Paper DS06

Harmonizing SDTM at the Source: Designing Collection Instruments that Support Sponsor Standards. 2 years later

Helen Scanlan, AstraZeneca, Cambridge, UK
Sue Huxtable, Covance, UK
Gerard Hermus, Covance, Netherlands

ABSTRACT

During Phuse 2017 we gave a presentation about a newly started collaboration around harmonizing SDTM at the Source. That presentation won the best in stream prize and many requested a follow up. Now we are back to tell you about how our partnership collaboration between Covance and AstraZeneca has identified a safe, fast way to build conformant SDTM, by collecting what you submit.

This paper will tell you what has happened with our vision we stated 2 years ago to have a shared set of standards, shared governance, with a shared goal driving advantages for both organizations. We will share what has been implemented from a standards perspective and discuss some of the challenges we faced. Has it saved time, been cost efficient and ended up in good quality as was our original vision? And where do we go from here?

INTRODUCTION

AstraZeneca (AZ) and Covance have for 2 years collaborated in a partnership around Data Management (DM) and Analysis and Reporting (A&R) in their Early Clinical Development (ECD) studies. We work alongside each other, to provide effective, seamless management of the Early Clinical Development clinical trials.

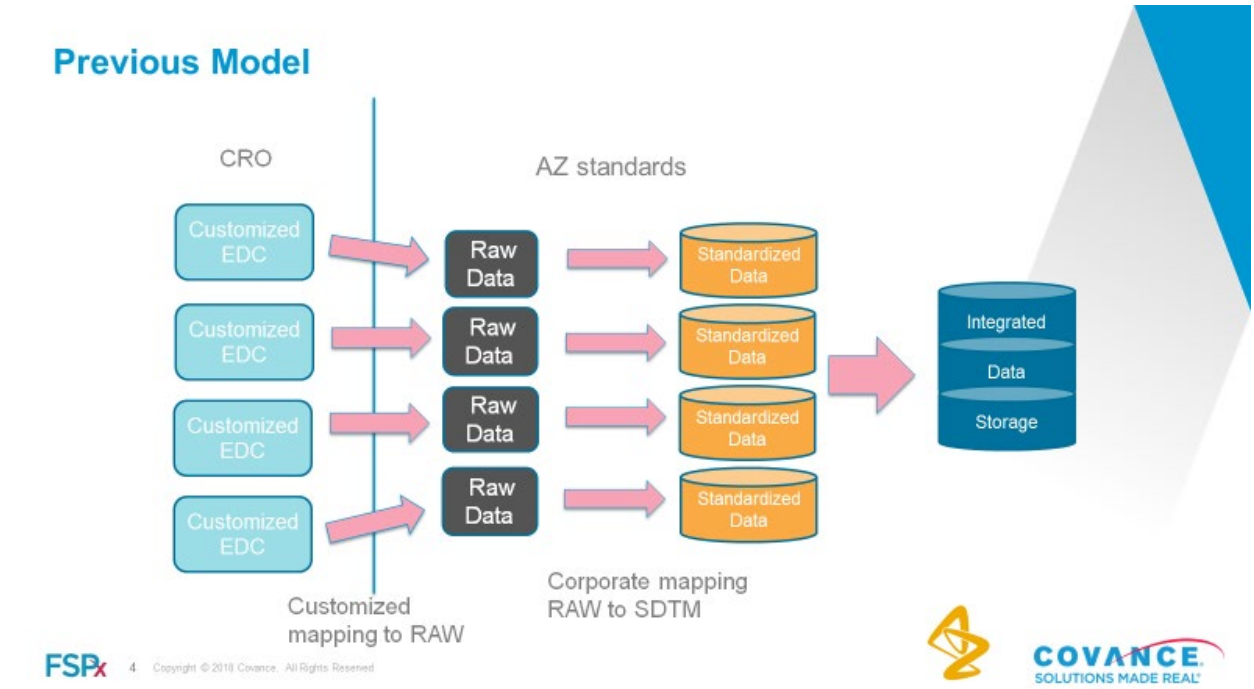
Whilst there are many strands to this relationship, this paper focuses on the work done from a standards perspective. AstraZeneca has a large and detailed library of standards, we have taken the best of the library that AZ has and combined it with Covance technical Electronic Data Capture (EDC) / SDTM expertise, so that we develop a seamless process from collection through to SDTM generation.

AstraZeneca's operational goal is to 'Get Closer to the Data', with an overall plan to define best practice for ECD studies.

HISTORY

AZ have clearly defined standards and a long-standing process for approval of new objects. Each TA has a standards team who manage their individual TA standards, there is a broad, cross-TA standards organization that has responsibility for CORE standards. AZ have implemented an off the shelf product, customized to AZ requirements, to manage library requests, including a submission request workflow. These managed standards in ECD projects, focus on the SDTM end, domains, variables and controlled terminology. The requests are determined to be allowable at the study level, or likely to be reused and therefore promoted to TA or CORE library standards. The approval process can take up to 10 weeks. AZ standards are defined to apply across all phases, all TAs, sometimes these are more focused / applicable to later phases than early development. No rules exist that are specifically focused on ECD type studies, which generally need more flexibility and quicker turnaround times.

The below model is used for late phase studies, formerly known as Global Medicines Development (GMD) studies. Standards are defined at both the RAW and SDTM mapping level, there is also standardised mapping between the RAW and SDTM definitions. This ensures consistency for TFL generation and reporting and forms the basis for ADaM generation. The standards are comprehensive and should include all data required to be collected per the protocol. Derived data is kept to a minimum. These standards are shared with the CRO. The CRO is required to submit the data compliant to the RAW data standard definitions.



For legacy ECD studies, the model implemented was slightly different and the CRO is allowed to work outside the RAW Data Standards, only providing data mapped to the AZ SDTM model. This gives the CRO an increased amount of flexibility around the eCRF design phase, needing a higher level of review than should normally be required if standards were implemented fully. Whilst this has allowed flexibility to enable the CROs to respond to challenging timelines and novel/exploratory study design, it has ultimately led to a significant level of differences of implementation at the EDC level. Therefore, much of the benefit that highly defined standards provide in consistency of front-end design and setup, has been lost. Increased variability from study to study, ultimately can slow down the review process and any gains from following a more flexible model may be lost.

WHERE WE ARE TODAY?

RAVE EDC LIBRARY

Working in partnership, a library of standards was developed. The original version contained around 25 eCRFs. The library has since been expanded to include 47 eCRFs, related edit checks, custom functions and derivations. Two versions have been published, with another planned for early 2020. The library eCRFs cover the base core set for a study and the mandatory AZ eCRFs.

Below are our originally defined overall design rules:

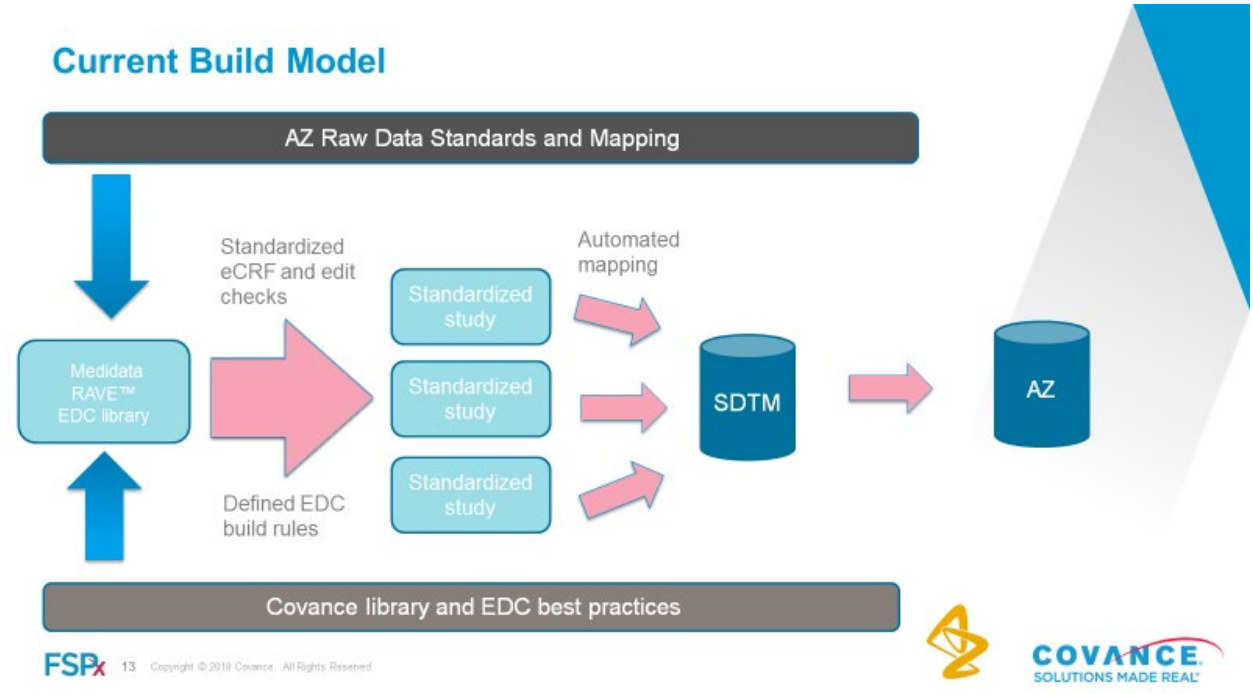
- Focus on mandatory/Highly Recommended fields on mandatory forms, optional (Recommended/Conditional) fields were considered if they were likely to be required for ECD studies.
- Refer to CDASH guidelines as required as reference point to ensure consistency and compliance.
- Design the screen with the end user in mind, ensure the collection flows, the screens are not too cluttered. Covance general EDC design rules were applied too.
- Wherever possible include the CT codes in the codelists.

- The codelists contain terminology that we considered was relevant to ECD studies, with the option to expand as required.
- Where the two companies differed between implementation, such as normalized vs, denormalized structure, we decided to go with the design we felt best suited entry.
- Where a denormalized structure was chosen, test codes were used wherever possible as the variable name.
- The Covance specification templates were used, as these are designed for Medidata RAVE™ definitions.

Whilst the AZ RAW data standards form an important part, they are no longer directly part of the study setup process. Along with the SDTM standards, the RAW standards are used as a point of reference, during eCRF development. The library does not include TA standards and for those we have gone back to AZ TA standards and reused where we were able to. Once incorporated in a study, study builders can refer to those and request approval to include in any new study. So we have tried to maintain the consistency across studies beyond our library forms.

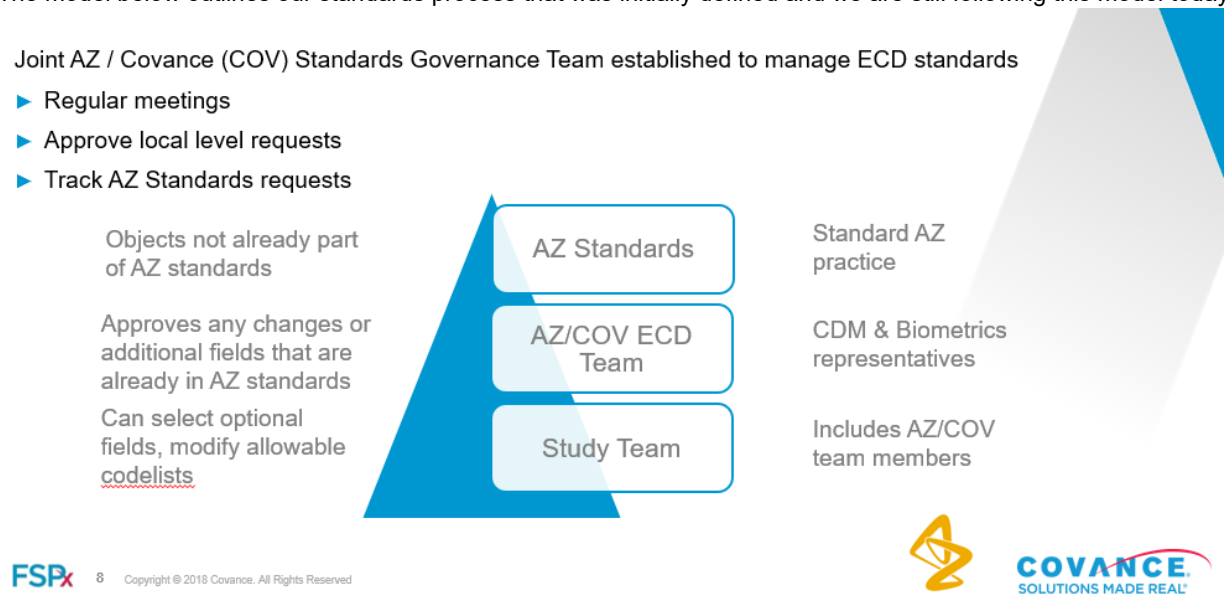
EDC SETUP

The diagram below outlines our initial build model and this is continuing today. Our standard EDC build process has a Kick Start Lead (KSL) role, who is responsible for creating specifications related to Rave EDC, including the eCRF specification, visit / matrix definition, edit check and derivation specification. We developed a standard set of eCRF specifications, which matches the EDC library. The KSL tailors the initial specification to match the protocol. The EDC library is used as the copy source for all study builds, the allowable modifications are clearly defined and controlled in both the specification and the library, the designer and study team are required to remain within the bounds of the standards. Any new requirements or exceptions need discussion and approval.



STANDARDS GOVERNANCE

The model below outlines our standards process that was initially defined and we are still following this model today.



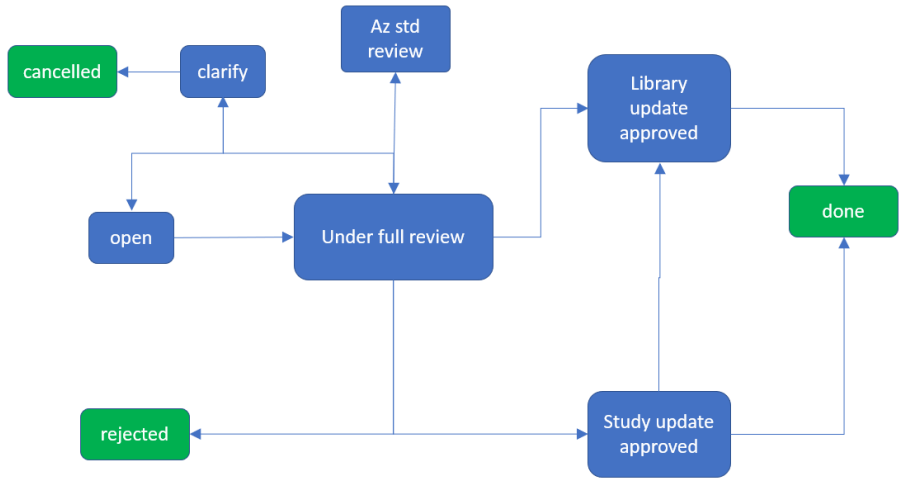
This is an important cornerstone to our model and compliance to standards is key to our success. It is important the exceptions / modifications to the standards at the study level are approved by a central standards team. After our initial wave of studies, it became clear that this was quite a task. In early eCRF development a standards representative attended study team review meeting, however this was not possible for amendments and activities post go live.

We decided to develop an online exception request system. The system would be open to both AZ and Covance team members to request exceptions to standards. We also included an option to ask questions, with the idea that in the future this could form an online query tool to find our 'frequently asked questions'.

JIRA™ TRACKING SYSTEM

JIRA™ is a cloud based workflow management system that can be customized to your needs. Our requirements are straightforward and are covered by the base package JIRA Core™. We determined the turnaround time on tickets for a first stage review is 24 hours. The range of complexity in requests we have seen so far is very broad. For easy requests that have minimal impact downstream, they can be approved relatively quickly. However, tickets that need further detailed investigation may require discussion with AZ partnership leads, including touch points within the AZ standards organization. Below is an outline of the workflow that was developed.

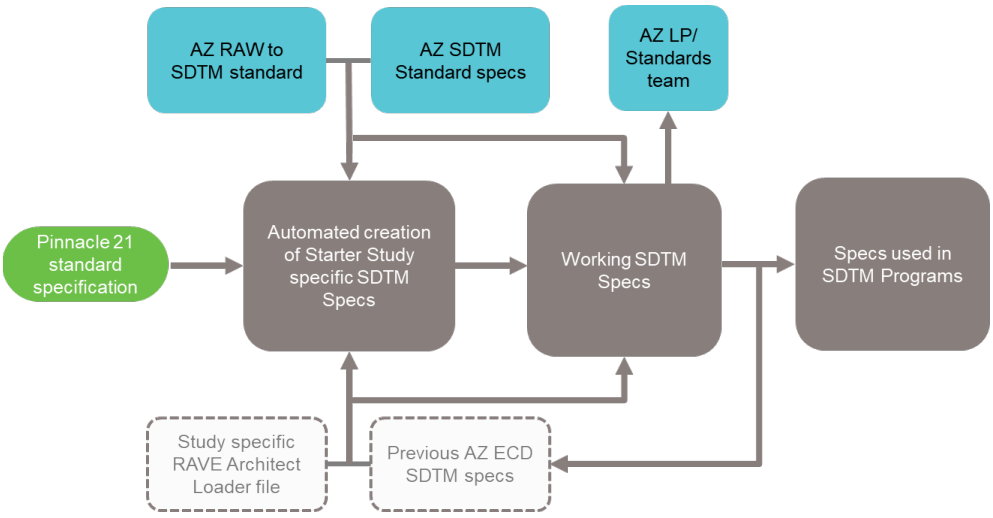
SAMPLE JIRA WORKFLOW



Generally, requests are entered by the KSL, but other roles do have access, there is a link between the request number back into the eCRF specification. A request is submitted at the highest level, for instance a new form is likely to include new fields, codelists etc. The customized workflow allows for approval of the request at the study level and as a library update. The library tasks are moved to a separate track and go through another round of approval. The advantage of tracking our library updates means that anyone can check those updates that have been approved and are pending a library release. The KSL can make these modifications in their studies without further approval. Below are the numbers of requests to date:

		First Response				
		Issues		Time (days)		
Issue Type	# Issues	n	(%)	Median	Min	Max
Form Level Update	46	43	93	1	0	3
New Edit Check	27	27	100	1	0	4
New Field on existing form	52	52	100	0.5	0	14
New Form	26	26	100	0.5	0	4
New derivation	3	3	100	1	1	2
Update field on existing form	47	45	96	0	0	3
Update to Codelist	27	25	93	1	0	1
Update to Edit Checks	23	23	100	1	0	5
Update to custom function	3	1	33	0	0	0

RAVE EDC TO SDTM



The AZ standards library has an excel file which provides their mapping from RAW to SDTM. However, as described above, AZ raw data standards are no longer directly used for our RAVE as this has been tailored for the relationship needs. We therefore had to setup a system which utilized AZ standards input as a starting point, but also incorporated all the mapping that the SDTM creation provided. The above process flow has been created to optimize the balance between automated mapping, manual mapping, utilizing the lessons learned from other studies and most importantly user friendliness.

The general idea is that the Study Architect Loader file is read into SAS, which is then held against the AZ SDTM standards documents and prior SDTM specs that were created for other studies. The system attempts to merge all RAW variables to known SDTM variables, and outputs a starter / blueprint specification that the SDTM programmer can use as basis for further updates. The system will provide the historical mappings from prior studies for variables that successfully merged and highlight the RAVE variables that have not been automatically merged. The SDTM programmer is expected to complete these specifications and to use the AZ SDTM standards documentation as source of input for further decision making. If there any needs to add new domains/variables, the Covance lead programmer will make a suggestion to AZ and the conclusion of that discussion is implemented.

The simplicity in this system comes from it utilizing the SDTM specifications from prior studies to retrieve new mapping information. This allows our SDTM programmers work comfortably within the pinnacle 21 specifications that they are familiar with. They make all required updates within that, and the system simply reads those back-in and mines them for new data. For each new study, we are reading in the latest specifications of all other studies, thereby ensuring we have the latest and most up to date mapping information.

It should be noted here that we do not want to claim that this is a machine learning process. It is simply a tool that utilizes the ability of computer systems to mine and compare data and is leaving interpretation and finetuning to the SDTM programmers. Although this is of course a solid platform to introduce machine learning as time goes on and our available data increases.

CURRENT STATUS

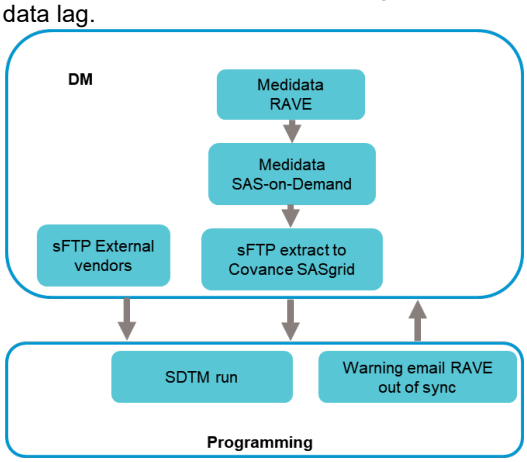
So far, 12 studies have been built using this model. We have successfully met our build timelines in all cases. There is a broad range of studies covered from complex oncology platform studies to a simpler observational study to support phase 2 recruitment.

CHALLENGES FACED

DATA FLOW INTO TO SDTM

As with most standardization efforts, the end goal is a reduction in workload and with that enabling short timelines between study start and first deliveries. Within the ECD relationship, the goal is to start daily SDTM deliveries within 6 to 8 weeks of FSI. And of course, with daily SDTM transfers, it is of importance to pay attention to time of data entry and availability of that data on AZ systems via SDTM.

Here was our first technical challenge. There are several systems that must be in sync with each other to minimize data lag.



We have RAVE SAS-on-Demand, a sFTP program to retrieve that data and push it to SASgrid and then finally an automated SAS process that runs all SDTMs. Steps had been taken to synchronize these systems in terms of careful planning and timing, but occasionally desynchronized due to the winter/summertime switch or IT issues. The easiest solution here was to create a SAS program which checks the synchronizations between the systems. E.g. it will report if the SAS-on-Demand more than a day before the sFTP picked up the data and also when the extraction from the sFTP is more than a day ago. Findings are sent directly to the group responsible for extracts and the Lead programmers. The end result is synchronization issues are immediately known and this cut down our resolution time significantly.

DAILY SDTM RUNS AND ONGOING DATA

Everybody working in our area will be sensitive to the fact that unqueried data can have its imperfections. And even more so, Programmers know how this can impact their ability to create SDTM. If left unsolved, it could stop the daily SDTM transfers to the sponsor and even become a cause of friction between DM and Programmers.

Our solution aims to increase of the co-operation between both groups by setting up a system that allowed programmers to create data-problem reports. When programmers create their SDTM programs, they can encounter data issues. Rather than gathering this information and sending it via a human made email, they can create a check dataset and output that to a permanently stored csv:

A	B	C	D	E	F
PROBLEM_text	PROBLEM_TYPE	SOURCE_TYPE	SOURCE_FILE	SUBJECT	FOLDERNAME
External LAB data but no sample in RAVE	MAJOR	EXTERNAL	LAB	xxxxx	Visit 2

In the above report a programmer noticed that external data contained more information than RAVE. They created code that would push such cases into a dataset. This dataset is then pushed into a macro that:

1. Determines whether the dataset contains issues that need to be reported
2. If such cases are within the dataset it:
    a. Pushes one warning to the SAS program log. This ensures that the SAS log will not be clean if data issues are present
    b. It creates a CSV file showing the data + user description of the issue

All programmers are instructed to try and implement such check reports in areas where we could get duplicates or difficult to interpret mismatches. Once all SDTM programs have been run, and several CSVs with data issue reports have been created there is a macro that takes all those CSVs and creates an XLSX file of that. This report is sent along with the daily SDTM transfer for reference and it is sent to Covance DM via email every Monday.

The benefit of these reports goes both ways. DM will have additional information at their disposal which helps the data-cleaning process, and programmers only have to think of a check once. As the quality of the RAW data improves, the data issue reports will eventually no longer find cases of interest, which means that the warnings disappear from the SAS log.

All groups involved have been uniformly positive about the process especially with the extra insight it provides on external data. Next steps will be implementing a feedback loop to include DM answers on old findings into the new reports.

Please note that we fully realize that some of these checks will overlap with edit-checks. However, these checks are not meant to replace edit-checks, and we do not want to get to a situation where programmers are unsure whether they should create a check. We are therefore on purpose generating some duplicate work, which through a study can be resolved.

RAVE LAB MODULE

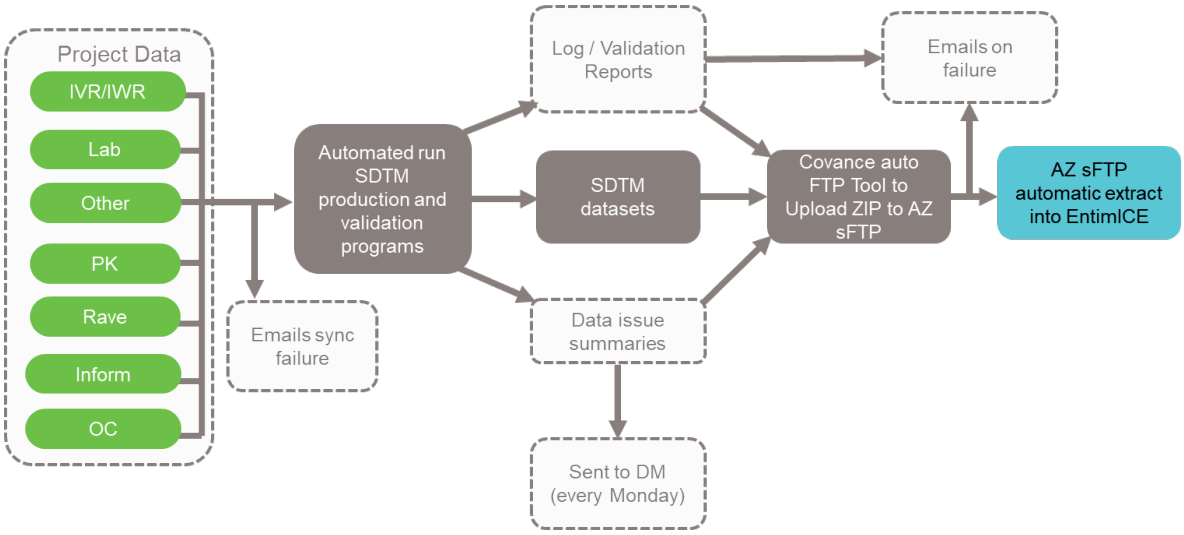
We all know how complex it can be to work with Local labs. The RAVE LAB module does remove a lot of the issues that you would normally encounter by providing a clear framework and metadata driven data-entry. But there it can still go wrong through perhaps insufficiency of the said metadata, or data entry issues. It can be quite laborious to work through the data and to support DM a SAS process was created to guide reviewers through several efficient steps.

1. Check whether the site have specified their lab name
    a. If sites have not entered this, the lab module will not be able to provide raw units to enter.
2. Per site, check how many results were entered without raw units
    a. Obviously, no raw units means that it can't be converted to standard units.
3. Per site/analyte, check how many results with raw units lack standard units
    a. This check will highlight whether the backend metadata contains all required conversions
4. Per analyte, check all standardized normal ranges against each other.
    a. We found examples of one analyte having a normal range of ~1-5 pmol/L for one record and 500-2500 pmol/L for others. This check flags normal range discrepancies like this to help finding incorrect raw unit assignments / incorrect conversions.
5. Per analyte, check the Coefficient of Variation of the raw unit summaries versus the Coefficient of variation of the standardized value
    a. The theory behind this check is that you would not expect large increases in the CV value when combining raw units into one standardized unit. If the CV value does increase significantly, you can assume that you have been combining different groups of data. We found examples of where the CV value increased more than tenfold and found this was due to sites entering the wrong raw unit. This would not have been picked up in an outlier analysis as the interquartile range was impacted by this as well.
6. Per analyte, an outlier analysis / top / bottom 5 for the more generic data checks

For EDC studies, DM has setup a specialized group of medics who are working with this report and are supporting the correction of the Lab module backend metadata and sending out the required queries. This is now being used across the entire group of studies.

DELIVERY OF DAILY SDTM

We already discussed many parts of our standardization of RAVE, creation of SDTM, synchronization checks and methods on dealing with data issues. This all needs to be tied together when we are creating our SDTMs for the daily delivery. Our initial deliveries were done with some manual steps and it was very soon clear that this could not be maintained. The following flow provides a general overview of the automated process that was introduced soon after:



First all raw data is gathered with synchronization issues being emails to stakeholders. After that SDTM is run, all data issues, log findings and mismatches with validation programs are gathered / summarized. The summaries are analyzed on whether it is ok to be sent. Currently this is a simple rule on that the production SDTMs are not allowed to have SAS log errors. After everything is approved to go, all raw data, SDTM datasets and quality reports are packaged into a zip file which is stored in a central location. This central location is checked by an automated FTP process which then sends it to the AZ sFTP. The automated FTP process triggers an error if there was a problem delivering the file to the AZ sFTP.

The process as a whole does not only provide the data, but also gives an open and honest assessment of what is being delivered. It has been a successful setup with 12 studies having performed deliveries totaling over 3000 zip files and within all those over 100.000 SDTM datasets sent.

WHAT THE FUTURE LOOKS LIKE?

Within the partnership, we are constantly looking for ways to improve our processes and looking to technologies that could support that aim. Having established a workflow and knowledgebase that allows us to successfully transfer SDTM datasets, our focus turns to any areas of automation that may better support any manual intervention.

XCELLERATE

Following Chiltern becoming part of the Covance organization, the data management team had the opportunity to share best practices with our new colleagues. Covance's tool **Xcellerate** consists of a number of end-user applications connected to a **clinical data repository** that supports near-real-time acquisition, mapping and integration of clinical trial data from any germane source. The solution leverages this clinical data repository to streamline and automate the process of delivering high-quality study data and associated metrics for downstream use.

It consists of the following four web-based applications:

(i) 'Discrepancy Manager', which facilitates review of programmed discrepancy check output and bulk handling of the resulting queries to sites and third-party data vendors;
(ii) 'Page Tracker', which facilitates reporting of missing EDC pages and the backlog of outstanding work for page review, verification and locking;
(iii) 'Query Tracker', which provides reporting and analytics for query management and root cause analysis;
(iv) 'Subject Tracker', which tracks subject data cleanliness and readiness for database lock and helps data management teams manage outstanding work.

We saw this as an opportunity to further automate some of our data flows and processes.

DISCREPANCY MANAGER

The management of the exception reporting is one area of focus. By loading these exceptions into the Discrepancy Manager, we can centrally manage and report on issues, these then feed into the Subject Tracker tool, as an overall subject status. Automating the loading of these exceptions, means resolved issues are no longer present and are auto-closed. Data Managers can send Rave queries, emails to vendors from the tool.

LOADING AND EXTRACTING DATA

Underlying Xcellerate is the clinical data repository. As our studies are being transitioned into this repository, it makes sense for us to investigate further potential areas for improvement. Clearly one of these is better consistent handling of multiple raw datasources. Xcellerate has functionality to manage the loading of external data, including error and exception handling for data from multiple sources. This can be further automated, so files can be delivered and loaded, minimizing any manual effort.

We have another work stream in progress, looking at documentation and processes with our Covance Labs colleagues, an expected outcome is to have the ability to establish nightly transfers of safety data. So we anticipate being able to treat lab data in a similar fashion to EDC data, removing the time lag associated with monthly transfers. Providing the ability to query in a similar timeframe to EDC data.

Our future plans also include looking at the integration points with our various technologies and databases. During 2020, we will investigate moving the extraction of the individual raw datasets from the clinical data repository.

CONCLUSION

Whilst standards form the basis of this model, what is of equal importance is the ability to be able to optimize and standardize processes. As a partnership, we have worked together to refine our goals and our processes, staying within the AZ standards, but having the flexibility to innovate. Two years on, the transfer of SDTM datasets is a stable and repeatable process. This stability enables our staff being able to focus on the bigger topics, rather than the day to day minutiae that can fill our days. As our studies move towards key milestones, we expect to see more efficiencies gained, whilst maintaining our high level of quality. Whilst there have been challenges to this project, overall the figures demonstrate that we have successfully met our original goals. Data is available nightly to the AZ organization, therefore allowing the study teams to have accessibility to the data at all times.

What have we learnt? This has been a positive and collaborative relationship and both sides have learnt a great deal. The initial plan was to have Covance team members sit alongside AZ study team members, this has been successful, is continuing to date and forms the backbone to the success of the partnership. Much of the success is due to this co-operation we ensure everyone works together towards a common goal. This not only has positive impacts on the quality of the delivery, but also in the engagement of the staff.

Back to the original AZ goal 'Get closer to the Data', by use of stable, optimized processes and technologies, we are making significant strides towards meeting this goal.

REFERENCES

1. **A cross-source, system-agnostic solution for clinical data review**

   Daniel Kowalski, Fang Du, Lalit Mohanty, Paul Konstant, Joseph Ciervo, Victor S Lobanov, Dimitris K Agrafiotis Database, Volume 2019, 2019, baz017, https://doi.org/10.1093/database/baz017

Published: 18 February 2019

CONTACT INFORMATION

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

Author Name:	Helen Scanlan	Gerard Hermus	Sue Huxtable
Company:	Astra Zeneca	Covance	Covance
Address:	Melbourn Science Park	Schipholweg, 111	Opsrey House
City / Postcode:	Melbourn, Hertfordshire,	2316 XC Leiden, The Netherlands	Maidenhead, Berkshire
Email:	Helen.blackburn@astrazeneca.com	Gerard.Hermus@covance.com	Susan.Huxtable@covance.com