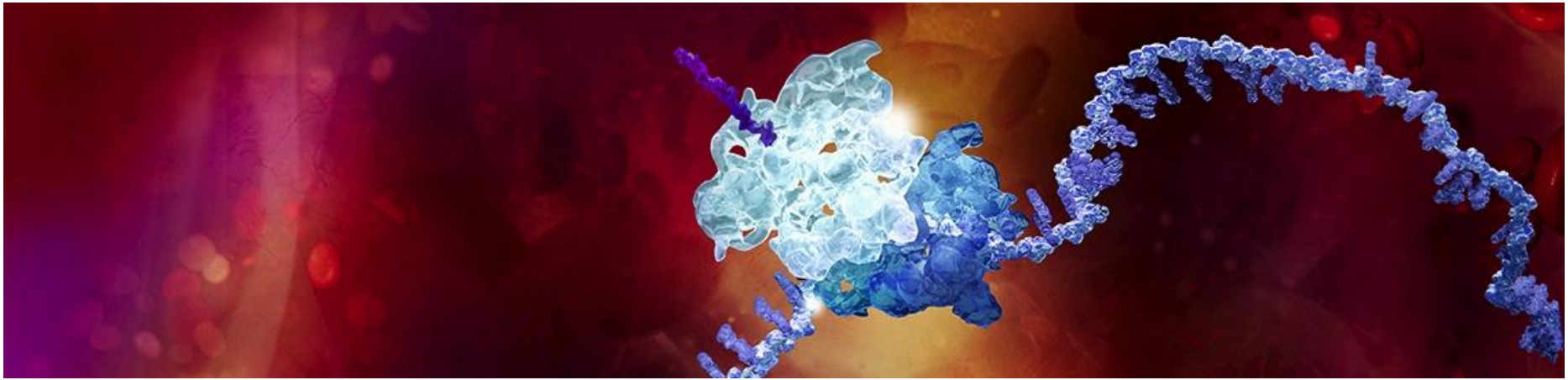
DH03 – Converting Legacy Data to "ADaM-like" Data for Making Efficient Regulatory Response

Bengt G Fältström



About the presentation

1. The Question from PRAC
2. The Preparations
3. Legacy data revisited
4. Starting the programming work
5. Anchor points, Specifications and Parallel work streams
6. The Result

The Question from PRAC

EMA / PRAC

- PRAC = **P**harmacovigilance **R**isk **A**ssessment **C**ommittee
- Division within EMA (European Medicines Agency)
- Responsible for assessing and monitoring the safety of human medicines

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

The Question from PRAC

- Data on kidney function measurements from all trials (of a certain class) that have measured eGFR* (or Creatinine clearance or other measures of renal function) at baseline and during follow-up.

- Please provide an **analysis of the change in kidney function for each treatment group**. Treat kidney function measurements as a continuous variable (do not convert them to a categorical variable based on threshold values).

(*eGFR = estimated Glomerular Filtration Rate*)

The Preparations

Forming a Cross-functional Core Team

- ✓ Clinical Operations
- ✓ Physician
- ✓ Regulatory

- ✓ Publishing
- ✓ Statistics
- ✓ **Programming**

- Representation from each involved function
- Lead by Clinical Operations
- Weekly meetings and ad-hoc meetings if necessary

Searching for the studies in scope

- ✓ Blinded and Randomized
- ✓ Placebo or Active control
- ✓ Treatment period at least 4 weeks
- ✓ Cross-over studies NOT included
- ✓ The protocol should prescribe collection of eGFR or Creatinine both at baseline and after at least 4 weeks of treatment
- ✓ Data and at least Study Report available

What we found

- 38 studies were found to be suitable
- Only a few were according to modern CDISC standard
- The majority conducted more than 10 years ago, many during previous millennium
- The oldest was run in 1983

Legacy data revisited

Legacy data – old data and/or in obsolete format

- Finding it
- Different formats, units etc.
- No formats
- Variables without labels
- Several instances of data
- Obsolete systems or platforms

Making Legacy data re-useable (“the fruit-rule”)

Find

Retrieve

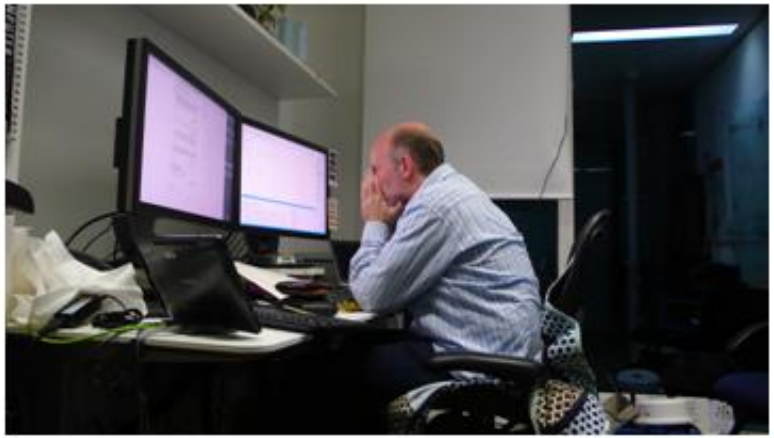
Understand

Incorporate

Translate

What can be done to understand?

- "Veteran" programmers might still be available
- Annotated CRFs or individual patient CRFs
- Study Protocols/Reports, Statistical Analysis Plans, other documents, data from other studies etc.



Starting the programming work

What data did we need to answer the question?

eGFR (CKD-EPI)* *(unit = mL/min per 1.73m^2)*

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

k = 0.7 for females, k = 0.9 for males
a = -0.329 for females, a = -0.411 for males
For min and max choose either Creatinine/k or 1 according to each criteria
Creatinine in unit mg/dL

** CKD-EPI (Chronic Kidney Disease – Epidemiology Collaboration)*

An alternative formula to calculate eGFR

eGFR (MDRD)* *(unit = mL/min per 1.73m^2)*

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

Creatinine in unit mg/dL

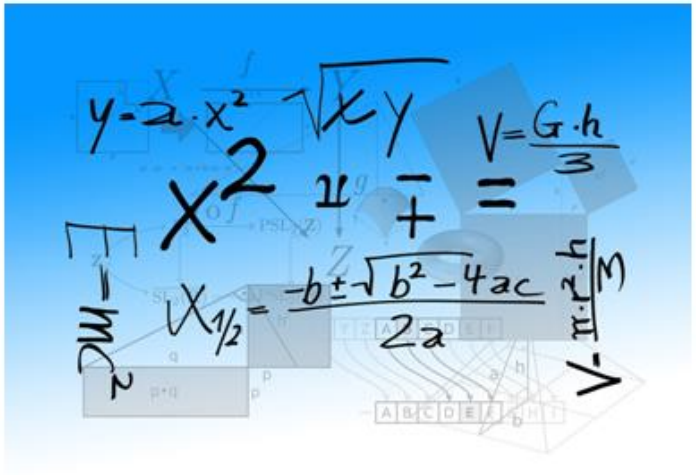
* *MDRD (Modification of Diet in Renal Disease)*

Which eGFR method should we use?

- MDRD is the preferred formula among Clinicians
- CKD-EPI is the more modern formula
- CKD-EPI considered more accurate, especially for higher eGFR values
- CKD-EPI performs less well for certain sub-populations;
  - black women
  - elderly
  - obese
- **Decision – We will provide tables and analyses for both eGFR formulas**

The data that we need for the eGFR calculations

- Demographic data
  - Gender
  - Age
  - Race
- Creatinine data
  - Baseline value
  - End of treatment value
- Supporting data
  - Date of Birth, Treatment start date, Treatment stop date



Can we go from original Raw Data directly to Analysis tables?

Original Raw Data from 38 studies

```
data a; set study1;
if .... ;
data b; set study2;
if .... ;
data tot; set a b ;

proc report ... ;
```

	eGFR change from Baseline Medication class in scope		eGFR change from Baseline Comparator	
S 1	N	Statistics	N	Statistics
S 2	N	Statistics	N	Statistics
S 3	N	Statistics	N	Statistics
…	…	…	…	…
All	N	Statistics	N	Statistics

Issues we encountered for our legacy raw data (I)

Missing format catalogue or dataset

- could be missing for all data in a study or for some data
- check with CRF, Protocol, Report or compare with other similar studies

Different formats for same variable

- RACE – changes during time, between regions
- LABCODE – changes during time, sometimes missing and only test name available which could also vary
- Randomization data – huge variation, very study-specific, sometimes just A, B etc
- check with CRF, Protocol, Report or compare with other similar studies

Issues we encountered for our legacy raw data (II)

Variable names with no labels
- understandable in most cases, but could be like V1, V2, V3 etc.
- check with CRF, Protocol, Report or compare with other similar studies

Different names for same topic
- especially date variables
- check with CRF, Protocol, Report or compare with other similar studies

Data located in different source modules/datasets
- e.g. randomization data
- detective work…

Issues we encountered for our legacy raw data (III)

Implied data

- e.g. last treatment day could be implied to be the same as last visit day
- check with CRF, Protocol, Report or compare with other similar studies

Different units for same lab data

- lot´s of variation, e.g. local studies
- need for Google search to find specific mapping to mg/dL for one locally used unit for Creatinine

These raw data issues could occur alone or in different combinations. Could be very time consuming to make data conformant

Question – use original raw data to do table outputs?

24 table outputs was requested by Physician and Statistician

- baseline, end-of-study, change, eGFR x 2, long-term studies, short-term studies, all studies

The time constraint

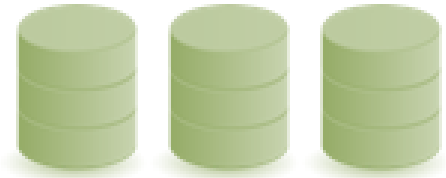
- only to locate, retrieve, select and validate all raw data would require all the allocated time

It became very clear that we needed different approach

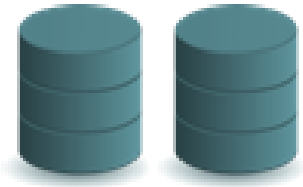
Anchor points, Specifications and Parallel work streams

Defining The Anchor Points

Consolidated Pooled Raw Data



“ADaM-like” Analysis Data



Going from Original Raw data to Analysis Table via defined Anchor points

Original raw data from 38 studies

Consolidated raw data for each study with standardized variables

Pooled raw data

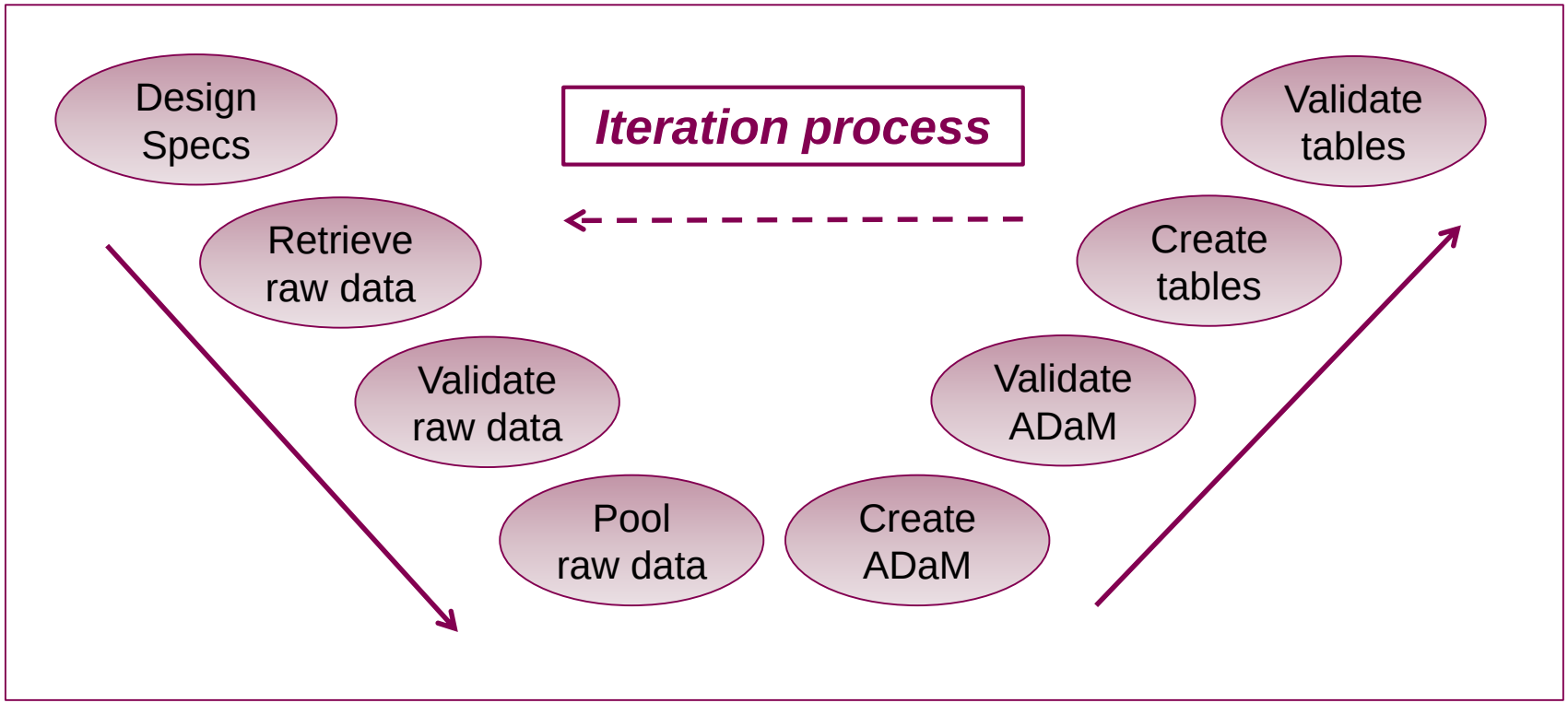
"ADaM-like" Analysis data

	eGFR change from Baseline Medication class in scope		eGFR change from Baseline Comparator	
S 1	N	Statistics	N	Statistics
S 2	N	Statistics	N	Statistics
S 3	N	Statistics	N	Statistics
…	…	…	…	…
All	N	Statistics	N	Statistics

The possibilities using Anchor points

- ✓ Possibility to organize the work in parallel streams, quite independently from each other
- ✓ Draft consolidated Raw data can be used for creating draft “ADaM” data
- ✓ Draft “ADaM” data can be used for creating draft tables
- ✓ Iteration process possible – concerns at later stages can be fed back to previous steps for re-design of the Specs

Data flow and parallel work streams



What to program where?

Pooled Raw data

- ✓ Finding the data
- ✓ Map to defined variables
- ✓ No derivations
- ✓ No exclusions (e.g. Treatment dates)

ADaM data

- ✓ Calculations
- ✓ Lab unit conversions
- ✓ Selection of data
- ✓ Derivations
- ✓ Make analysis ready

Consolidated Raw Data – (per study and pooled)

DM
One record per subject

STUDYID
SUBJID
SEX
BRTHDTC
AGE *(if collected)*
RACE
RACEOTH
RFSTDTC

EX
"Many" records per subject

STUDYID
SUBJID
EXTRT
EXDOSE
EXDOSU
EXSTDTC
EXENDTC

LB
"Many" records per subject

STUDYID
SUBJID
LBTEST
LBTESTCD
LBDTC
LBORRES
LBORRESU

The first "ADaM-like" analysis datasets

ADSL *(sub-set of variables)*
One record per subject

STUDYID
SUBJID / USUBJID
SEX
AGE
ENDAGE
RACEGR1
TR01PG1
EXSTDTC
EXENDTC

- **Some derivations:**

- ENDAGE – calculated for studies > 6 months.
- RACEGR1 – can have values Black or Non-black. Based on RACE in DM and specific rules set up by Statistician if unclear or ambiguous data.
- TR01PG1 – can have values "Medication class" or "Comparator"
- EXENDTC – last observation from EX or according to specified rules if missing or unclear

The second "ADaM-like" analysis datasets

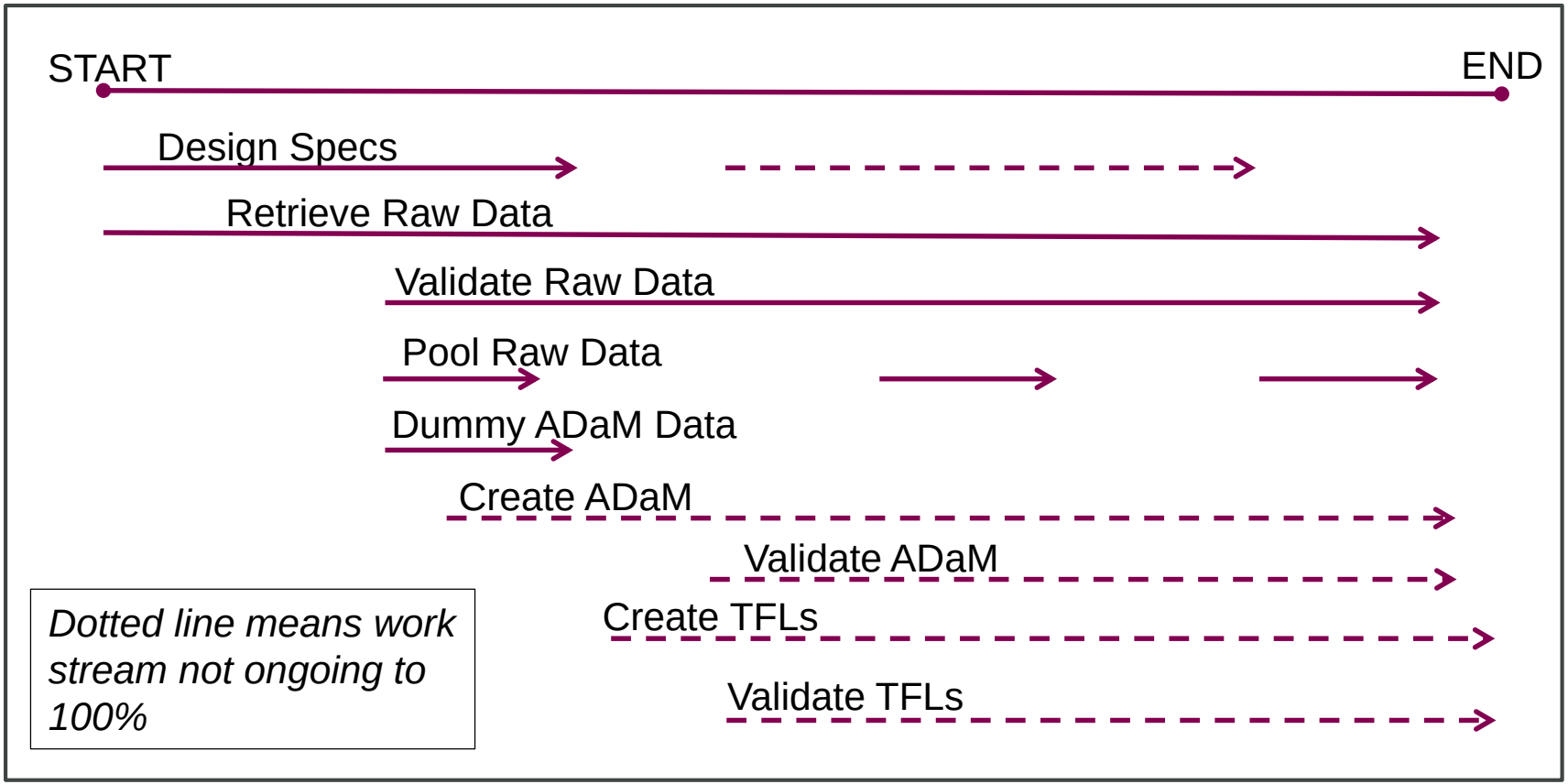
ADLB *(sub-set of variables)*
Several records per subject

STUDYID
SUBJID / USUBJID
AVISIT
PARAM, PARAMCD
LBSTRESN, LBSTRESU
AVAL, BASE, CHG
ANL01FL

- **Some derivations:**
- AVISIT– can have values "Baseline" and "End of Study".
- PARAM – can have values Creatinine, eGFR-MDRD and eGFR-CKD/EPI
- AVAL, BASE, CHG – calculated for both eGFR definitions
- ANL01FL – indicates which records that can be used for analysis

- *For complete set of variables and derivations see paper 2017 - DH03*

Work streams running in parallel



The Result

The delivery

- ✓ We delivered a set of analysis tables in time for review and interpretation by Statistician and Physician
- ✓ The response to EMA PRAC was delivered in time and was accepted

Our Key Learning's – Success Factors

- ✓ Have control of your Legacy Data.
- ✓ Planning and Preparations – Every hour spent on planning early on might save days of work in the end.
- ✓ The importance of well defined and understandable Specs.
- ✓ Team work – regular meetings and follow-up, encourage discussions and questions.
- ✓ Co-operate with other functions.

What about Time, Resources and Quality?

- ✓ TIME – Yes, parallel work is faster than sequential…
- ✓ QUALITY – Yes, a team of 4 – 5 persons is more knowledgeable, intelligent and creative than a team of 1 or 2. Discussions and follow-up reduced the risk of making mistakes. Regular iterations.
- ✓ RESOURCES – Yes, maybe or quite probably. Focus, Energy, Dedication, Possibility for instant help and resolution etc.
- ✓ BONUS – We had fun doing it ☺

Summary

- ✓ 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 our work and we may also have *reduced the total amount of resources* needed for delivering the final analyses.

Many thanks for your time and attention.

All questions welcome.

bengt.faltstrom@astrazeneca.com