsequent reconciliation run is needed (often weekly or monthly), issues identified from using the latest EDC and vendor data are merged with the master file (**Figure 5**), ensuring that any populated comments and statuses for ongoing issues remain attached to the records.

```
data Curr_old Curr Old;
    merge Curfile(in=a) Master_dm(in=b);
    by Subjid Edcvisit Edcform Edcdataset Edcdate Edctime Edctimepoint
    Extvisit Extdate Exttime Exttimepoint Issuedescription;

    /*----Issues still existing (New & Old identified issues)---*/
    if a and b then output Curr_old;

    /*---Only new issues---*/
    if a and ^b then output Curr;

    /*---Only resolved issues---*/
    if b and ^a then output Old;
run;
```

Figure 5: New Listing & Master File Merging Step

Discrepancies that still exist are flagged as old, newly identified discrepancies are marked as new with the current reconciliation date, and issues that no longer trigger any mismatch checks are moved into the resolved section. The final output in this phase consists of two sheets (**Figure 6**): the first lists all active issues (old & new) and notes which EDC and vendor data is used, while the second sheet preserves a record of resolved issues for any auditing or historical purposes. The updated output also goes through the validation process and gets resubmitted to Data Management for review, repeating this cycle until all discrepancies are addressed and the study database is ready to be locked.

Data Management					EDC							Vendor Data File (Lab)									Data Management				
STUDYID	Issue Number	Reconciliation Date	Reviewer Name	New Issue Found?	Subject ID	Subject Sex	EDC Visit	CRF Name	Date	Time	Timepoint	Subject ID	Subject Sex	Specimen ID	Specimen Type	EXT Visit	Vendor File	Date	Time	Timepoint	Issue Description	DM Comments or Site Query Confirmation	Vendor or Site Response (Response to the DM Query)	Status (Open or Closed or Added to Vendor Log or Irresolvable)	Screen Failure Flag
Sample Study	1	3-Nov-2024		N	ADC01-001	F	Day 2	Plasma PK	3/1/2024	11:03	24H	ADC01-001	F	J027AJQ2	Plasma	Day 2	ADC_lab_01NOV2024	3/1/2024	11:08	24H	Vendor Sample collection time does not match EDC sample collection time	Queried Vendor	Sample not received yet	Open	
Sample Study	2	3-Dec-2024		N	ADC01-002	F	Day 3	Plasma PK	6/20/2024	10:26	48H	ADC01-002	F	J05HPON7	Plasma	Day 3	ADC_lab_01DEC2024	6/15/2024	10:26	48H	Vendor Sample collection date does not match EDC sample collection date for a Visit	Should resolve next round		Open	
Sample Study	3	3-Jan-2025		Y	ADC01-003	M	Day 36	Plasma PK	10/28/2024	10:05	Predose	ADC01-003					ADC_lab_01JAN2025				Records present in edc but not in Vendor data				

EDC01JAN25_EXT02JAN25_03JAN25 | old_resolved_issues

Figure 6: Sample Lab Output

Process for more Complex Recons: SAE Reconciliation

Although Phases I & II remain as the foundation for every EDC and vendor data reconciliation, serious adverse events (SAEs) data require additional attention. Multiple CRF pages capturing data such as adverse events and deaths must be consolidated into one EDC dataset and then be merged with the external vendor safety data. The method of aligning the key variables from both sources persists, however, the final output consists of a complex layout that is arranged into four sheets. The first sheet (**Figure 7**) displays the total number of SAEs from the EDC and vendor data files and also indicates how many records that pass through the mismatch checks are considered a match.

STUDYNAME Total SAE Counts	#
Total Number of SAE's from EDC	98
Total Number of SAE's from Vendor	98
Matched SAE's	80

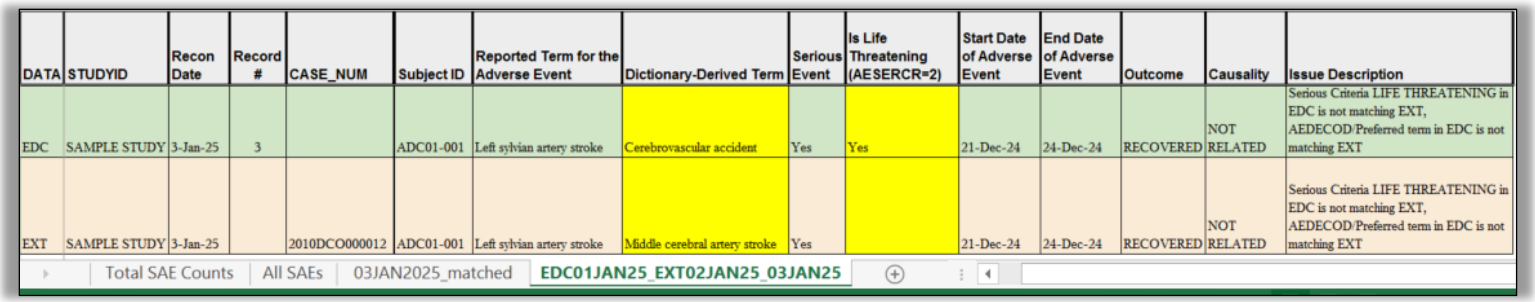
Total SAE Counts | All SAEs | 03JAN2025_matched | EDC01JAN25_EXT02JAN25_03JAN25

Figure 7: Total SAE Counts Tab

The second sheet provides a complete list of all current records from both sources. The third sheet displays only those events that successfully pass the mismatch checks (**Figure 8)** and are considered resolved as of the current listing date. The fourth sheet presents records that continue to show discrepancies, including previously flagged issues that have been reviewed by the DM and vendor/site teams with updated comments with these issues clearly highlighted (**Figure 9**).

List of Issues to be output for this Listing as below	
Issue 01: Records present in edc but not in Vendor data	Issue 06: Serious Criteria DISABILITY in EDC is not matching EXT
Issue 02: Records present in Vendor data but not in edc	Issue 07: Serious Criteria CONGENITAL ANOMALY/ BIRTH DEFECT in EDC is not matching EXT
Issue 03: AEDECOD/Preferred term in EDC is not matching EXT	Issue 08: Serious Criteria OTHER IMPORTANT MEDICAL EVENT in EDC is not matching EXT
Issue 04: Serious Criteria LIFE THREATENING in EDC is not matching EXT	Issue 09: Death date in EDC is not matching EXT
Issue 05: Serious Criteria HOSPITALIZATION in EDC is not matching EXT	Issue 10: Relationship to Study Drug in EDC is not matching EXT

Figure 8: Sample SAE Issue Descriptions



DATA	STUDYID	Recon Date	Record #	CASE_NUM	Subject ID	Reported Term for the Adverse Event	Dictionary-Derived Term	Serious Event	Is Life Threatening (AESERCR=2)	Start Date of Adverse Event	End Date of Adverse Event	Outcome	Causality	Issue Description
EDC	SAMPLE STUDY	3-Jan-25	3		ADC01-001	Left sylvian artery stroke	Cerebrovascular accident	Yes	Yes	21-Dec-24	24-Dec-24	RECOVERED	NOT RELATED	Serious Criteria LIFE THREATENING in EDC is not matching EXT, AEDECOD/Preferred term in EDC is not matching EXT
EXT	SAMPLE STUDY	3-Jan-25		2010DCO000012	ADC01-001	Left sylvian artery stroke	Middle cerebral artery stroke	Yes		21-Dec-24	24-Dec-24	RECOVERED	NOT RELATED	Serious Criteria LIFE THREATENING in EDC is not matching EXT, AEDECOD/Preferred term in EDC is not matching EXT

Figure 9: Unmatched tab

An additional feature of this reconciliation output is its horizontal layout. Each record from the EDC is directly followed by its corresponding vendor record below, which simplifies the Data Management review process. This thorough approach ensures that the safety databases are accurately reconciled, providing essential information on adverse events and safety concerns that is most crucial for maintaining the accuracy of the study outcomes.

BRIDGING THE GAPS THROUGH COLLABORATION

Our programmatic reconciliation approach, built on utilizing vendor DTAs, CRFs, and carefully drafted specifications, actively fosters close collaboration between data management and programming teams. Working closely from the beginning of this process, DMs provide vital clinical context clarifying permissible discrepancies such as allowing screening date variations or special handling of incomplete data. Programmers are then able to translate these rules into code that systematically identifies and reports the mismatches. DMs have mentioned that early alignment has been particularly beneficial when introducing new types of vendor data, as it allows both sides to address format or mapping details well in advance. This ongoing communication also supports adaptability throughout study progression; if additional variables or mismatch checks become necessary, or if certain checks are deemed irrelevant, teams can swiftly update specifications and programs without disrupting existing workflows. As a result, final listings undergo thorough validation with minimal manual oversight, which boosts confidence among sponsors and vendors. Ultimately, this synergy between DM and programming teams promotes an environment in which data is reconciled efficiently and reliably, safeguarding the overall quality of study results and avoiding last-minute hurdles before database locks.

Mitigating the Risks of Delayed Data Integration: Insights from Data Management

Unresolved discrepancies can become increasingly problematic, and costly as clinical trials progress. When serious data mismatches are discovered too close to the database lock deadline, sponsors risk facing extended timelines, heightened regulatory scrutiny, and diminishing confidence regarding the outcomes of the clinical trial. Frequent reconciliation cycles, conducted weekly or monthly while ramping up as major milestones or DBLs approach, provide a safeguard against such risks. By catching issues in smaller, regular increments, teams can manage corrections efficiently and maintain a clear version history of all changes. As a result, both trial budgets and credibility are better protected.

Recent feedback from a data manager underscores just how vital these timely programmatic reconciliations can be. The DM noted that during the manual reconciliation process, it can take at least one to two days for even basic checks, such as identifying records present in the EDC database but missing from the vendor. The process can quickly become unmanageable as the volume of data swells into hundreds of thousands. For instance, in one study, the arrival of substantial new vendor data less than two weeks away from the

DBL triggered a series of inconsistencies that ultimately delayed the lock by several weeks, increasing the costs and straining resources. According to the DM, these crises have since been avoided by scheduling automated reconciliation runs more frequently in the final phase of the study, thereby reducing last-minute hurdles. The DM also reported that mismatches in visits, time, or time point fields which are often difficult to spot manually, are now more quickly and reliably flagged along with having the option to record retained comments directly in the reconciliation. Over time, this iterative process sharpens the alignment between vendor and EDC data, drastically reducing discrepancy rates especially during critical times such as milestones and DBLs, in part thanks to automated reconciliation listings.

CONCLUSION

In a fast-paced clinical environment where each data point can shape outcomes that directly affect countless individuals, the SAS-based reconciliation approach outlined in this paper is more than just a technical task, it is a critical commitment to accuracy, trust, and timely progress. By leveraging vendor DTAs, key variables, and well-crafted specifications along with integrating data from multiple sources into one cohesive system with early, regular monitoring, teams establish a foundation of reliability that shields studies from last-minute upheavals.

Ultimately, careful oversight of trial data leaves a lasting impression that can be carried forward into future studies. Above all, the end result showcases the power of proactive planning, transparent communication, and shared accountability which modern research demands.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Piyush Makkapati

Company: Vita Global Sciences

Address: 281 Winter St, Waltham, MA 02451

Email: Pmakkapati@softworldinc.com

Author Name: Swathi Nattam

Company: Vita Global Sciences

Address: 281 Winter St, Waltham, MA 02451

Email: Snattam@softworldinc.com