PhUSE 2017

Paper DH03

Converting Legacy Data to "ADaM-like" Data for Making Efficient Regulatory Response

Bengt G Fältström, AstraZeneca, Gothenburg, Sweden

ABSTRACT

Legacy data continues to be an important aspect for all pharmaceutical companies and will certainly continue to be so in the future. It is often very challenging and time-consuming to interpret and make the data understandable and possible to use for further analysis. This paper discusses a way of working to respond to a question from the EMA PRAC division that concerned a number of studies conducted as far back as the 1980-s and 90-s. By defining well structured and specified anchor points – including uniform pooled raw data and "ADaM-like" analysis data – the work could be organized in independent parallel streams. When all legacy data was reviewed, applicable data extracted and validated it was then possible to produce the desired tables and listings with no delays and the response to EMA PRAC could be sent as required.

INTRODUCTION

WHAT IS EMA PRAC?

The PRAC (Pharmacovigilance Risk Assessment Committee) is a division within EMA (European Medicines Agency), which is responsible for assessing and monitoring the safety of human medicines. This means all aspects of risk management including

- the detection, assessment, minimization and communication of the risk of adverse reactions, while taking the therapeutic effect of the medicine into account
- design and evaluation of post-authorization safety studies
- pharmacovigilance audit

For further information about the PRAC you can read more on the EMA webpage. See References section at the end of the document.

THE QUESTION FROM PRAC

The question from PRAC asked for data on kidney functions measurements from all trials for a certain class of drugs that have measured eGFR (estimated Glomerular filtration rate) or Creatinine clearance or other measures of renal function at baseline and during follow-up.

An analysis should be made of the change in kidney function for each treatment group, comparing the medication class with any comparator which could be either placebo or an active comparative treatment. The kidney function measurement should be kept as a continuous variable and should not be converted to a categorical variable based on threshold values.

PURPOSE AND SCOPE OF THIS PAPER

The purpose of this paper will be to discuss the difficulties the Programming Team faced when providing answers to the question above, and how we managed to overcome them by organizing the team and work processes in the best way.

STARTING UP

SELECTION OF STUDIES

The following criteria were set up internally for the studies to be included in the investigation, to be able to properly answer each of the questions.

1. Studies should be blinded and randomized
2. Studies should have either a placebo or active comparator control group
3. Treatment period should be at least 4 weeks
4. Cross-over studies would not be included
5. The study protocol should prescribe collection of eGFR or serum-Creatinine at both baseline and at a time point after at least 4 weeks of treatment
6. Data and at least the study report should be available

THE EGFR CHALLENGE

Question 3 from the PRAC asked for an analysis of the change in kidney function (i.e. eGFR – estimated glomerular filtration rate) from baseline to end of treatment. The eGFR values could be either measured values, or calculated from the Creatinine values. It turned out that we did not have measured eGFR value for any patient in any study so all eGFR values had to be calculated from the collected Creatinine values.

The eGFR can be calculated from the Serum Creatinine values in different ways using slightly different formulas. The different methods each have their advantages and drawbacks, and are used in different situations and for different populations or patient groups. Two of the most common methods are the MDRD and the CKD-EPI formulas. Both of these use the Creatinine value together with Age, Gender and Race for the subject to calculate the eGFR. For both methods the unit of S-Creatinine should be in mg/dL and the calculated eGFR value will then be in mL/min per 1.73 m^2.

The MDDR formula:

$$\boldsymbol{eGFR = (186) \times (\text{S-Creatinine}^{-1.154}) \times (\text{Age}^{-0.203}) \times (1.210 \text{ if Black}) \times (0.742 \text{ if Female})}$$

The CKD-EPI formula:

$$\boldsymbol{eGFR = (141) \times \min(\text{S-Creatinine}/k \text{ or } 1)^{a} \times \max(\text{S-Creatinine}/k \text{ or } 1)^{-1.209} \times 0.993^{Age} \times (1.018 \text{ if Female}) \times (1.159 \text{ if Black})}$$

- k = 0.7 for females and k = 0.9 for males
- a = -0.329 for females and a = 0.411 for males
- For min and max, choose either S-Creatinine/k or 1 whichever fulfills the criteria

The MDRD formula (Modification of Diet in Renal Disease) is the most popular among clinicians, even though that the newer CKD-EPI formula (Chronic Kidney Disease Epidemiology Collaboration) is considered to be the most accurate, especially for higher eGFR values. But the CKD-EPI formula performs less well in certain sub-populations, e.g. black women, elderly and obese, which could explain that MDRD is still the preferred formula.

MORE CHALLENGES

There were a number of additional challenges that the programming team could foresee already from start. Based on the criteria for study selection we ended up in retrieving raw data for 38 studies in total. The vast majority of these studies had been conducted quite a long time ago, many of them from the previous century and the oldest one as far back as 1983. Next section will elaborate more on the specific problems with legacy data.

In addition it was decided that each analysis would be made for both eGFR methods described above, and it would also be done separately for short-term studies (up to 6 months of treatment) and long-term studies (6 months or more of treatment). We also wanted to present baseline, end-of-treatment and change from baseline in separate tables. In total we ended up in 24 different analyses and tables to create.

The time to have final tables ready to be interpreted by physicians and statistician and writing of response letter was not more than 4 weeks. We could really feel the challenge.

LEGACY DATA REVISITED

DEFINITIONS AND PROBLEMS

Legacy data can be defined as either old data or data stored in an obsolete format. Often both of these apply. In the best of worlds we can let the legacy data stay safely where it is, but sometimes we are required to make new use of it. The need can come from our own organization where we want to make new use of some data. It can also be an external request where Academia or a Regulatory Authority wants to have access to the data or to have an analysis performed.

It is often associated with difficulties to a greater or lesser extent to understand the data adequately in order to make further use of it. In many cases the format of the data has become obsolete and changed during time, and some sort of mapping procedure is needed to make it possible to use further together with data from other studies and from later times. It might not always be possible to have a complete 1:1 mapping which will add further to the complexity, and make some interpretation decisions necessary.

Some data might not have any formats at all associated to it, perhaps just numbers, 1, 2, 3 etc. which were understandable at the time for the programmers and statisticians involved, but not now, perhaps many years afterwards. Variables might be named just V1, V2, V3 etc. with no label associated to them, which again probably was not a problem for the persons involved at the time, but for us nowadays trying to understand will cause great problems.

Units are another area which could cause problems, especially for lab data. There can be a large diversity of units over time and over different regions.

To add to the complexity, legacy data might also be stored on platforms or systems not in regular use any longer, which will make it problematic even to retrieve it for further use. Furthermore there might be several instances of data available, and it may not be entirely clear which is the correct version to select.

WHAT WE CAN DO

There are of course a lot of things we can do. Ideally there are still some “veteran” programmers and statisticians around in the organization, who are familiar with the data we are trying to understand. Hopefully there is also documentation in place, such as annotated CRFs, individual patient CRFs, study protocols, analysis plans and study reports etc. With these to help it will in almost all cases be possible to solve the puzzle and to be able to interpret the data. One useful way could be to run frequency tables on the data and compare with the numbers in the study report.

IMPLEMENTING THE WORK

LET´S GET STARTED

The basic instinct when facing an extensive work task like this, and with limited time available, is probably to get started with the work right away. And so it was also for us… From Clinical Operations function we had received a list of all studies together with some documentation available, such as study protocols and study reports. So we started to look at the data and write SAS® programs to extract the needed data out of the original datasets. Without having given much thoughts how to proceed with the data we have gathered.

Very soon we realized that this was not a feasible way to continue. We encountered more or less every situation described in the legacy section above, and we estimated that only to select, extract and interpret all needed data from all studies would take up the entire allocated time. A new approach was needed.

LET´S SIT DOWN AND PLAN

So, we sat down together in the programming team to analyze the total scope of work and see what could be done. From the statistician and physician we had a well defined layout of the desired tables, but the raw data we had available was not possible to use directly to produce these tables. Not by any far means.

The possibility of first extracting the raw data from all studies and then produce the outputs was immediately rejected. First, the time did not allow this approach, and second, there were so many exceptions and special cases regarding the data points so taking care of these in the table programs was considered not a good idea. We needed to do some customization and standardization of the data before creating the table programs, and which also should make it possible for us to work in parallel instead of in sequence, and make each contributor much less dependent on other work. And thus make it possible to deliver in time.

THE ANCHOR POINTS

Since we have quite a lot of experience doing table programs using ADaM datasets, we decided that one of the defined anchor points should be one or more “ADaM-like” datasets for the tables to be based on. The other anchor point would be the pooled raw data with well defined structure, but where we kept calculations, selections and derivations to an absolute minimum. The reason for this was that this part was considered the most time critical.

The table programs should also be “clean” and not include any selections or derivations of data, only the statistical calculations. That meant that all data manipulations needed had to take place in the programs creating the ADaM datasets.

THE SPECIFICATIONS

TABLES

Below is the basic layout for the table showing the change of eGFR from baseline to end-of-study. Similar tables were also done for eGFR at baseline and end-of-study, for the two different products, two eGFR formulas and for short-term and long-term studies. In total that summed up to 24 individual tables.

	eGFR change from Baseline Medication class in scope					**eGFR change from Baseline** **Comparator**				
Study 1	N	Mean	Median	Stand. dev	Min/Max	N	Mean	Median	Stand. dev	Min/Max
Study 2	N	Mean	Median	Stand. dev	Min/Max	N	Mean	Median	Stand. dev	Min/Max
Study 3	N	Mean	Median	Stand. dev	Min/Max	N	Mean	Median	Stand. dev	Min/Max
…	…	…	…	…	…	…	…	…	…	…
All studies	N	Mean	Median	Stand. dev	Min/Max	N	Mean	Median	Stand. dev	Min/Max

CONSOLIDATED RAW DATA

The raw data was organized in three different datasets, one for demographics (DM), one for exposure and treatments (EX) and one for the serum-Creatinine samples (LB). As might be seen we tried to use SDTM terminology as much as possible. A few recent studies were already in that format so that was considered the most appropriate way. Limiting the raw data to only variables necessary to produce the output was definitely one key to success.

For EX and LB we did not do any selection of which records to include. This selection was made when deriving the ADaM datasets.

The consolidated raw data was derived in two separate processes, first at study level (DM_study, EX_study and LB_study) and then pooled for all studies together (DM_pooled, EX_pooled and LB_pooled).

DM – Demographics (subject level structure)

Variable name	Variable label	Origin/Derivation
STUDYID	Study Identifier	
SUBJID	Subject Identifier	
SEX	Sex	From “Demographic” CRF dataset
BRTHDAT	Date/Time of Birth	From “Demographic” CRF dataset
AGE	Age	If Age collected in “Demographic” CRF dataset, else empty
RACE	Race	From “Demographic” CRF dataset. If not collected then set as UNKNOWN
RACEOTH	Other Race Specification	Populated If Race=OTHER and collected, otherwise empty
RFSTDTC	Subject Reference Start Date	First treatment date. If not collected set as applicable visit date according to study protocol

EX – Exposure (multiline structure)

Variable name	Variable label	Origin/Derivation
STUDYID	Study Identifier	Match DM dataset
SUBJID	Subject Identifier	Match DM dataset
EXTRT	Name of Actual Treatment	Actual treatment exactly as it is entered in source dataset
EXDOSE	Dose per Administration	Populated if available in source dataset
EXDOSU	Dose Unit	Populated if available in source dataset
EXSTDTC	Start date of Treatment	
EXENDTC	End date of Treatment	Populated if available in source dataset

LB – Laboratory data (multiline structure)

Variable name	Variable label	Origin/Derivation
STUDYID	Study Identifier	Match DM dataset
SUBJID	Subject Identifier	Match DM dataset
LBTESTCD	Lab Test Short name	Actual Lab Code present in source dataset
LBTEST	Lab Test	Actual Test Name present in source dataset
LBDTC	Date of Specimen Collection	Date for lab sample collected from source dataset
LBORRES	Result in Original Unit	Result for lab sample collected from source dataset
LBORRESU	Original Unit	Unit for lab sample collected from source dataset

"ADAM-LIKE" ANALYSIS DATA

The analysis data was organized in two different datasets. One subject level ADSL dataset where all derived demographic data was stored together with treatment information including first and last date of treatment. The ADLB dataset was organized with Parameters, Analysis Visits and Analysis Flags appropriate for the table outputs.

As mentioned there were also some calculations, selections and derivations done when transferring the pooled raw data to analysis data. See list below for examples.

- Convert to standard unit (mg/dL) for serum-Creatinine if in other unit
- Select first and last treatment date
- Estimate last treatment date if unclear or not present according to specified rules (i.e. using study length)
- Calculate age
- Calculate end age to be used for studies with more than 12 months of treatment
- Derive race from different formats, and when unclear or missing according to specified rules
- Select baseline and end-of study Creatinine value according to specified rules
- Calculate eGFR from the Creatinine value and demographic data using the formulas for eGFR (MDDI) and eGFR (CKD-EPI)

The ADSL dataset

Variable Name	Variable Label	Code list	Derivation from Consolidated Raw Data
STUDYID	Study Identifier		
USUBJID	Unique Subject Identifier		Concatenated STUDYID/SUBJID
AGE	Age		If Age is populated in Raw data then ADSL.AGE = DM.AGE else ADSL.AGE is calculated using DM.RFSTDTC and DM.BRTHDAT
AGEU	Age Unit	Year	Set by the program

Variable Name	Variable Label	Code list	Derivation from Consolidated Raw Data
ENDAGE	Age(years) at last Creatinine Sample collected		IF STUDYTYP = LONG then ENDAGE calculated when last Creatinine sample was taken. When STUDYTYP = SHORT then ENDAGE set as empty
SEX	Sex	(SEX)	From DM dataset
RACE	Race	(RACE)	From DM dataset
RACEOTH	Other Race Specification		From DM dataset
RACEGR1	Pooled Race Group 1	Black Non-black	If DM.RACE in (“Black”, “African-American”, etc.) then set as Black, else set as Non-black. For some special cases or empty values follow derivation rules from Statistician.
RACEGR1N	Pooled Race Group 1 (N)	1 = Black 2 = Non-black	
TR01PG1	Planned Pooled Trt 1 for Period 1	Medication class Comparator	Derived from EX.EXTRT according to agreed derivation rules.
TR01PG1N	Planned Pooled Trt 1 for Period 1 (N)	1 = Medication class 2 = Comparator	
EXSTDTC	Start Date of Treatment		First observation from EX.EXSTDTC
EXENDTC	End Date of Treatment		Last observation from EX.EXENDTC or according to specified rules if not present
STUDYTYP	Type of Study	SHORT LONG	Will be coded based on information in Study Protocol/Report. Study length > 6 months will be set as LONG, otherwise SHORT.
STUDYLEN	Planned Study Length (Days)		Coded based on information from Study Protocol/Report, e.g. planned treatment = 4 weeks will give STUDYLEN = 28
RFSTDTC	Subject Reference Start Date		From DM.RFSTDTC

The ADLB dataset

Variable Name	Variable Label	Code list	Derivation
STUDYID	Study Identifier		Match ADSL
USUBJID	Unique Subject Identifier		Match ADSL
AVISIT	Analysis Visit	Baseline End of Study	
AVISITN	Analysis Visit (N)	1 = Baseline 2 = End of Study	
PARAM	Parameter	S-Creatinine eGFR (CKD-EPI) eGFR (MDRD)	

PARAMCD	Parameter Code	CREAT
		EGFR1
		EGFR2
			CREAT = S-Creatinine
			EGFR1 = eGFR (CKD-EPI)
			EGFR2 = eGFR (MDRD)

LBORRES	Result in Original Unit		For Parameter Code = CREAT set as LB.LBORRES
LBORRESU	Original Result Unit		For Parameter Code = CREAT set as LB.LBORRESU
LBSTRESN	Numeric Result in Standard Unit		For Parameter Code = CREAT calculate to unit mg/dL which is used in the formulas
LBSTRESU	Standardized Result Unit	mg/dL	

Variable Name	Variable Label	Code list	Derivation
AVAL	Analysis Value		For Parameter Code = CREAT then set as LBSTRESN For Parameter Code = EGFR1 or EGFR2 value is calculated according to formula
BASE	Baseline Value		Set as AVAL at Baseline
CHG	Change from baseline		CHG = AVAL-BASE
ABLFL	Baseline Record Flag		
ANL01FL	Analysis Flag 01		Analysis valid record
ADY	Analysis Relative Day		LB.LBDTC - RFSTDTC
ADT	Analysis Date		From LB.DTC

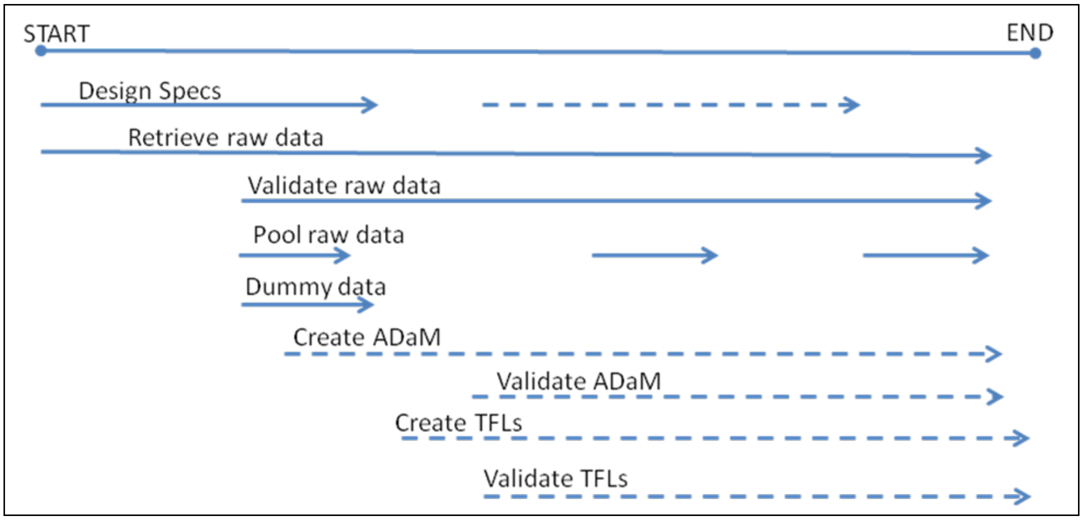
PARALLEL WORK STREAMS

The defined anchor points together with the well defined specification for each of the anchor points made it possible to work more or less in parallel and quite independently from each other. The team consisted of five persons where two got the main responsibility for the raw data including the validation; one person was responsible for creating the ADaM datasets and one for creating the table programs. These two team members also cross-validated each other's work. The fifth person was leading the project and was responsible for the cross-functional interactions, the specifications, documentation and for overall project oversight and management.

THE PARALLEL STREAMS

- Design the specifications
- Retrieve raw data
- Create draft pooled raw data to be used by ADaM programmer
- Validate raw data
- Pooling of all retrieved and validated raw data
- Create draft ADaM datasets to be used by table programmer
- Create ADaM datasets
- Validate ADaM datasets
- Create table outputs
- Validate table outputs

The below figure will give a hint how the parallel streams were distributed during the work period.



THE RESULT

THE RESPONSE

With combined efforts and dedicated work we managed to deliver the tables in time to Clinical Operations who was leading the overall cross-functional project team. The results could then be reviewed and interpreted by physician and statistician and the response document could be sent in time to EMA PRAC.

KEY LEARNINGS

A number of learnings could be made after working with this delivery. The first was of course the importance of having all legacy data stored and available for use in easily accessible environments. A few years ago AstraZeneca performed a project where a number of products including quite old ones were searched and locations for data and key documents were saved in a database. This preceding work was invaluable for the success of performing this task.

Another key issue was to really take the time to do a solid planning and preparation work in the beginning. Every hour spent on planning at an early stage can probably save days of work in the end. As part of the preparation was of course the production of well defined and understandable specifications of importance. This helped us focus on the data and variables that really mattered.

Keeping all the team members informed about how work progressed in the other streams was also important. We had regular meetings several times each week during the work where we shared information and progress and discussed issues to be solved. We also tried to ensure that every team member had a full understanding of all aspects of the project since all work streams were connected in the end. It happened several times that good solutions were suggested by team members who were not directly involved in a particular stream but who could see a way of solving a problem based on experiences in another stream.

It was also essential that we maintained a good cooperation with the other functions – Clinical Operations, Physicians, Regulatory and Statisticians. What we could have been a bit more proactive about was to have assurance from the other functions that all the studies that we had received as suitable for inclusion really were so. It turned out for some of the studies, after quite a lot of hours spent on programming, that they actually were not suitable for various reasons.

PROFITS IN TERMS OF TIME, RESOURCES AND QUALITY

From a management perspective it is of course interesting to speculate if there were any further gains to be cashed in apart from a timely delivery when organizing the work in parallel streams and using well known industry standards (CDISC) such as ADaM and to some extent SDTM. What about time, resources and quality? Can we have it all?

To start with the last one I would say that the quality of the delivery was improved by working in this way. Five persons is a perfect size for a manageable group and the total amount of knowledge, intelligence and creativity will of course be larger than what one or two persons could have contributed with. The regular discussions and follow-up work we had decreased the risk of making mistakes or taking any wrong decisions how to interpret our data material.

Coming to the time aspect it was of course necessary to organize the work to work in parallel in this or in another similar way. Otherwise it would not have been possible to have met the timelines set up.

Considering the probably most interesting aspect for the manager – Can we save resources? – Then it becomes a bit trickier to evaluate. The team of five persons completed this work in four weeks by using parallel streams, with other assignments allocated as well. Assuming that the team spent 80% of the working time dedicated to this work (=16 weeks) could then one person (or rather two since validation is needed) do all of this in strict sequence during four months? I am not sure. It would probably take a little longer time. We definitely had a great advantage working together in a group and helping each other out when getting stuck on particular issues. This probably saved quite a lot of time in the end. We also got energy from each other and became a really focused and efficient team. In addition, and as a bonus, we also had fun doing it.

The gain in total resources needed in terms of FTE (Full Time Employee) might be in the region of 20-25 % according to a very rough and unscientific estimation. Let´s say that instead of 5 FTE-months we now actually used up 4 FTE-months. From a manager´s perspective this is perhaps worth considering also for other types of deliverables not necessarily subject to these stressed timelines.

CONCLUSION

By organizing the work in independent parallel work streams and using well known industry standards like ADaM and SDTM to create well defined specifications as anchor points to connect the streams, it was possible to deliver the required analyses in time to be sent in to Regulatory Authorities (EMA PRAC) for a question involving a number of legacy studies. In addition we probably improved the quality of the response and we may also have reduced the total amount of resources needed for delivering the final analyses.

REFERENCES

Link to EMA (PRAC) webpage:
http://www.ema.europa.eu/ema/index.jsp?curl=pages/about_us/general/general_content_000537.jsp&mid=WC0b01ac058058cb18

ACKNOWLEDGMENTS

I would like to express my sincere thanks to the programming team (Jane Lu, Henrik Aronsson, Jesper Olsen and Zahra Sadeghi) for a marvelous group achievement, and to my manager Monika Malmros for constant support and mitigation efforts, not least during the process of writing this paper.

CONTACT INFORMATION

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

Bengt G Fältström
AstraZeneca
Gothenburg, Sweden
+46 70 967 00 63
bengt.faltstrom@astrazeneca.com