Paper SI13

Automating Module 5 Packaging with Customizable Metadata

Dhananjay Chhatre, Gilead Sciences, Foster City, USA
Kevin Miller, Gilead Sciences, Foster City, USA

ABSTRACT

In today's regulatory environment, study data packages included in regulatory submissions are being thoroughly reviewed and assessed for adherence to CDISC data standards requirements, and future audits focused on dataset and submission package development are the next logical focus area for inspectors. To mitigate audit risks and improve efficiencies and quality through automation, Gilead has embarked on a journey to develop a suite of utilities to automate Module 5 packaging while implementing Biometrics and Regulatory Affairs/Operations checks using customizable defaulted metadata as the starting point. This paper will focus on the strategic process and flow for how utilities work together, enabling automation and improving efficiencies while allowing for study statistical programmer customization for their specific study. It will also present details on the Python utilities/packages driving the automation, as well as the built-in quality-by-design functionality to identify potential issues upfront.

INTRODUCTION

The U.S. Food and Drug Administration's (FDA's) binding guidance document publications require Sponsors to submit standardized study data for future regulatory submissions. [1] [2] [3] The intent of these standards is to enable the Agency to more efficiently review submissions by leveraging technologies and automation, increasing alignment with the Center for Drug Evaluation and Research (CDER) 21st Century Review Process. [4] Standardized data allows regulatory reviewers to more fully understand and characterize the efficacy and safety of each medical product. These standardized study datasets must be packaged with accompanying metadata and ancillary documentation following the electronic Common Technical Document (eCTD) standard for regulatory submissions to CDER (as well as to the Center for Biologics Evaluation and Research [CBER]). [5] Within the eCTD's modular structure, Module 5 (M5) contains the tabular listings of clinical studies, study reports, case report forms, and datasets. [6]

While various tools exist to create the appropriate eCTD content leveraged by a Sponsor's Regulatory Affairs/Operations teams, the M5 dataset package content itself is often prepared and vetted by the Biometrics department. While this can be done manually, this leaves room for errors and missed checks which may get overlooked downstream. [7] To supplement existing dataset and define.xml validation, Gilead is developing a process-based, metadata-driven automation solution to reduce the burden on both Biometrics and Regulatory Affairs/Operations, while also seeking to improve efficiencies and quality. [8]

AUTOMATION STRATEGY

Leveraging metadata and automation was based heavily on maximizing flexibility, both for the study programmer creating the final M5 package contents, as well as those groups (e.g., Regulatory Operations, eSubmissions Teams) overseeing quality of the final package. Understanding the eCTD study dataset folder directory structure allows us to easily map our working area folders to the target eCTD M5 folders (Figure 1). Additionally, the formal strategy would need to account for manual touch points (accounting for study-specific changes) and the automated processes to promote and facilitate user adoption.

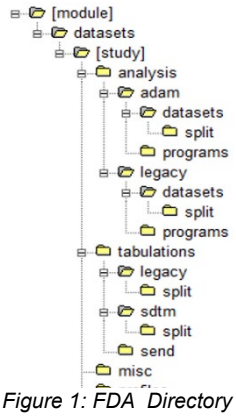

Figure 1: FDA Directory Structure for Datasets

PROCESS FLOW

By defining standard source-to-target mappings, we can pre-populate metadata based on a user's specified working location. This metadata approach allows the user to update beyond the automation assumptions, to handle any nuances needed for their study. To support this flexibility, formalized process flows were developed to describe how the automated steps would support manual interventions, and vice versa (Figure 2). Within that flow, we considered standard components of the M5 package (i.e. Study Data Tabulation Model [SDTM] and Analysis Data Model [ADaM]), as well as optional components (e.g. Miscellaneous and Profiles content) that

may be needed for a study submission. [9] The end-to-end process is transparent for potential audits, and clearly documents traceability between a user's working area and the final M5 package contents submitted to a regulatory agency.

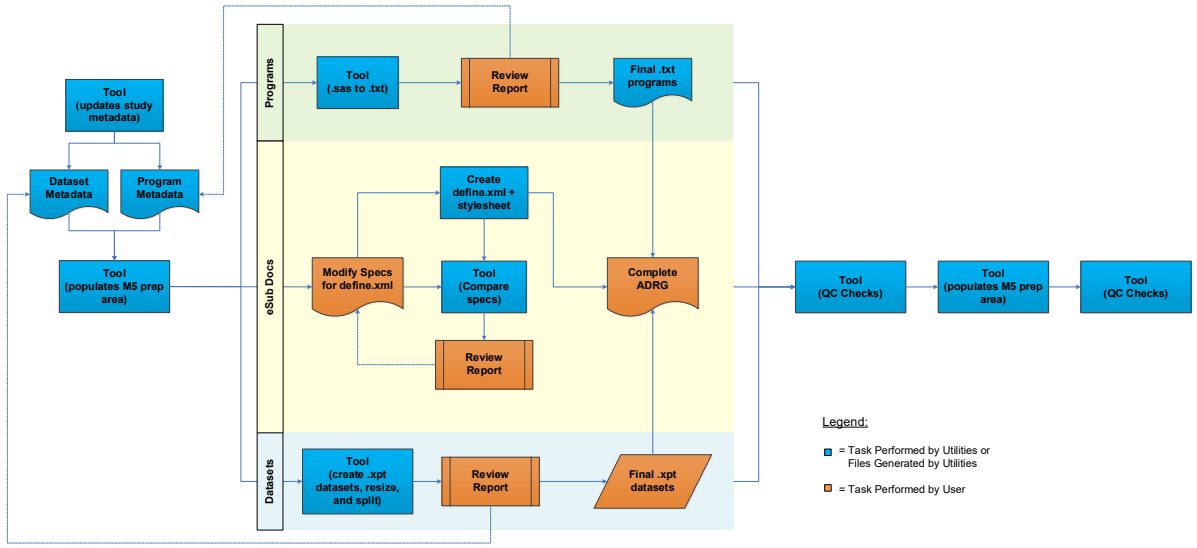

Figure 2: Process Flow for Automation Utilities in M5 Package Generation

SDTM would follow a similar flow as an ADaM package for a study, as well as for Miscellaneous and Profiles content. The nuance is that each of these four categories (i.e. SDTM, ADaM, Miscellaneous, Profiles) would have their own metadata file. In this way, if a change is made in a working area that leads the population in one package type, the other metadata files are not updated. This would preserve date/time stamp traceability between the source (i.e. metadata file) and the target (i.e. M5 package contents).

METADATA SOURCE FILES

Each of the four categories of M5 package content have their own respective metadata file that will drive automation and traceability from working area to regulatory submission. While the structure can be customized by each Sponsor, we took a minimalist approach and focused on a simple "File Name", "Source Folder", and "Target Folder" trio that would apply to all M5 package files. Understanding the content allowed us to default "Target Folder" for specific file types and derive the "Source Folder" based on the run location of the initial metadata-populating tool. Generally, we captured the below metadata (Table 1), with blue-highlighted cells depicting pre-populated values that should not be altered, to further limit the changes that the study programmer needs to make for study customization.

	Active	File Name	Source Folder	Target Folder	Split	Split Name	Where Clause
Datasets	Y	Y	Y	Y	Y	Y	Y
M5 Documents	Y	Y	Y	Y			
Programs	Y	Y	Y	Y			

Table 1: Summary of Metadata Being Captured, by File Type

The Active column allows the user the flexibility to adjust or update content from their working area which should or should not be included in the final M5 package, allowing study customization from the default state. This simple strategy ensures correct implementation by the user, specifically when adding new rows within the standard metadata file. The dataset source files are unique in that users would need to populate metadata around the specific splitting of SDTM datasets (currently when greater than 5 gigabytes [GB]).

AUTOMATION UTILITIES

Development of automation utilities allows for maximum benefit from standardized metadata input files, enabling the user to update a single source and reap benefits through the now traceable M5 package generation process. In our process, we

identified seven automation utilities that needed to either be developed or customized to align with the new metadata-driven approach, with the following mandates:

1. Populate metadata for a given study based on run path and existing files
2. Copy files from source to target (M5 prep area), as specified in all metadata files
3. Convert .sas programs to .txt files
4. Compare define.xml specs in M5 prep area to working area for traceability purposes
5. Convert .sas7bdat datasets to .xpt
6. Resize .xpt datasets to maximum length used
7. Split .xpt datasets, as defined in metadata file
8. Perform quality control checks, as defined by metadata, on M5 prep area contents
9. Copy files from M5 prep area to final M5 location for package processing by Regulatory Affairs/Operations

PYTHON PACKAGES

In lieu of developing automation utilities in SAS, we elected to leverage the open source Python programming language. We felt the processing speed, requirements and volume lent itself to Python. The open-source nature of Python has given power to developers to create incredibly robust and useful modules. The eSubmissions (eSub) process lends itself to utilizing these modules. A few examples of these modules follow: using XlsxWriter to create user interactive Excel sheets; pandas for interaction with imported data frame sources; and pyPdf for sourcing metadata information about pdfs and other packages for integrating with documents like xpt data files, database info, and text files. Python modules allow for the quick and efficient movement and copying of files, directory checking and creation, as well as regular expression and searching capabilities. All of these are crucial necessities for the functionality and end-to-end processing of the eSub process.

METADATA-DRIVEN QUALITY CONTROL (QC) CHECKS

While developing the metadata files, we realized that we could also use metadata to drive our QC checks of M5 package contents, using various flags and folder locations as triggers for specific checks with varying severities (i.e. errors vs. warnings). In this way, maintenance and updates to checks are streamlined, allowing for easier future updates and improved planning for new future checks.

COMPILATION OF CHECKS

In developing our QC checks, we sought to build upon existing checks and regulatory expectations, pulling from various sources like the FDA, internal departments, and published papers. Specifically, we compiled existing checks from:

- Gilead/Submission package checklists
- Gilead/Phase I-specific checks
- Gilead/Regulatory Operations checks on M5 packages received from Biometrics
- Checks mentioned in various papers [10], [11]
- FDA technical rejection criteria [12]

As these did not constitute an exhaustive list, we then supplemented with checks defined during an internal review of FDA (as well as Japan's Pharmaceuticals and Medical Devices Agency [PMDA]) guidance documents, presentations, and supporting documentation that would enable generation of a clean M5 package. Although many checks were compiled, not all were programmatically feasible (or at least easily programmable). The manual QC checks were categorized and specified in an updated submission checklist that each study programmer would need to review and execute prior to M5 package finalization. Standard report outputs were designed to simplify the review of the QC process and provide users with a quick and efficient resolution to the findings. An example report template is below (Figure 3), showing the various metadata presented to allow for easy understanding of errors and warnings.

Status	Source	Tab	Type	Folder Path	File Name	Variable	Value	Check Description
ERROR	tabulations_data.xlsx	Datasets	SDTM	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\tabulations\sdtm\	suppae.xpt			The file is found in corresponding module destination location but is missing from prep destination location.
ERROR	analysis_data.xlsx	Datasets	ADaM	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\analysis\adam\dataset	adsl.xpt			Subject-level analysis dataset (ADSL) must be submitted.
ERROR	analysis_data.xlsx	eSub Docs	Legacy	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\analysis\legacy\datase	define.xml			The file is missing from the specified source location but is found in the specified destination location.
ERROR	analysis_prgm.xlsx	Programs	ADaM	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\analysis\adam\prograr	t-demog.sas			Only program files should be in the programs folder and all programs should be TXT files.
ERROR	general	N/A	N/A	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\analysis\	adam			The sdtm and adam folders must be present.
ERROR	general	N/A	N/A	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\tabulations\legacy\	sdrg.doc			Invalid file type found.
WARNING	tabulations_data.xlsx	Datasets	SDTM	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\module5\tabulations\sdtm\	lb.xpt	LBTESTCD	EGFRMDRD	Variable length for --TESTCD or PARAMCD cannot exceed 8 characters in length. Variable may be truncated.

Figure 3: Example QC Checks Report

In total, we identified over 100 checks looking at various components of the M5 package contents:

- Datasets
- Define.xml
- Structure (are items where they should be)
- PDFs
- Programs

A few examples of QC checks and why they were deemed important are described in subsequent sections.

EXAMPLE 1: PRESENCE OR ABSENCE IN METADATA AND/OR M5 FOLDER

Checks will fire when there is a disconnect between a file listed in the metadata file and what is present in the M5 prep or final folder areas. This can happen if someone accidentally deletes or renames a file in the folder or changes Active status to N (from Y). This check captures the very dangerous scenario where a statistical programmer thinks they are including a dataset, program, or other document, since it is in their working area, but it was accidentally deleted, renamed, or moved to the wrong M5 folder area. Missing a required file, or including a file that should not be present, could lead to future technical rejections or additional regulatory questions, in addition to extra processing and review time across departments.

EXAMPLE 2: DATE/TIME STAMP MISMATCH

The date/time stamp is an important element in determining the correct progression of events, from metadata updates to execution of automation utilities to copying of appropriate files into the M5 prep/final areas. Therefore, if the metadata file has a later stamp than the files for that category (e.g. SDTM datasets), then there is a break in traceability and needs to be investigated. Although it is possible that the metadata file has no significant changes, the user should confirm that the automation utilities do not need to be re-executed to account for a salient update (e.g. datasets need to be split per metadata guidelines). This check captures the dangerous scenario where a package is prepared, but some change is required in the metadata file, such as including an additional dataset, programs no longer being included, or renaming of files. If this gets missed, we would be submitting incorrect M5 package contents for regulatory review, potentially adversely impacting their decision.

EXAMPLE 3: SPLITTING OF DATASETS

Per FDA expectations, datasets >5 GB in size need to be split for their tools and machines to be able to open and process them. As such, even if users follow the formalized process, they may be generating .xpt datasets that are too large. Therefore, a check will fire reminding the user to update their metadata file to provide details on split dataset name(s) and the where clause by which the parent dataset should be split. This check captures the dangerous scenario where a submission with large datasets cannot be reviewed by the regulatory agency, requiring a pause in the review clock and a delay in approval.

METADATA-DRIVEN EXECUTION OF CHECKS

A separate metadata file exists which contains a list of all the checks being executed, which also contains various flags and triggers. This enables us to run only those checks applicable to a certain condition being met. For example, we can designate a given check to be run for either FDA or PMDA (or both), either in the M5 prep or M5 final folder areas (or both), for SDTM or ADaM (or both), and even for the Miscellaneous or Profiles folder areas (or both). This enables very specific checks to be run depending on what metadata files have been populated, depending on the run location of the initial automation tool. Additionally, the metadata file itself could drive the updating, removal, or addition of new checks which the automation tools could read in, removing the need for updating the tool programming.

CUSTOMIZABLE METADATA FOR STUDY USE

As mentioned previously, we wanted to provide a formal structure with standard rules and assumptions, while allowing users the flexibility to customize M5 package contents as needed for their study. By maintaining a standard metadata file structure and minimizing the required metadata, we hope for a smooth transition and adoption of this new process. Below are three different examples of how the metadata can be customized for unique study use.

EXAMPLE 1: SPLIT DATASETS

The study programmer can configure their tabulations metadata file when their study contains large datasets (typically laboratory data) greater than 5 GB (Figure 4). Specifically, they can define how many split datasets should be created, and the desired where clauses to be used to make the splits.

Active	File name	Split	Split name	Where	Source	Destination
Y	ae.sas7bdat	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
Y	cm.sas7bdat	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
Y	sv.sas7bdat	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
Y	dm.sas7bda	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
Y	vs.sas7bdat	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
Y	ds.sas7bdat	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
N	test.sas7bda	N			\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\
Y	lb.sas7bdat	Y	lb01	lbcat = "CHEMISTRY"	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\split\
Y	lb.sas7bdat	Y	lb02	lbcat = "MICROBIOLOGY"	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\split\
Y	lb.sas7bdat	Y	lb03	lbcat ^in ("CHEMISTRY", "MICROBIOLOGY")	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\sdtmdata\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\tabulations\sdtm\split\

Figure 4: Example of Metadata for SDTM Datasets

Once the appropriate utility is executed, per the above example, the parent dataset will be retained in the final M5 prep folder, and three additional datasets named lb01, lb02, and lb03 will also be created.

EXAMPLE 2: ADDING DATASETS IN MISCELLANEOUS FOLDER

There will be instances when contents of a study's M5 package need to include miscellaneous datasets which don't qualify as either tabulation, analysis or profile datasets. For example, data that is not captured in SDTM but is needed to create ADaM datasets, such as look-up tables or special deviations not collected via CRF, should live in the Miscellaneous folder. (13) There could also be study data to support or feed Tables, Figures, or Listings (TFLs) coming from another study that need to be included in this study's M5 package. As there would be no clear location, the Miscellaneous folder can be used.

EXAMPLE 3: INCLUSION OF NON-TRADITIONAL ELEMENTS (E.G. GLOBAL MACROS)

Companies likely have formal or informal processes or best practices as it relates to standard elements to include in an M5 study dataset package. These could range from datasets, local macros, support files with .inc extensions, etc. Similarly, there may be files that are generally not included, such as global macros, due to worry about intellectual property being released or shared without proper consent. Whatever the rationale or file itself may be, there will likely be scenarios when a regulatory agency (or the Sponsor themselves) decides to include an exempt file, such as a validated, global macro. In this case, given the process flow and business requirements of the automation utilities, the user simply needs to add in the global macro to be included, its source, and copy/paste the destination to match the other programs already pre-identified (Figure 5).

Active	File name	Source	Destination
Y	t-demog.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\prog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	l-demog.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\prog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	g-eff.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\prog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
N	header.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\prog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	init.inc	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\tools\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	newvisid.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\tools\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	a-adsl.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\adamprog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	a-adae.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\adamprog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\
Y	global1.sas	\\gsasdata\biometrics\projects\p123\s1234567\task\module5\adamprog\	\\gsasdata\biometrics\projects\p123\s1234567\task\esub\prep\analysis\adam\programs\

Figure 5: Example of Metadata for ADaM Programs

Once the metadata file is updated, the appropriate automation utility can be executed, and all downstream steps will include this new program (e.g. global1.sas).

CONCLUSION

We believe that metadata-driven automation in development of clinical trial datasets, as well as the packaging and execution of necessary QC checks, is the future, especially given the vast array of technologies already available to us now. Improvements in efficiencies and quality through smart automation relies heavily on consistent use of data standards not just within a therapeutic area or company, but across industry. As we continue to test and modify both the process and the tools at Gilead, we are aware of what next steps are required for scaling implementation across the company. Meticulous planning and preparation for the future includes an initial proof of concept, subsequent (successful) pilots, clear documentation, readily available training, and subject matter experts able to support users as needed. While not easy, we have begun to leverage data standards and novel technologies to better our processes for generating Module 5 packaging to more quickly deliver medicines into the hands of those that need it.

REFERENCES

1. **FDA.** *Providing Regulatory Submissions in Electronic Format – Submissions Under Section 745(a) of the Federal Food, Drug, and Cosmetic Act.* 2014.
2. —. *Providing Regulatory Submissions In Electronic Format — Standardized Study Data.* 2014.
3. —. *Study Data Technical Conformance Guide.* 2014.
4. —. *CDER 21st Century Review Process: Desk Reference Guide.* 2014.
5. —. *Providing Regulatory Submissions in Electronic Format - Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications.* 2019.
6. —. *eCTD Technical Conformance Guide.* 2019.
7. *Optimizing Standards Implementation for Successful eData Submissions to the FDA.* **Chhatre, Dhananjay.** 2015.
8. *Developing, Implementing and Governing End-to-End Standards at Gilead.* **Chhatre, Dhananjay.** 2019.
9. **FDA.** *Study Data Technical Conformance Guide.* 2019.
10. *Ready, Set, Go: Planning and Preparing a CDISC Submission.* **Dalton, Maria.** 2018.
11. *Quality Check your CDISC Data Submission Folder Before It Is Too Late!* **Busa, Bhavin.** 2017.
12. **FDA.** *Study Data Standards in eCTD: What You Need to Know About the New Technical Rejection Criteria.* 2016.
13. *Preparing ADaM Datasets and Related Files for FDA Submission.* **Minjoe, Sandra and Hari, Ragini.** 2018.

ACKNOWLEDGMENTS

We would like to acknowledge and thank the great data standards team at Gilead for their continued dedication and hard work in getting the standards published and governed, as well as the many volunteers who contribute to standards development concurrently with study work. In addition, thanks to Eric Du, Michael Calcagno, and Jennifer Britt for helping to review this paper, and Greg Campbell and Pei Sern for their oversight of the statistical programming technology and data science teams.

CONTACT INFORMATION

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

Dhananjay (DJ) Chhatre
Gilead Sciences
333 Lakeside Drive
Foster City, CA 94404
Work Phone: 650-389-8873
Email: Dhananjay.Chhatre@gilead.com
Web: www.gilead.com