

# Loading Oracle Clinical studies using SAS methodology

Curtis Wolf and Steve Light - DataCeutics Inc.

## ABSTRACT

Oracle Clinical (OC) is becoming a major player in the pharmaceutical industry, as a tool for clinical data management. As it becomes more widely utilized, migrating data into Oracle Clinical from SAS-based legacy systems presents a significant challenge for data management programmers to resolve. This paper discusses two different approaches to the Oracle Clinical batch load process and the upfront work that can make the migration process a whole lot easier.

The topics of this paper include developing a study conversion plan, preparing preload data, using the Oracle Clinical metadata to ensure accuracy, quality control of preload data, conversion of preload data to a load file format, and the actual Oracle Clinical load process.

## INTRODUCTION

This discussion covers two approaches to the SAS to Oracle Clinical migration process. The different approaches, which were partially based on client preferences, gave us experience in different methods. Ultimately these approaches concluded at the same endpoint – a validated migration of SAS legacy data into Oracle Clinical. Based on your needs and/or preferences one of these two approaches or a combination of both may make more sense for you.

## SAS TO OC MIGRATION PROCESS

The two methods that we will be discussing varied at several points, but shared a common set of steps that all migration project must have. These steps are detailed below.

- OC Structure is defined and created
- Annotated case report forms (crf's) corresponding to the OC structure, are produced
- Study plan is developed to remapped SAS legacy data to OC structure which includes renaming variables, merging or splitting data, transposing data, and transforming data values
- SAS legacy data is converted to OC preload format – one record per data value and all required OC identifying information
- OC metadata (study name, DCI, DCM, and other required fields) is joined with SAS Data
- OC fields such as CP Event, sub-event, and repeat sequence number are assigned
- Data is loaded into OC
- Final QC/validation of loaded data is conducted

## TWO METHODS – SAME CONCLUSION

The source data, for these migrations, was residing in SAS-based legacy system. There are several differences to each of the two approaches, but the major point of difference could be summarized in the point at which the validation effort was focused. One approach placed more of the validation effort at the tail end after the OC load and the other focused more on upfront or pre-load validation.

The actual steps involved in migrating the data from SAS to OC are tedious and must be performed with the utmost care so as to

not lose or alter any of the legacy data. Both of our client projects ended in very successful migrations, but along the way we made various adjustments to improve the process. The validation aspect was crucial in ensuring the ultimate success of these projects and as that infamous car repair commercial stated, “You can pay me now, or you can pay me later”. We’re not talking oil filters though. We’re talking about clinical data from perhaps dozens of different studies with a variety of data types and collection schedules.

## METHOD 1 – POSTLOAD QC

The first method is outlined below.

- SAS programs generate DCM equivalent datasets
- OC Questions, DCI, and DCM names assigned based on CRF OC annotations
- OC text load files generated by merging OC metadata via ODBC link with SAS
- OC text file is loaded, errors fixed and reloads conducted
- OC database validated against raw SAS data

In this first method, the bulk of the validation at the post-load point of the process. The earlier steps in the process relied primarily on the accurate annotation of the crf's.

This process worked well as long as there were not too many annotation changes along the way as comparisons were made from one study to the next. The documentation of the SAS legacy system was also very challenging to understand for anyone not familiar with it. This required very tight coordination between the few client staff members who were familiar with the legacy data structures and our staff who was responsible for mapping the data to the new OC specifications.

Once the data was mapped based on the annotations, the data was converted to a predetermined OC load file structure by joining the SAS data with the OC metadata on a real-time basis via ODBC, output into text files, and then loaded into OC. Repeat sequence numbers were automatically assigned during this step as well based on certain key values. Some errors were caught during the conversion process, but most errors were not detected until the actual OC load was conducted. This resulted in multiple re-loads, which were very time-consuming based on network traffic and other system performance issues.

Eventually, we added an additional step that compared the OC metadata to the converted legacy data and produced a mismatch report that we could use to check and correct errors before the actual load was conducted. OC views were generated to extract the OC data back into SAS datasets, which were then compared against the SAS legacy data. This process was repeated until the data was accurately migrated.

## METHOD 2 – PRELOAD QC

The second method is outlined below.

- OC Metadata extracted into SAS Dataset
- SAS program generates pre-load DCM dataset and assigns question, DCI, and DCM names
- OC metadata merged and compared against SAS data
- SAS/OC mismatches corrected and rerun
- Validated SAS results converted into text file and

loaded into OC

- Final testing of loaded data is conducted

The second migration method emphasized validation of the OC load files before actually loading the OC database. OC metadata was extracted into a SAS dataset that was then merged with each converted legacy SAS dataset as part of the load file preparation phase. For each OC DCM we generated a 'pre-load' SAS dataset with the full contents of an OC load file. This dataset was saved and a full QC cycle was run on the data to verify an accurate migration.

This method identified errors in CRF annotation, OC implementation or load file preparation, allowing us to correct errors before any loads were conducted.

Once the pre-load datasets were produced and validated against the extracted OC metadata, the SAS datasets were converted into the standard OC load text file format and then loaded into OC. Problems sometimes occurred at this juncture when the OC definitions had been altered at all since the last extract of the metadata used in load file preparation. This was the main downside to this method.

## CONCLUSIONS

These two methods, while taking somewhat different paths, end up at the same endpoint: a validated OC database. Each method was directly affected by the accuracy of the OC Data Base Coordinator's work, either in the OC CRF annotations or in the actual OC database implementation. The SAS programming steps required in the first method are quicker and can be done by programmers with little knowledge of Oracle Clinical. The second method requires more SAS programming and must be done by staff with some familiarity with OC.

The first method delayed the validation until after the load but also ensured that the most current version of the OC metadata was utilized at the time of the load. The second method relied on validation of the data prior to loading, eliminating repeated load steps and capturing errors earlier in the process. So, as you can see, neither method is perfect or necessarily superior to the other. There are several factors, which should be evaluated prior to choosing one or the other, such as

- **SAS skills of the staff on the project team** – method 1 requires relatively simple preparation of legacy data and uses standard SAS macros to do much of the OC-specific processing. Method 2 requires additional SAS programming but allows greater control over the exact structure of each load file.
- **Stability of the OC database definition** – either method is impacted by changes to data base specifications. Method 1 relies on OC annotations only in preparation of load files, so annotations must be kept current and accurate. Method 2 uses the OC metadata in addition to CRF annotations, but the extracted OC metadata must be kept current.
- **Network performance and OC execution time** – method 1 is more likely to require repeated OC load cycles, depending on network performance and the size of the existing OC data base this can be a constraint.
- **SAS programmer access to the OC metadata** – method 1 requires programmer access to the OC database to prepare a load file, method 2 allows a single user to access the data base to extract metadata and then distribute SAS datasets to programmers, who can use the extracted metadata to generate final load files.

## CONTACT INFORMATION

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

Curtis Wolf

DataCeutics Inc.

1610 Medical Drive,

Suite 208

Pottstown, PA 19464

Phone: 610-970-2333

Email: [Wolfc@dataceutics.com](mailto:Wolfc@dataceutics.com)

