PhUSE 2007

Paper RA08

All roads lead to SDTM

Which one shall we take?

Author – Ian Fisher, AstraZeneca, Alderley Park, UK
Co-author - Meena Rahman, AstraZeneca, Alderley Park, UK
Co-author - Donna Thompson, AstraZeneca, Alderley Park, UK

A shared vision in the Pharmaceutical industry is of a near future where all studies acquire and report clinical data in a globally standard format. The changing regulatory environment and the emergence of the Study Data Tabulation Model (SDTM) and Analysis Dataset Model (ADaM) from CDISC have provided companies with both the incentive and tools to make this vision a reality.

AstraZeneca (AZ) has embraced these new standards and the long-term goal is for all studies to acquire data in a SDTM compliant format, potentially taking advantage of a set of global data collection fields developed via the Clinical Data Acquisition Standards Harmonization (CDASH) project. Reporting datasets will follow AdaM conventions and programmers will have access to a comprehensive library of standard programs and macros allowing automated production of datasets and output.

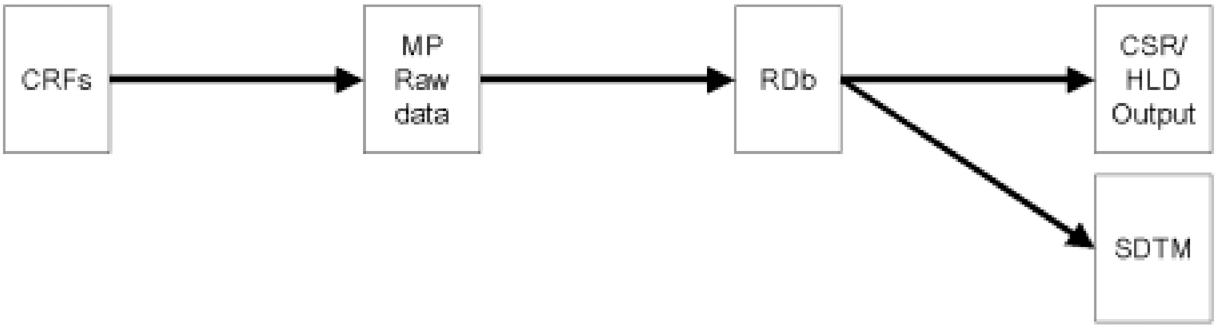
Realisation of this strategy would represent a significant development from 3 years ago, when data acquisition was via a large and flexible bank of corporate, therapeutic and project standards. Emphasis was more on standardisation of the structure of the data rather than standardisation of the contents of the data. Care was taken not to enforce too strict an adherence to the standards in the belief that flexibility in data acquisition and reporting allowed optimal delivery of each study, something which strict standardisation might compromise.

This paper will discuss the AZ approach and rationale to where, when and how SDTM was implemented. The journey so far has been difficult, fraught with many challenges, disagreements and ultimately some compromise, but a clear path has been forged and AZ is well on the way to realising their vision. The paper will map the road AZ have taken in their efforts to accommodate SDTM into their own standards, and offer some valuable insight, advice and guidance to other companies embarking on a similar journey.

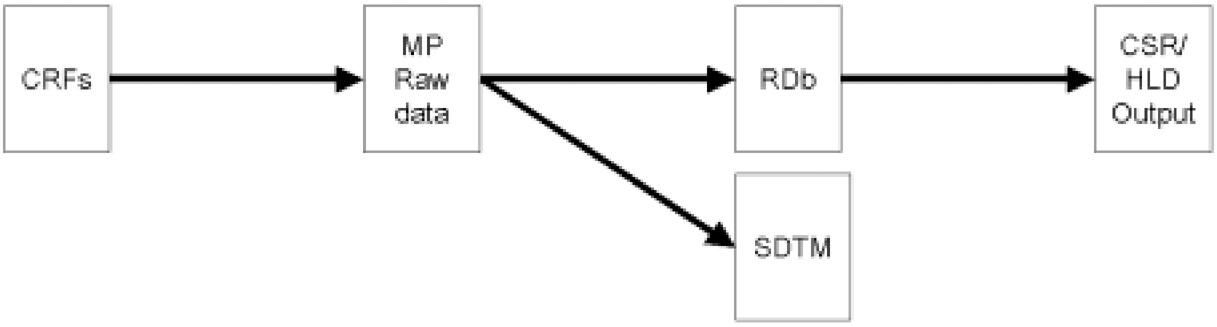
Careful consideration went into developing AZ's approach to SDTM implementation, but it is acknowledged that other paths could have been chosen. In this section, the paper will discuss where those paths may have led:

Would a complete redesign of the CRF package have been a more efficient approach?

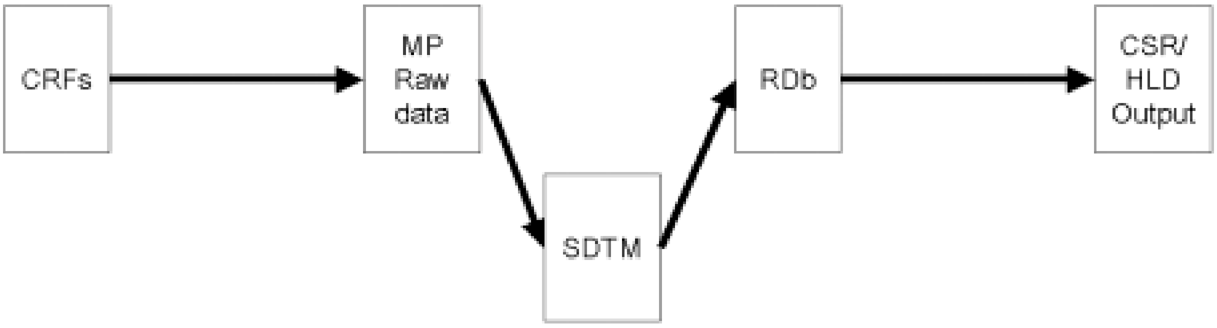
The CDISC Implementation Guide is a vital companion for understanding SDTM, but it does not specify how to transform the raw data collected from a clinical trial into SDTM compliant datasets. There are many correct methods for accomplishing this, each with their own benefits and drawbacks. The most elegant solution would be a complete redesign of the raw module package (MP) to ensure all data was collected in a SDTM compliant structure. The obvious benefits of this approach would be that SDTM datasets were available directly from the raw database, data management staff can communicate easily with end-users regarding any data issues as everyone would be looking at the same structures of datasets and programmers can concentrate on analysis and reporting. The significant disadvantages of this approach though are the high cost of redesign, the disruption for ongoing drug projects and the disturbance of the existing database management systems, which could lead to a loss of data quality. Also, as SDTM implementation is relatively uncharted territory, it is likely the strategies, maps and models will evolve over time and to ensure this does not impact on data management and data quality, AstraZeneca decided to take advantage of the flexibility of SAS® and implement SDTM post-data collection and considered the following options:



Create SDTM datasets post-reporting database (RDb) creation. This would have the benefit of no impact on data collection, RDb or Clinical Study Report (CSR) or Higher Level Document (HLD) production and SDTM creation need only happen at submission time. There would be, however, no direct link between the outputs and the SDTM datasets, possibly leading to review difficulties.



A parallel approach where SDTM datasets were created at the same time as the RDb would have the same pros and cons as the post-RDb approach but may result in duplication of derived variables.



Creating SDTM datasets pre-RDb would ensure the RDb was completely described in terms of the SDTM source, which would be desirable for the reviewer and no final SDTM construction would be required. Some changes to the existing RDb specification would be required, though. AstraZeneca chose this pre-RDb approach to SDTM implementation, as it was believed to offer the least disruption to existing processes and would be the most congruent with the systems and tools already in place.

Has the method used for mapping CRF data to SDTM format been the most effective?

A precise map, detailing how to transform every collected raw MP variable into a SDTM compliant structure, is necessary for AstraZeneca's approach to succeed. This map has been built, first at a corporate level for those variables common across all therapeutic areas (TA) and, further to that, at a TA level. This map takes the form of a spreadsheet with 26 columns and nearly 1900 rows, just at the corporate level, with up to an additional 2000 rows depending on the TA. A 26 column, 3900 row spreadsheet is a complex document, which has taken many people many hours to produce and, just as importantly, maintain and update. A representative example is shown here:

SDTM Sort Variable	SDTM Class	SDTM Domain	SDTM Variable	SDTM Label	SDTM Type	SDTM CT / Format	SDTM Origin	SDTM Role	SDTM Core	AZ Origin of SDTM Variable	MP Module	MP Variable	MP Label
101	Special	DM	STUDYID	Study Identifier	Char		CRF	Identifier	Req	MP	DEM	STUDY	Study Code

The mapping process has been complicated by several factors. To ensure quality, a definite set of rules is required, as is a final and absolute standard to compare against. With SDTM, neither of these has been available. The team responsible for the initial stage of mapping the corporate MP variables found that early versions of the SDTM implementation guide were often ambiguous, resulting in several correct ways to define a mapping. Whereas these core raw modules were well understood and familiar to everyone party to the mapping process, the TA mapping often required the team to consult at length with subject experts to fully understand what was being collected prior to mapping to SDTM. Especially in the early stages of the SDTM map development, these issues led to greater time than expected being spent on deciding how best to map certain modules; often those modules most key to a TA. Also, until very recently, there has not been a yardstick against which to measure the degree of compliance to SDTM for a submission. At least one such submission checking service is now on the market and tools like this will be invaluable for those seeking registration for the first time using SDTM datasets.

The evolutionary nature of SDTM has also proved challenging when building the map. As with most learning processes, improvements have usually been incremental and reactionary. For example, early SDTM Implementation Guides have been seceded by more specific and less flexible versions providing a clearer set of rules for teams to use when mapping variables and modules. A driver for the increased specificity of these guides is the feedback and questions from various implementation teams. A more stringent set of guidelines could lead to either validation of a previous decision based on ambiguous rules, or, as has often been the case, could lead to a full remapping of modules and variables. Within AstraZeneca, knowledge and understanding of SDTM principles has matured greatly and resulted in an increased probability of high quality SDTM mapping. This has also led to the revisiting of previous decisions causing some disruptions for drug projects marked as ‘Early Adopters’ for SDTM implementation.

The construction of the map from raw MP modules and variables to SDTM has been accepted by end-users in AstraZeneca as the most effective method for SDTM implementation. However, lessons have been learned which could have resulted in optimisation of the map creation process.

Which are the correct projects and studies to pilot with the new standards?

Two projects have been identified as ‘Early Adopters’. One small multi-drug project and one large oncology product developing for launch.

A suitable drug project to pilot SDTM implementation would not be on the critical path when starting the implementation, but would have a high profile in the business. A project with several studies with a quick turnover allows prompt illumination of benefits and difficulties and also affords the ability for improvements to be made in time for the next set of studies. An example of such a project at AstraZeneca is a signal searching pilot project, which conducts mainly phase II efficacy studies with small patient populations. This project implemented SDTM for 7 Phase II studies with patient numbers ranging from 64-96 across two oncology compounds.

The large oncology product, conducting its final two phase I trials and having recently launched a large phase III program was chosen as the drug project to adopt AstraZeneca’s SDTM implementation strategy. The two phase I trials were marked as ‘Early Adopters’ and would also be the first to implement tools and systems developed as part of a wider analysis and reporting transformation aimed at globalising information science in AstraZeneca.

With the corporate map created, the TA mappings added and the project specific mapping integrated, the study programmer has an initial task of ensuring every variable collected is mapped and that a study specific SDTM mapping document accurately describes the transformations intended. Paring down the project standard SDTM to a study map proved a straightforward task and programming the SDTM datasets under the direction of the map has proved very successful on these two phase I studies.

Having chosen two relatively similar pilot studies, the benefits of standardisation were evident. The vast majority of SDTM, RDb and reporting programs were directly transferable from one study to another. Also, given their standard structure, SDTM datasets which had one raw data source in one study, but multiple raw data sources in the other, were easily programmable and proved to be much more suitable for data review. An example was the EX domain, with one study using the study drug alone and the other study using the study drug in combination with several chemotherapy regimens. For review of the data, composite datasets like EX provide an excellent picture of patient level data.

However, implementing SDTM programming on two reasonably similar studies also helped to highlight some significant issues when they arose on one or both studies. Firstly and perhaps most importantly, SDTM structured datasets are designed for storage and not for reporting. The lack of any formats in SDTM follows through into RDb datasets with recreation of formats and value lists required for the correct reporting of data coming at the reporting stage. Also, composite SDTM datasets such as LB, which contains blood and serum laboratory tests results alongside urinalysis results, have proved very complex to treat with a standard approach. Additionally, as SDTM datasets contain no numeric variables, the combination of essentially numeric (i.e. character versions of numeric results) and textual results (e.g. ‘positive’, ‘negative’, ’trace’) has occasionally proved frustrating when deriving report-required variables for one class of result, which are not applicable for another result class. Changes from baseline would be an example where this has caused issues. A further example of the non-report readiness of SDTM structured dataset is the DS domain. A compilation of several single line per subject raw data sources, reporting from DS requires extensive restructuring and the creation of a new derived single line per subject reporting dataset. Such issues are confirmation of the expected results of inserting the SDTM dataset creation process before the RDb dataset creation process.

On the whole, selecting this oncology product as an ‘early adopter’ has proved the correct decision. Lessons can be learned on a mature project most easily. To learn lessons about SDTM implementation, one needs a stable platform of raw data standards to work from and a reasonably stable set of outputs to work towards. One can therefore validate assumptions one has made regarding the mapping of variables to SDTM without the worry of changes to source data and also one can evaluate the continuing validity of the templated standard outputs. SDTM is essentially outside the sphere of influence for programmers, but outputs can be changed, with acceptance from internal reviewers. The two phase I studies have afforded the opportunity to put in place the vast majority of programs prior to implementation on the much larger phase III studies. This has bred confidence in the programmers and the project team in the implementation process and that any effect on a critical package of studies will be negligible.

Aside from the SDTM mapping, what other activities have been required?

Formal CDISC training on the fundamental principles of SDTM was provided to all programming personnel involved in SDTM implementation. A comprehensive understanding of SDTM in terms of findings, interventions, supplementary data etc. was vital for those formulating the mapping from raw data structures to SDTM, less so for programmers solely working with the map. What have been important for programmers to understand are the trial design specifications. These will be individual to each study and will require updating for every new type of study.

To account properly for the derivation of the large number of standard variables in SDTM e.g. baseline flags, changes from baseline and days from reference start and end dates, a suite of macros supporting SDTM was required and, therefore training was also required. The macro library has been one of the more variable parts of the SDTM. A bank of non-study non-project aligned programmers developed and maintained these macros, which has been important in terms of support provided to study programmers, who might not have sufficient time to create such a library of macros, but also a non-project aligned team ensured complete consistency in approach and adherence to SDTM guideline which is paramount to the aims of such a macro library.

Complete regeneration of the RDb specifications has also been necessary. Required for submission, the accuracy of this document is crucial and the variability of these specifications during the development of the SDTM map has been a challenging experience for those maintaining the specifications at the global, project and study level.

Training in the use and purpose of the SDTM macros and the new SDTM-influenced RDb specifications was necessary for all programmers. With training at the point of need being a priority, programmers for the ‘early adopter’ studies were trained first with phased training sessions as needed for other programmers depending on individual project intentions on when SDTM implementation would occur.

What lessons have been learned so far?

The decision to create SDTM datasets programmatically from the raw datasets and prior to RDb creation was correct for AstraZeneca. In the years preceding SDTM implementation significant improvements in the stability and standard nature of the raw data standards has provided a good platform on which to construct a map. Furthermore, the implementation team recognised in advance that processes would be iterative and this approach would cause least disruption to project and study deliverables, in comparison to the available alternative implementation strategies.

In terms of data storage, maintaining consistencies and enabling compilation of data domains, SDTM is truly an elegant solution to many issues encountered by clinical programmers. Annual IND update reports, IDMC reviews and pooling of data for submissions will be able to be supported by much simpler programming methods and in much shorter periods of time. So although the true aim of SDTM is to aid the regulatory review process, many in-house programming tasks will undoubtedly benefit too.

In terms of data reporting, the picture is slightly more opaque. A RDb based on SDTM structured datasets is not immediately reporting friendly and will require back engineering of previously collected variables which were not mapped to SDTM, reprogramming formats which cannot be stored in SDTM, construction of new derived datasets and more data manipulation in reporting programs than had occurred previously. Furthermore, pre-SDTM RDb datasets were designed to support the SAS® outputs. These outputs were the result of years of experience, advice and views gained from countless studies and input from statisticians, physicians and other medical scientists. With the implementation of SDTM the drivers for RDb structure become more ambiguous.

Perhaps rather than stating that SDTM is not report ready, one could take the view that the approach to reporting might need to be rethought. SDTM is most likely here to stay and when AdaM arrives the essential SDTM structure will be retained. What AstraZeneca is already finding is that the drive towards standardisation of data storage and its subsequent impact on reporting places the study programmer in a challenging position. Adherence to external and internal programming standards and ensuring in-house output reviewer satisfaction is often incongruous. Therefore, internal customers who are not fully aware of the constraints of SDTM sometimes see programmers as less than flexible.

Another lesson learned is that just as a house is only as stable as its foundations, a SDTM map is only as stable as the raw data standards. Any change in structure of raw data collection, deviation from standard raw data naming conventions not only affects the map but can also have an effect on the use of SDTM macros. A reactive and knowledgeable support group involved in raw data standard maintenance is a must as is an analogous maintenance and support team for the SDTM map. Communication between these teams and study and project programmers should be often and timely. Ideally, any escalation process or proper direction of queries should be clearly defined in a communication plan.

As has been previously stated, the AstraZeneca global SDTM map and each subsequent project SDTM map evolved rather than simply appeared. This presented a moving target for programmers and the only thing worse than aiming at a moving target, is not to know it has moved. Therefore, once again, the importance of elevation of issues found by programmers to governance teams and the dissemination of decisions made by governance teams to programmers cannot be overstated.

In addition, it was found that 'early adopter' studies had their programming timelines effectively governed by the implementation team rather than the study teams. The availability of the SDTM map and the response to queries dictates the pace of work and the confidence of meeting deadlines.

Implementation of a SDTM map is a steep learning curve for all involved and if AstraZeneca were to repeat the process, more time would be budgeted for. Decisions made early in the implementation process were often revisited, not because they were incorrect but usually because knowledge of SDTM and how it could best be applied to AstraZeneca's benefit had grown. Efficiencies are found and implementation and governance teams should not be afraid to change their minds, provided proper communication of such decisions is ensured.

Working towards externally mandated standards has resulted in a greater incentive towards overall standardisation across all areas of programming within AstraZeneca. The realisation of having to produce datasets across all studies with exactly the same format has raised the level of overall applicability of SAS® programs and macros produced. Again, this presented a significant learning curve and required significant programming resource for the first 'early adopter' studies, but benefits have already been seen. Programming times for the second 'early adopter' study were 50% of those for the first and it is estimated that as steep as the learning curve was, the drop in subsequent programming timings will be just as sharp as programmers can work with 'off the shelf' solutions from standard codes libraries.

What recommendations would AZ have for others planning the same journey?

Although SDTM data structures are standard, the decision as to where and how to implement them into the programming process is not. Although for AstraZeneca, SDTM dataset creation post-data collection and pre-RDb was the most suitable approach, which may not be the case for all companies. AstraZeneca had a stable set of raw data standards, which were seen as a good foundation upon which SDTM datasets could be built. Additionally, reporting of study data was more variable across projects and studies and would therefore be less affected as a degree of flexibility in reporting programming existed. Furthermore, an analysis and reporting transformation had been planned to coincide with the implementation of SDTM, meaning that certain synergies could be planned for. Were a company to have extremely variable raw data standards but a very stable and standardised reporting procedure, perhaps post-RDb creation of SDTM datasets would be more suitable. An exhaustive conversation concerning the risk benefit ratios for each implementation scenario should take place.

If creating SDTM datasets pre-RDb, carefully evaluate the appropriateness of the intended output. For example, summaries of patient disposition, which include counts of reasons for withdrawal or discontinuation of randomised treatment, may have an associated value list e.g.

1 = Incorrect Enrolment/ Eligibility Criteria not Fulfilled
2 = Adverse Event
3 = Condition Under Investigation Worsened
4 = Voluntary Discontinuation by Subject
5 = Condition Under Investigation Improved / Subject Recovered
98 = Other

To correctly order such variables, it could be beneficial to store code-decode pairs or reprogramming of formats will be required. The effects of no numeric variables or formats in SDTM should not be overlooked when considering the potential overhaul required for existing reporting processes.

Assuming a mapping document specifying the route from raw (or RDb) datasets to SDTM datasets is created, an accompanying change tracking document must exist and all changes and outstanding queries relevant to the validity and appropriateness of the mapping contained therein should be logged. Also, the tool used to create the SDTM map should be carefully considered. There is a propensity to immediately choose Microsoft Excel for such a task, but given the evolutionary nature of the mapping process and the importance of communication of changes, tools such as Microsoft Word or Adobe Acrobat, with in-built tracking tools could be a more suitable choice.

In addition to the requisite education plans for programmers and implementation personnel in the fundamentals of SDTM, it is also important to communicate the imminence of SDTM and relevant implementation activities to other skills groups who, though not directly involved in the process, will be affected by the resource demands, availability of data and consequences of the when, where and how implementation will occur.

Throughout the whole process of SDTM implementation, a myriad of decisions need to be taken. The aim of the process is ultimately to achieve consistency, which can only be achieved by strong and effective leadership of the implementation teams. The clearer and more widely communicated the messages are from the governance and implementation teams, the more confidence programmers have in the stability of their work and their willingness to commit to deadlines. This confidence in turn will spread to the study teams who will then view the SDTM implementation process for its benefits rather than perceiving it to be a drag on timelines.

Is the original vision of full and complete standardisation of data acquisition, manipulation and reporting still seen as achievable and what is the timeframe?

The only answer is yes. Although in its early stages, SDTM has been implemented at AstraZeneca and the benefits are already being seen. There is a growing feeling that this is only the first step on a longer journey. At AstraZeneca, the decision to implement SDTM post data collection and pre-RDb has created a centre point of standardisation in clinical information science processes. If our submitted raw data is standard and this will save us and regulators time and effort, surely standardised collection such as that potentially available via the implementation of CDASH would result even greater gains? Furthermore, if our RDb were built on a standardised structure surely a similar and analogous process, perhaps AdaM, would be the logical next step?

In January 2006, CDISC stated that by 2010 or earlier, there would be a single CDISC standard for the full life cycle of a clinical trial or study from protocol representation through the capture of source data to analysis, submission and archiving. Assuming this aim is met, the driver for the industry to achieve complete standardization will be the adoption of this standard by the regulatory authorities. 2010 would be a challenging target to meet, but encouraging experiences in SDTM implementation such as those being enjoyed at AstraZeneca and the lessons learned therein can only spur the industry forwards.

The achievability of the goal is unquestionable, but the paths taken could be many and their lengths variable. The greatest challenge to overcome when attempting complete standardisation will be the ever-changing terrain. Even the slightest change in one set of standards will have a knock on effect to every other set of standards in place. Communication, governance and maintenance will become paramount within clinical information science as will the availability of broadly skilled technical personnel. Flexibility and creativity of reporting must never be lost in the drive for standardisation though. Standardisation should not be seen as a requirement for fewer programmers, but as a method of freeing existing expertise to focus on the interesting, investigative and exciting reporting of data which could give companies the edge in understanding their data and prove vital in the increasingly competitive market.

Author Contact Details:

Ian Fisher
Senior Programmer
UK Programming
TE2 C/2
Parklands
Alderley Park
01625 232294
Ex 32294
mailto:ian.fisher@astrazeneca.com

References

1. Study Data Tabulation Model prepared by the CDISC Submission Data Standards Team, Version 1.1.
2. New WebSDM™ Hosted Service Offering Enables Customers to Validate Clinical Trial Data Prior to Submission – Phase Forward® Press Release.
3. Implementation of the CDISC SDTM at the Duke Clinical Research Institute – Jack Shostak, Duke Clinical Research Institute, Durham NC.
4. CDISC Roadmap Discussion Document – www.cdisc.org
5. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration.