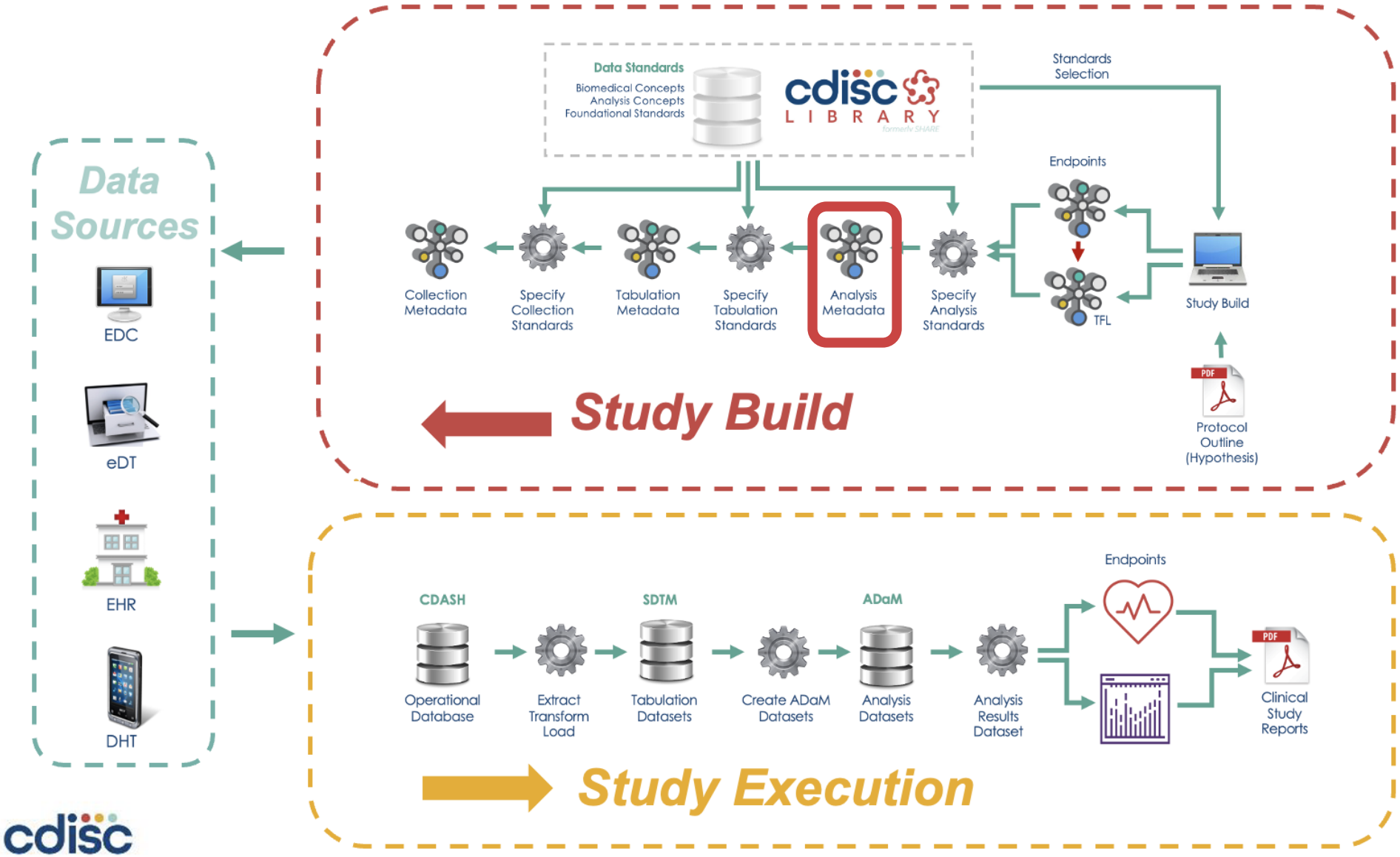
# Analysis Concepts
## – Initial perspectives from the CDISC working group

Bess LeRoy
Head of Standards Development, CDISC
05-September-2025

# Standardizing Analysis Metadata




cdisc

# CDISC 360

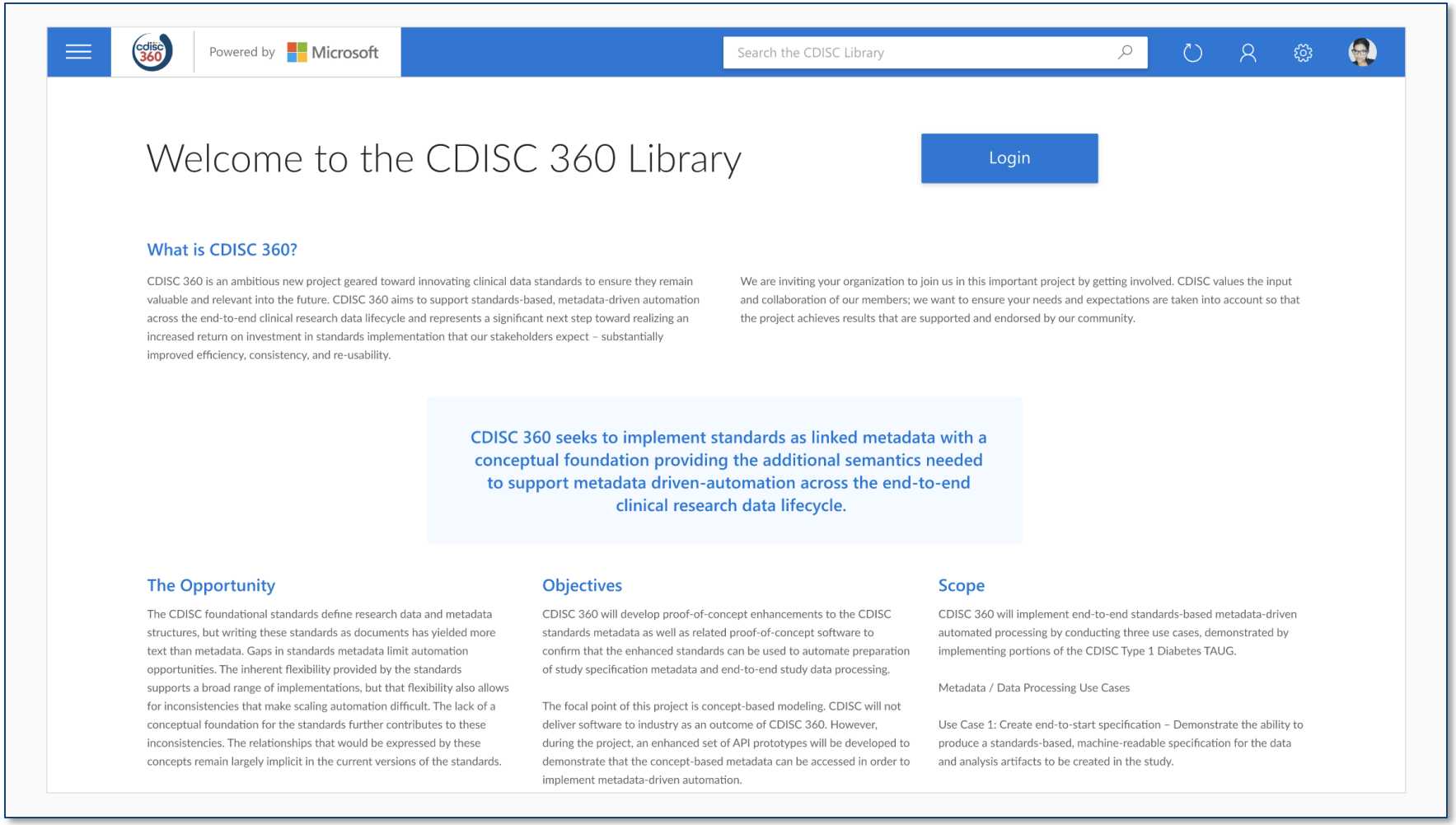
- CDISC 360 was a proof of concept that sought to implement standards as linked metadata with a conceptual foundation providing the additional semantics needed to support metadata-driven automation across the end-to-end clinical research data lifecycle.

- This will enable software developers to develop new tools (proprietary and open source) that consume this novel metadata to ease standards' implementations, while increasing data processing efficiencies.

- Reduce unnecessary variation and lower the barrier to adoption.

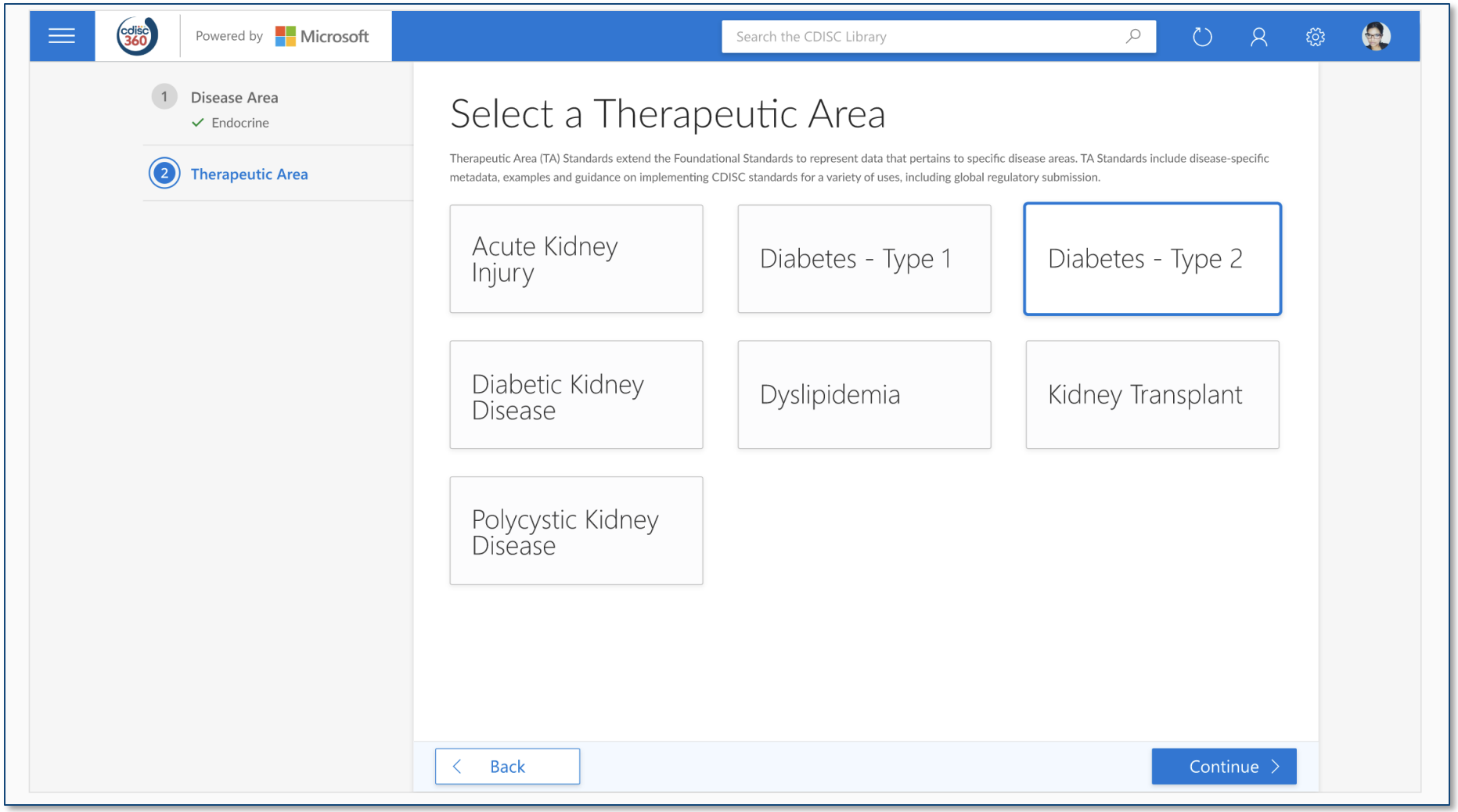
White Paper: https://www.cdisc.org/sites/default/files/2021-06/CDISC_360_Project_White_Paper.pdf

# CDISC 360: The Art of the Possible

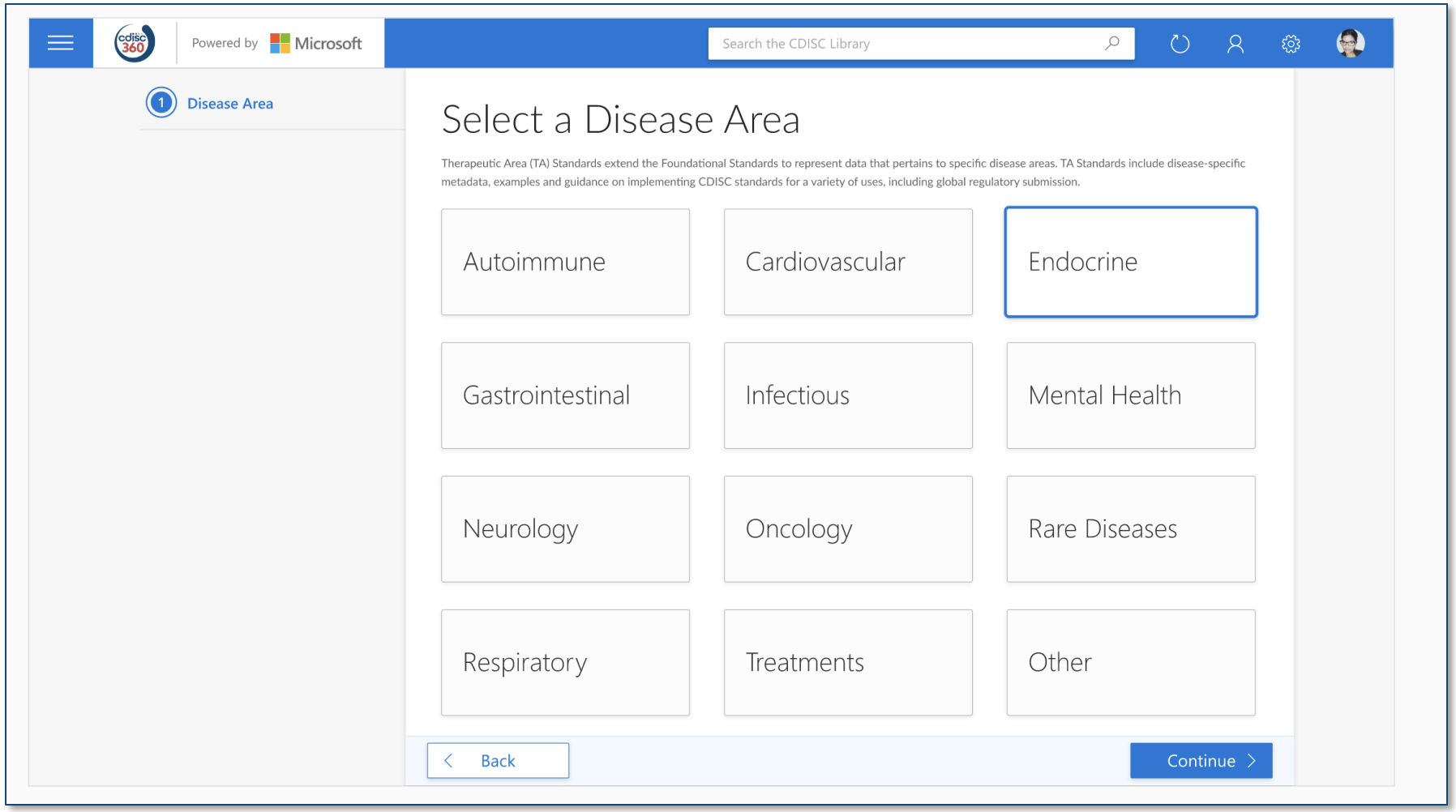



Reference: '*CDISC 360 - The Journey so Far and the Road Ahead', Peter Van Reusel, 28th April 2020*

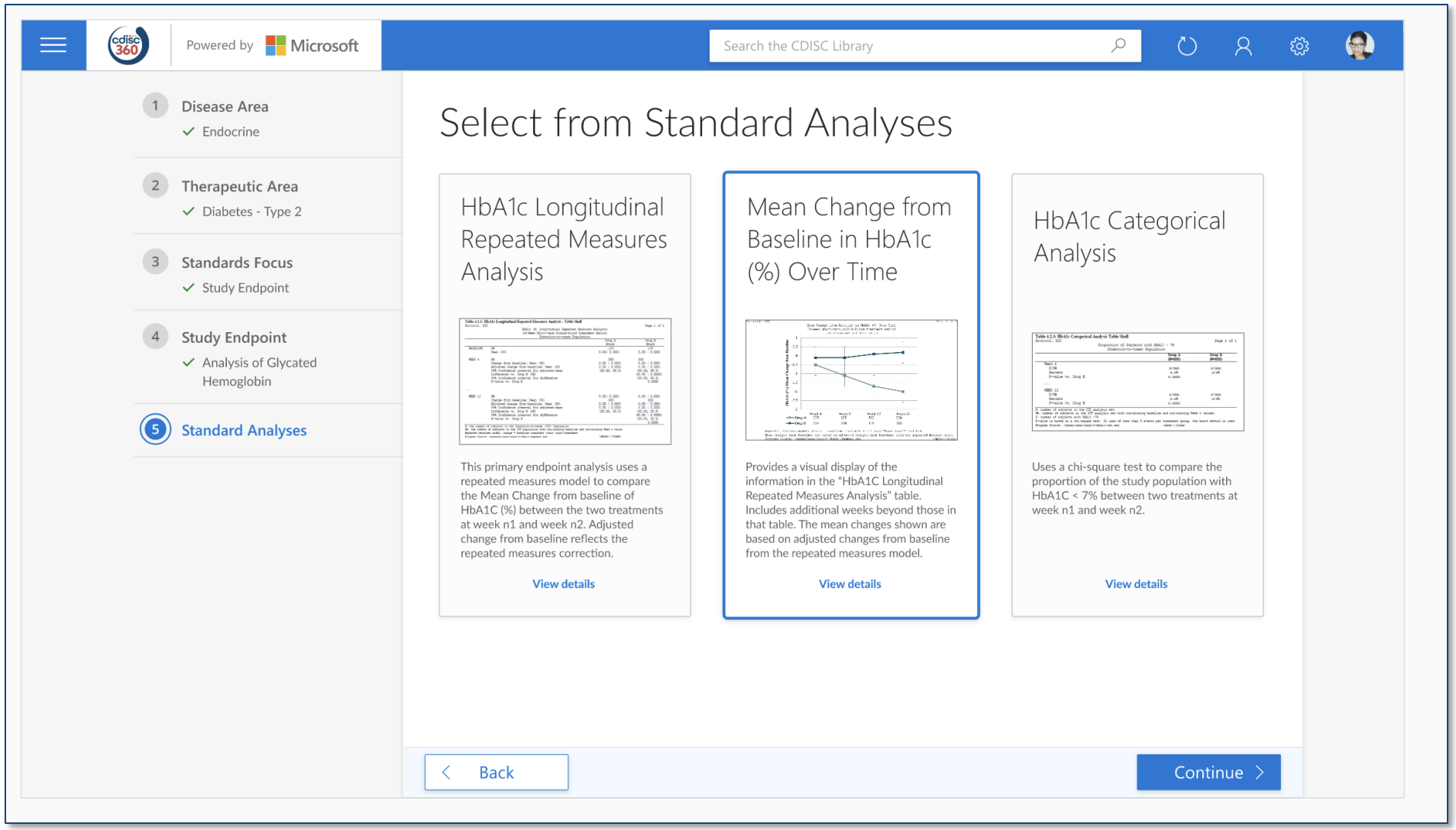
# CDISC 360: The Art of the Possible



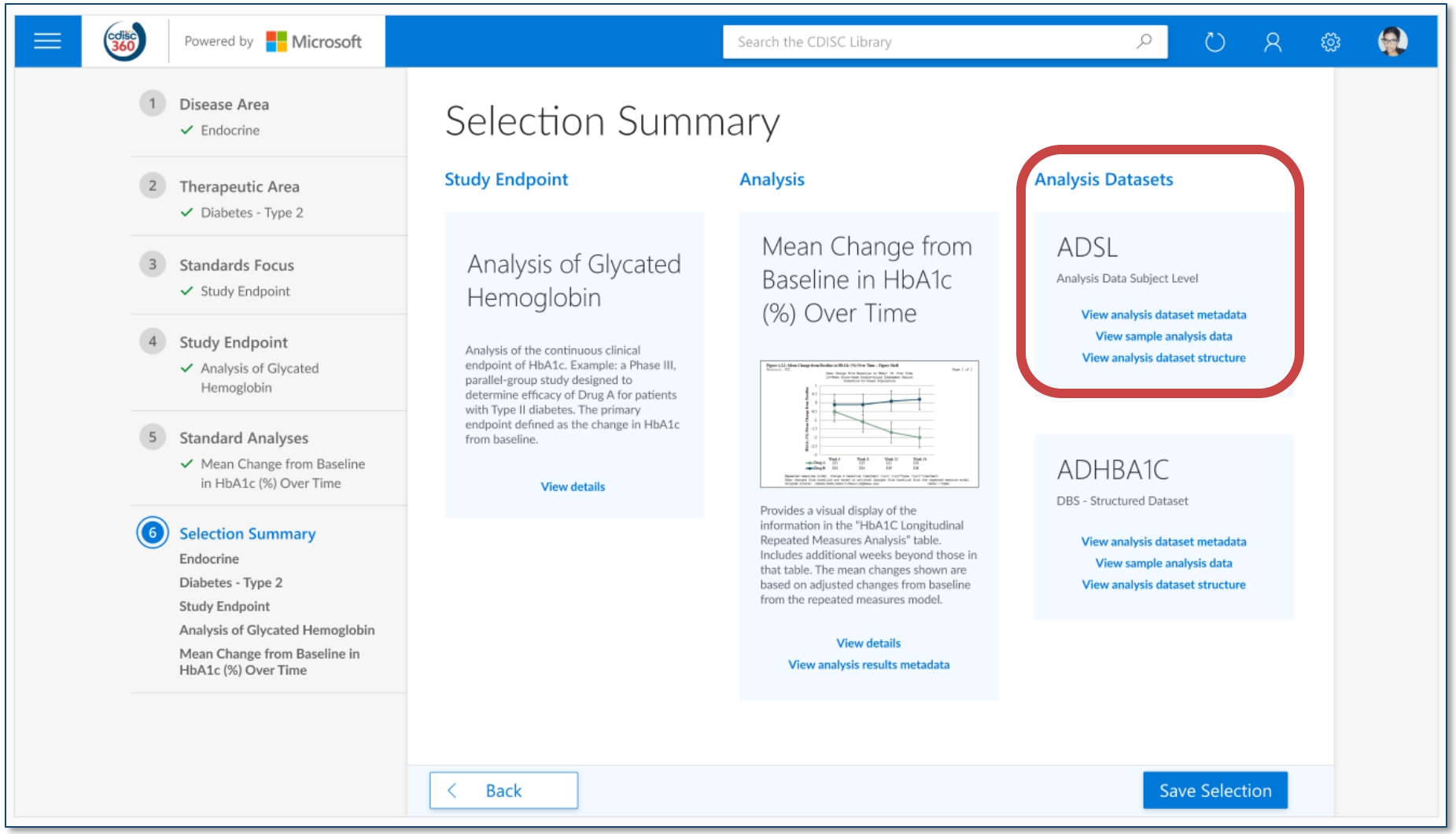
# CDISC 360: The Art of the Possible

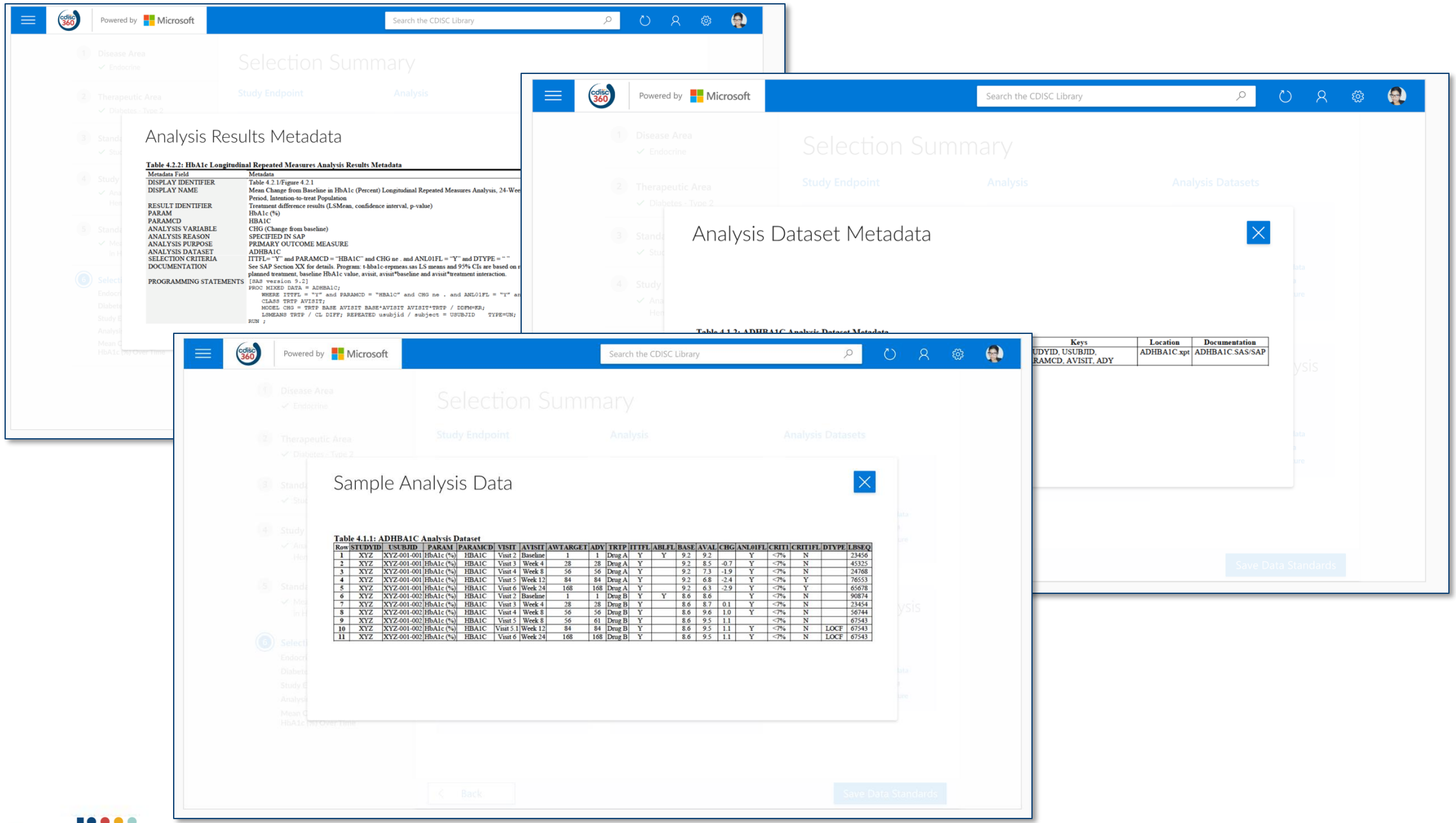


# CDISC 360: The Art of the Possible

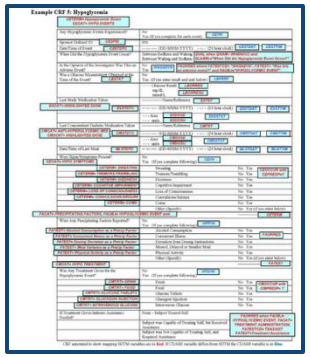


# CDISC 360: The Art of the Possible



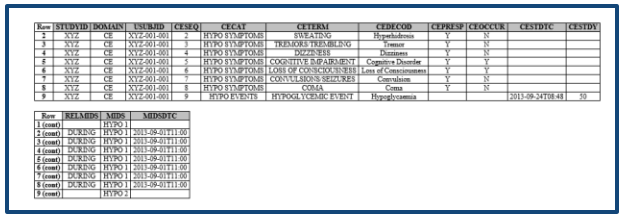

Powered by Microsoft
Search the CDISC Library
Selection Summary
Study Endpoint
Analysis
Analysis Datasets
Analysis Results Metadata
Table 4.2.2: HbA1c Longitudinal Repeated Measures Analysis Results Metadata
Metadata Field
Metadata
DISPLAY IDENTIFIER
Table 4.2.1/Figure 4.2.1
DISPLAY NAME
Mean Change from Baseline in HbA1c (Percent) Longitudinal Repeated Measures Analysis, 24-Wee
Period, Intention-to-treat Population
RESULT IDENTIFIER
Treatment difference results (LSMean, confidence interval, p-value)
PARAM
HbA1c (%)
PARAMCD
HBA1C
ANALYSIS VARIABLE
CHG (Change from baseline)
ANALYSIS REASON
SPECIFIED IN SAP
ANALYSIS PURPOSE
PRIMARY OUTCOME MEASURE
ANALYSIS DATASET
ADHBA1C
SELECTION CRITERIA
ITTFL= "Y" and PARAMCD = "HBA1C" and CHG ne . and ANL01FL = "Y" and DTYPE = " "
DOCUMENTATION
See SAP Section XX for details. Program: t-hba1c-repmeas.sas LS means and 95% CIs are based on r
planned treatment, baseline HbA1c value, avisit, avisit*baseline and avisit*treatment interaction.
PROGRAMMING STATEMENTS
[SAS version 9.2]
PROC MIXED DATA = ADHBA1C;
WHERE ITTFL = "Y" and PARAMCD = "HBA1C" and CHG ne . and ANL01FL = "Y" an
CLASS TRTP AVISIT;
MODEL CHG = TRTP BASE AVISIT BASE*AVISIT AVISIT*TRTP / DDFM=KR;
LSMEANS TRTP / CL DIFF; REPEATED usubjid / subject = USUBJID TYPE=UN;
RUN ;
Analysis Dataset Metadata
Keys
Location
Documentation
UDYID, USUBJID,
RAMCD, AVISIT, ADY
ADHBA1C.xpt
ADHBA1C.SAS/SAP
Sample Analysis Data
Table 4.1.1: ADHBA1C Analysis Dataset
Row STUDYID USUBJID PARAM PARAMCD VISIT AVISIT AWTARGET ADY TRTP ITTFL ABLFL BASE AVAL CHG ANL01FL CRIT1 CRIT1FL DTYPE LBSEQ
1 XYZ XYZ-001-001 HbA1c (%) HBA1C Visit 2 Baseline 1 1 Drug A Y Y 9.2 9.2 Y <7% N 23456
2 XYZ XYZ-001-001 HbA1c (%) HBA1C Visit 3 Week 4 28 28 Drug A Y 9.2 8.5 -0.7 Y <7% N 45325
3 XYZ XYZ-001-001 HbA1c (%) HBA1C Visit 4 Week 8 56 56 Drug A Y 9.2 7.3 -1.9 Y <7% N 24768
4 XYZ XYZ-001-001 HbA1c (%) HBA1C Visit 5 Week 12 84 84 Drug A Y 9.2 6.8 -2.4 Y <7% Y 76553
5 XYZ XYZ-001-001 HbA1c (%) HBA1C Visit 6 Week 24 168 168 Drug A Y 9.2 6.3 -2.9 Y <7% Y 65678
6 XYZ XYZ-001-002 HbA1c (%) HBA1C Visit 2 Baseline 1 1 Drug B Y Y 8.6 8.6 Y <7% N 90874
7 XYZ XYZ-001-002 HbA1c (%) HBA1C Visit 3 Week 4 28 28 Drug B Y 8.6 8.7 0.1 Y <7% N 23454
8 XYZ XYZ-001-002 HbA1c (%) HBA1C Visit 4 Week 8 56 56 Drug B Y 8.6 9.6 1.0 Y <7% N 56744
9 XYZ XYZ-001-002 HbA1c (%) HBA1C Visit 5 Week 8 56 61 Drug B Y 8.6 9.5 1.1 <7% N 67543
10 XYZ XYZ-001-002 HbA1c (%) HBA1C Visit 5.1 Week 12 84 84 Drug B Y 8.6 9.5 1.1 Y <7% N LOCF 67543
11 XYZ XYZ-001-002 HbA1c (%) HBA1C Visit 6 Week 24 168 168 Drug B Y 8.6 9.5 1.1 Y <7% N LOCF 67543
Back
Save Data Standards

# CDISC Foundational Standards



Data Collection

**CDASH**

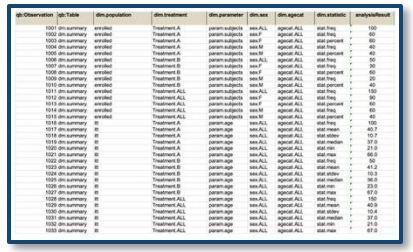


| Row | STUDYID | DOMAIN | USUBJID | CESEQ | CECAT | CETERM | CEDECOD | CEPRESP | CEOCCUR | CESTDTC | CESTDY |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | XYZ | CE | XYZ-001-001 | 2 | HYPO SYMPTOMS | SWEATING | Hyperhidrosis | Y | N | | |
| 3 | XYZ | CE | XYZ-001-001 | 3 | HYPO SYMPTOMS | TREMORS TREMBLING | Tremor | Y | N | | |
| 4 | XYZ | CE | XYZ-001-001 | 4 | HYPO SYMPTOMS | | Dizziness | Y | N | | |
| 5 | XYZ | CE | XYZ-001-001 | 5 | HYPO SYMPTOMS | COGNITIVE IMPAIRMENT | Cognitive Disorder | Y | Y | | |
| 6 | XYZ | CE | XYZ-001-001 | 6 | HYPO SYMPTOMS | LOSS OF CONSCIOUSNESS | Loss of Consciousness | Y | Y | | |
| 7 | XYZ | CE | XYZ-001-001 | 7 | HYPO SYMPTOMS | CONVULSIONS SEIZURES | Convulsion | Y | N | | |
| 8 | XYZ | CE | XYZ-001-001 | 8 | HYPO SYMPTOMS | COMA | Coma | Y | N | | |
| 9 | XYZ | CE | XYZ-001-001 | 9 | HYPO EVENTS | HYPOGLYCEMIC EVENT | Hypoglycaemia | | | 2013-09-24T08:48 | 59 |

| Row | RELMIDS | MIDS | MIDSDTC |
|---|---|---|---|
| 1 (cont) | | HYPO 1 | |
| 2 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 3 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 4 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 5 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 6 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 7 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 8 (cont) | DURING | HYPO 1 | 2013-09-01T11:00 |
| 9 (cont) | | HYPO 1 | |

Data Aggregation

**SDTM**



Analysis

**ADaM**

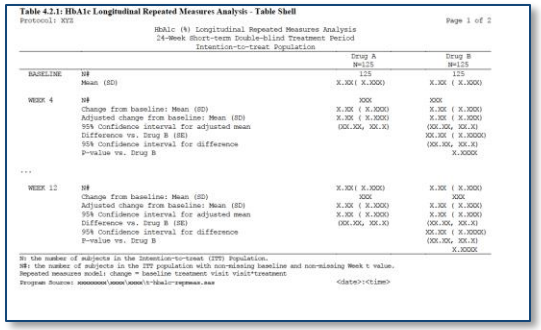


Table 4.2.1: HbA1c Longitudinal Repeated Measures Analysis - Table Shell

Protocol: XYZ

Page 1 of 2

HbA1c (%) Longitudinal Repeated Measures Analysis
24-Week Short-term Double-blind Treatment Period
Intention-to-treat Population

| | | Drug A N=125 | Drug B N=125 |
|---|---|---|---|
| BASELINE | N# | 125 | 125 |
| | Mean (SD) | X.XX( X.XXX) | X.XX ( X.XXX) |
| WEEK 4 | N# | XXX | XXX |
| | Change from baseline: Mean (SD) | X.XX ( X.XXX) | X.XX ( X.XXX) |
| | Adjusted change from baseline: Mean (SD) | X.XX ( X.XXX) | X.XX ( X.XXX) |
| | 95% Confidence interval for adjusted mean | (XX.XX, XX.X) | (XX.XX, XX.X) |
| | Difference vs. Drug B (SE) | | XX.XX ( X.XXXX) |
| | 95% Confidence interval for difference | | (XX.XX, XX.X) |
| | P-value vs. Drug B | | X.XXXX |
| ... | | | |
| WEEK 12 | N# | X.XX( X.XXX) | X.XX ( X.XXX) |
| | Change from baseline: Mean (SD) | XXX | XXX |
| | Adjusted change from baseline: Mean (SD) | X.XX ( X.XXX) | X.XX ( X.XXX) |
| | 95% Confidence interval for adjusted mean | X.XX ( X.XXX) | X.XX ( X.XXX) |
| | Difference vs. Drug B (SE) | (XX.XX, XX.X) | (XX.XX, XX.X) |
| | 95% Confidence interval for difference | | XX.XX ( X.XXXX) |
| | P-value vs. Drug B | | (XX.XX, XX.X) |
| | | | X.XXXX |

N= the number of subjects in the Intention-to-treat (ITT) Population.
n#: the number of subjects in the ITT population with non-missing baseline and non-missing Week t value.
Repeated measures model: change = baseline treatment visit visit*treatment
Program Source: xxxxxxxx\xxxx\xxxx\t-hba1c-repmeas.sas <date>:<time>

Results

**???**

# CDISC Analysis Result Standards – Released April 19, 2024!

cdisc
Clear data. Clear impact

## Analysis Results Standard (ARS) v1.0

cdisc ARS ANALYSIS RESULTS STANDARD

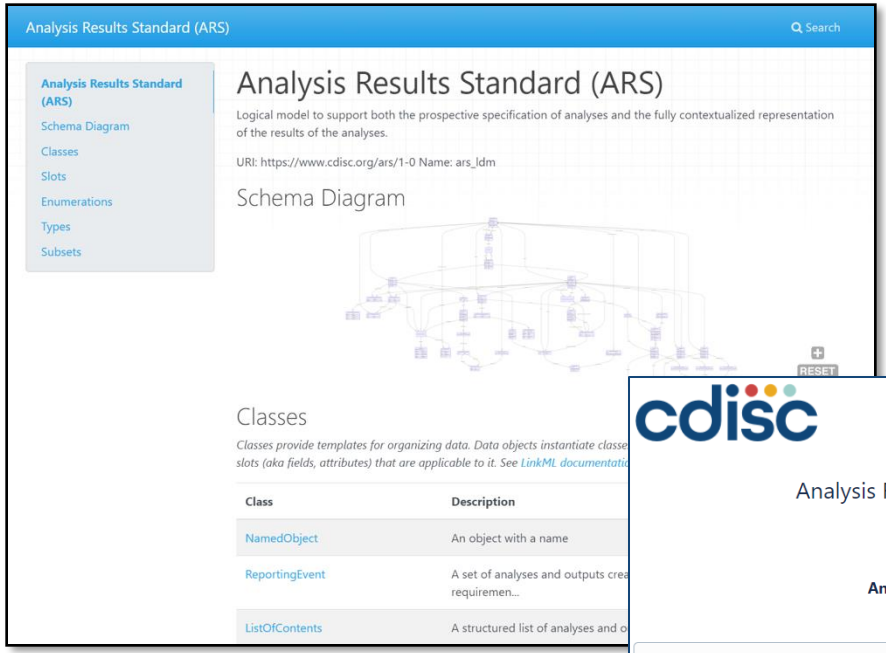
Large trials generate many analysis results in the form of tables, figures, and written reports, yet these results are rarely output in a form that is machine-readable. Previously, there has been no standard way of describing and organizing these results, making it difficult to automate their generation, make them reproducible, trace their origin, or enable them to be reused in other outputs.

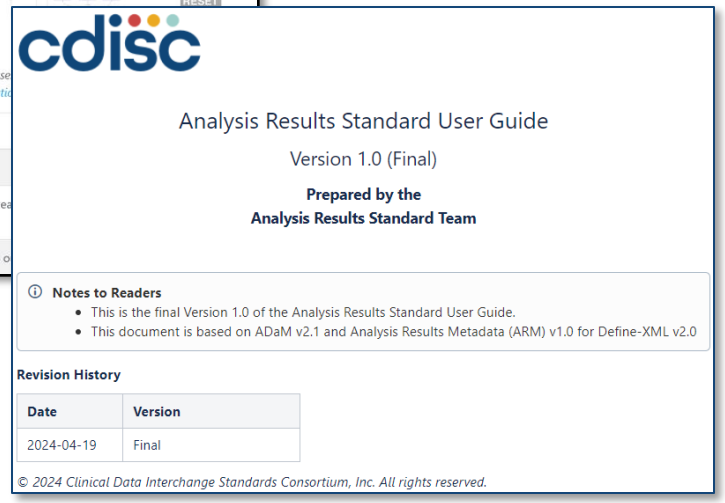
To address these inefficiencies, CDISC has developed the Analysis Results Standard (ARS), which aim to facilitate automation, reproducibility, reusability, and traceability of analysis results data.

### Features of ARS v1.0

- A Logical Data Model that describes analysis results and associated metadata.
- A User Guide to illustrate and exercise the model with common safety displays.

https://cdisc-org.github.io/analysis-results-standard/




https://wiki.cdisc.org/display/ARSP/Analysis+Results+Standard+User+Guide+v1.0

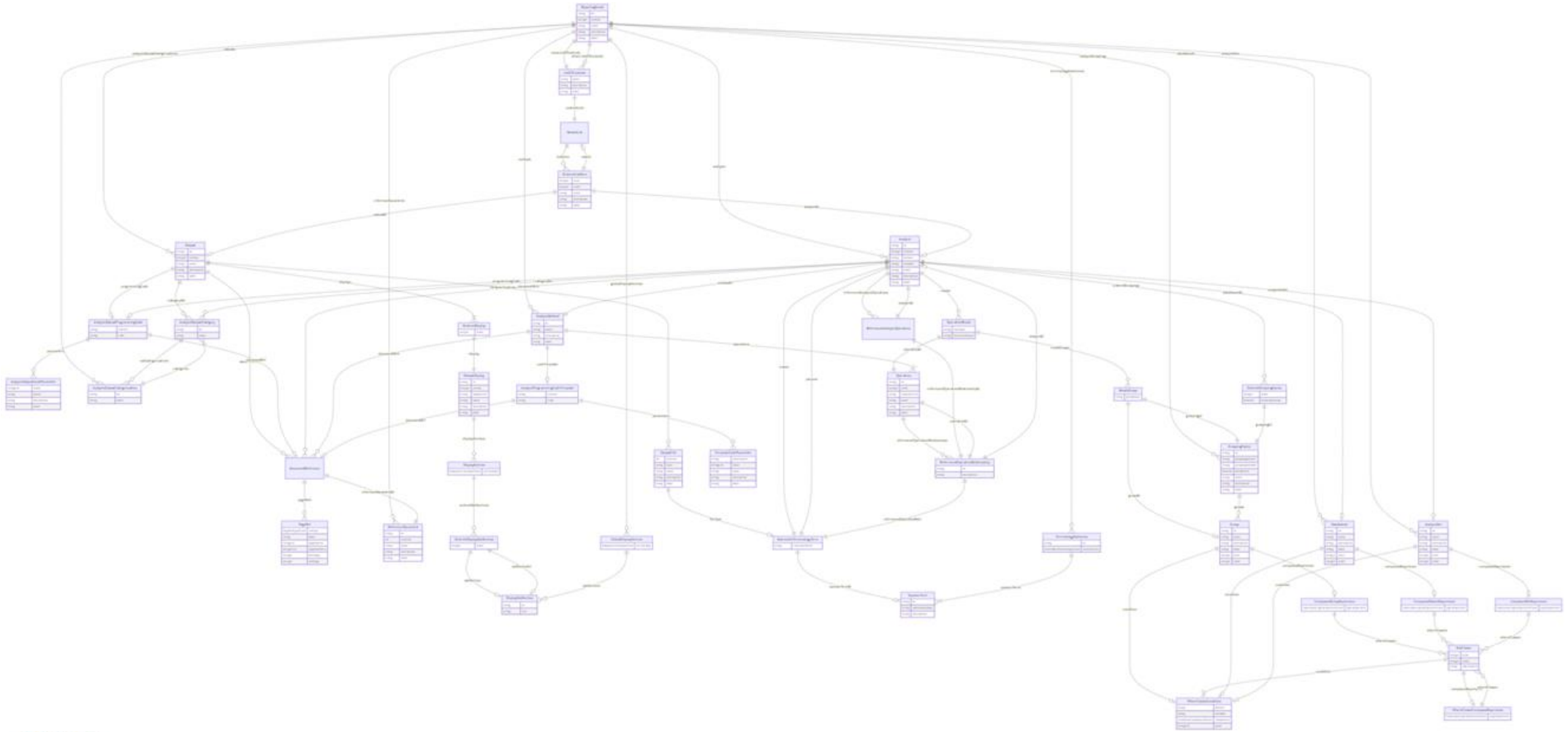
cdisc 25 YEARS ANNIVERSARY

# Analysis Results Key Objectives

Leverage analysis results metadata to drive the automation of results

Support storage, access, processing, traceability and reproducibility of results

# ARS Logical Model Schema Diagram



# Today's specification of analysis

- Written in plain text
  - not machine readable
- Some details are left out
  - often (at best) retrievable from the analysis executable program/define.xml
  - reason why regulators requires SAS programs to be submitted
- Specifications are not software agnostic
  - SAS versus R syntax
  - different set of options

| ***Primary analysis from CDISC pilot study**** | **Issue** |
|---|---|
| *The primary analysis of the ADAS-Cog (11) at Week 24 will use the efficacy population with LOCF imputation for any missing values at Week 24.* | define.xml revealed that it was Change from baseline (CHG) that was analysed. |
| *An ANCOVA model will be used with the **baseline score**, **site**, and **treatment** included as **independent variables**.*<br><br>***Treatment** will be included as a **continuous** variable, and results for a **test of dose response** will be produced.*<br><br>***Interaction terms will not be investigated**.* | The define.xml specifies how the analysis is done in SAS, bot how to translate to R or other programming language? |
| *If the test for **dose response is statistically significant**, pairwise comparisons among the 3 groups will be performed and evaluated at a significance level of 0.05.* | Pairwise estimates made – but test for dose response was significant ($p < 0.05$) |

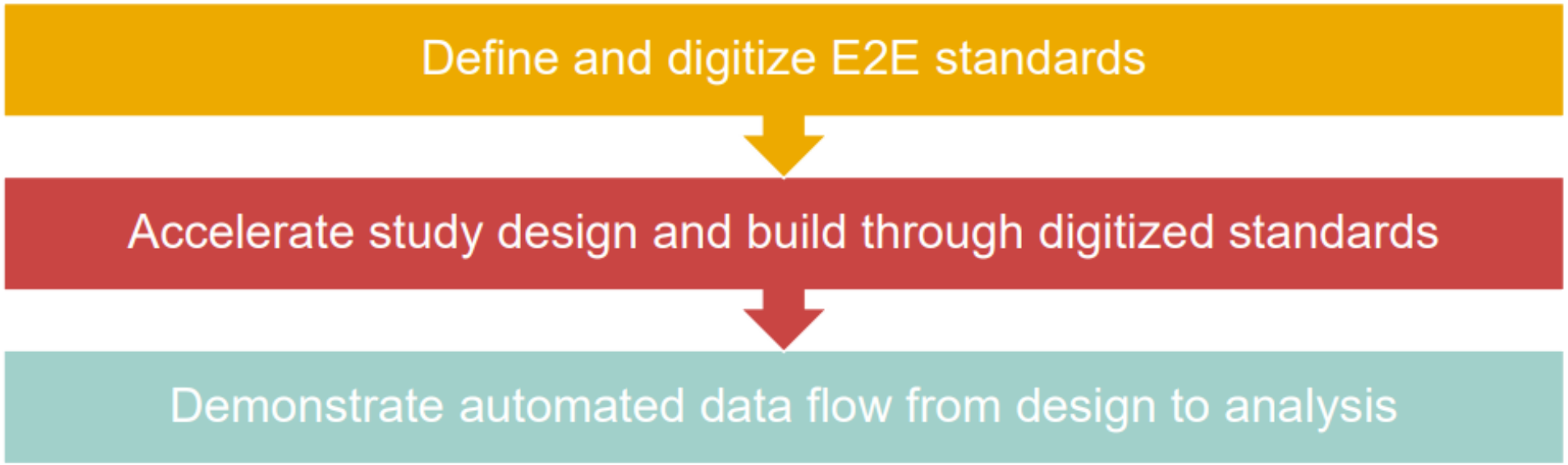
*know the document is old, but very likely the same is seen in newer documents

# Why Develop Analysis Concepts?

- Reduces ambiguity
  - Analysis Concepts with standardized metadata structure enforce precision in specifying analysis settings and assumptions
- Enables machine-readable analysis plans
  - Analysis Concepts structured as metadata rather than narrative text, can be directly linked to statistical programming code
- Supports traceability
  - Analysis Concepts help maintain clear linkage between protocol objectives, endpoints, and the specific analytical methods applied
- Streamlining collaboration
  - Analysis Concepts provide a common language between statisticians, clinicians, data managers, and other stakeholders

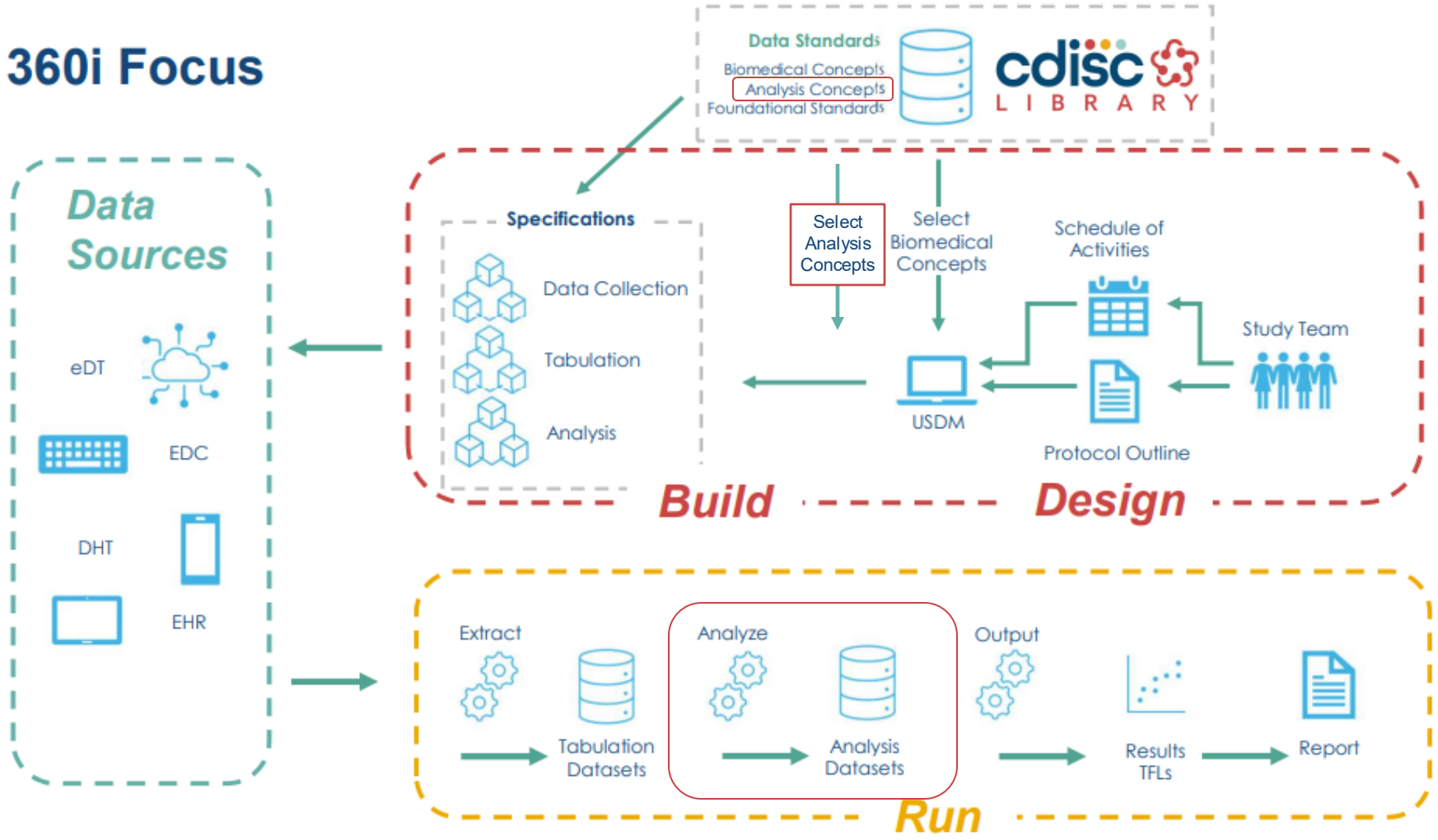
# Analysis Concepts – scope of CDISC 360i

## 360i Journey: Ideas → Implementation → Common Practice




From 360i-Program-Kickoff webinar – 2025-FEB-18

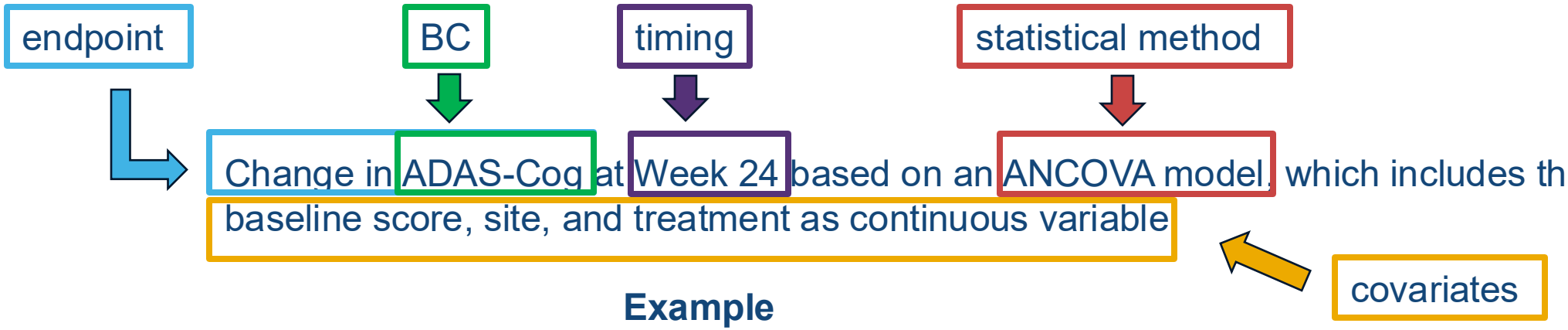
# Analysis Concepts in 360i: Use for Downstream Automation



# Analysis Concepts

- What is an Analysis Concept?
  - Collection of standardized configurable elements & methods that are needed to describe a final output or answer a clinical question
  - Standard definitions in a digital statistical analysis plan

- What is the purpose of an Analysis Concept?
  - Supports configuration and automation of analysis
  - Informs what needs to be collected to answer the clinical question




**Example**

# Working Towards a Logical Model

endpoint

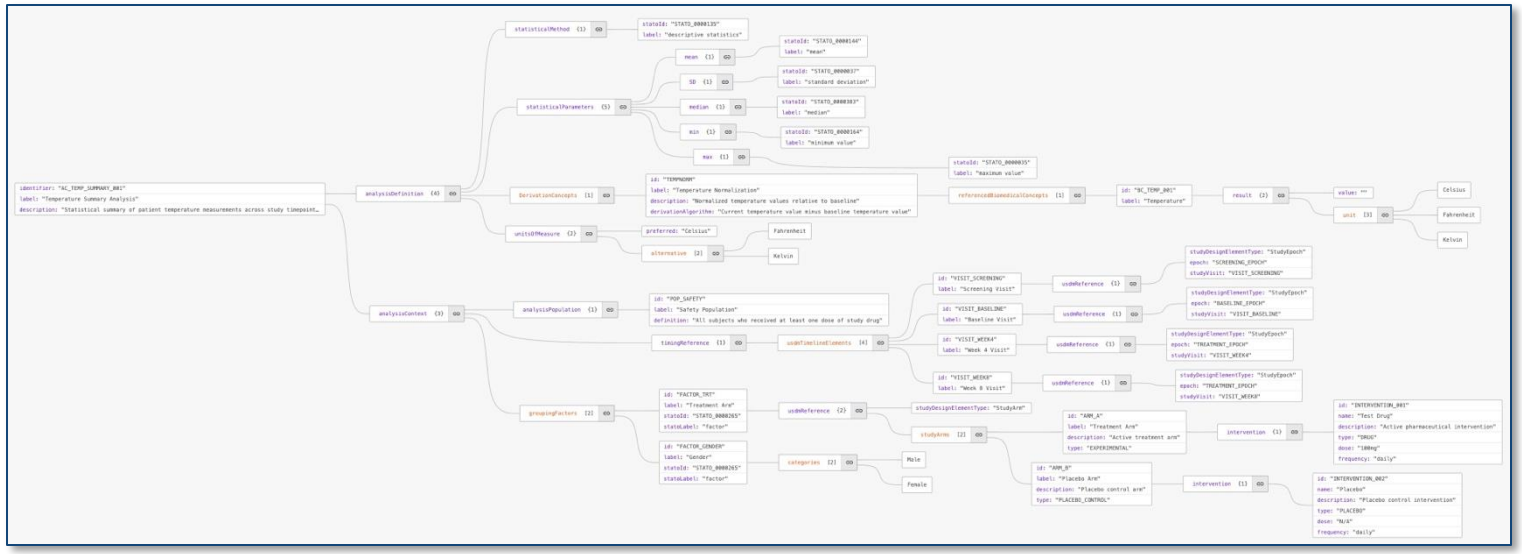
BC

timing

statistical method

Change in ADAS-Cog at Week 24 based on an ANCOVA model, which includes the baseline score, site, and treatment as continuous variable

covariates

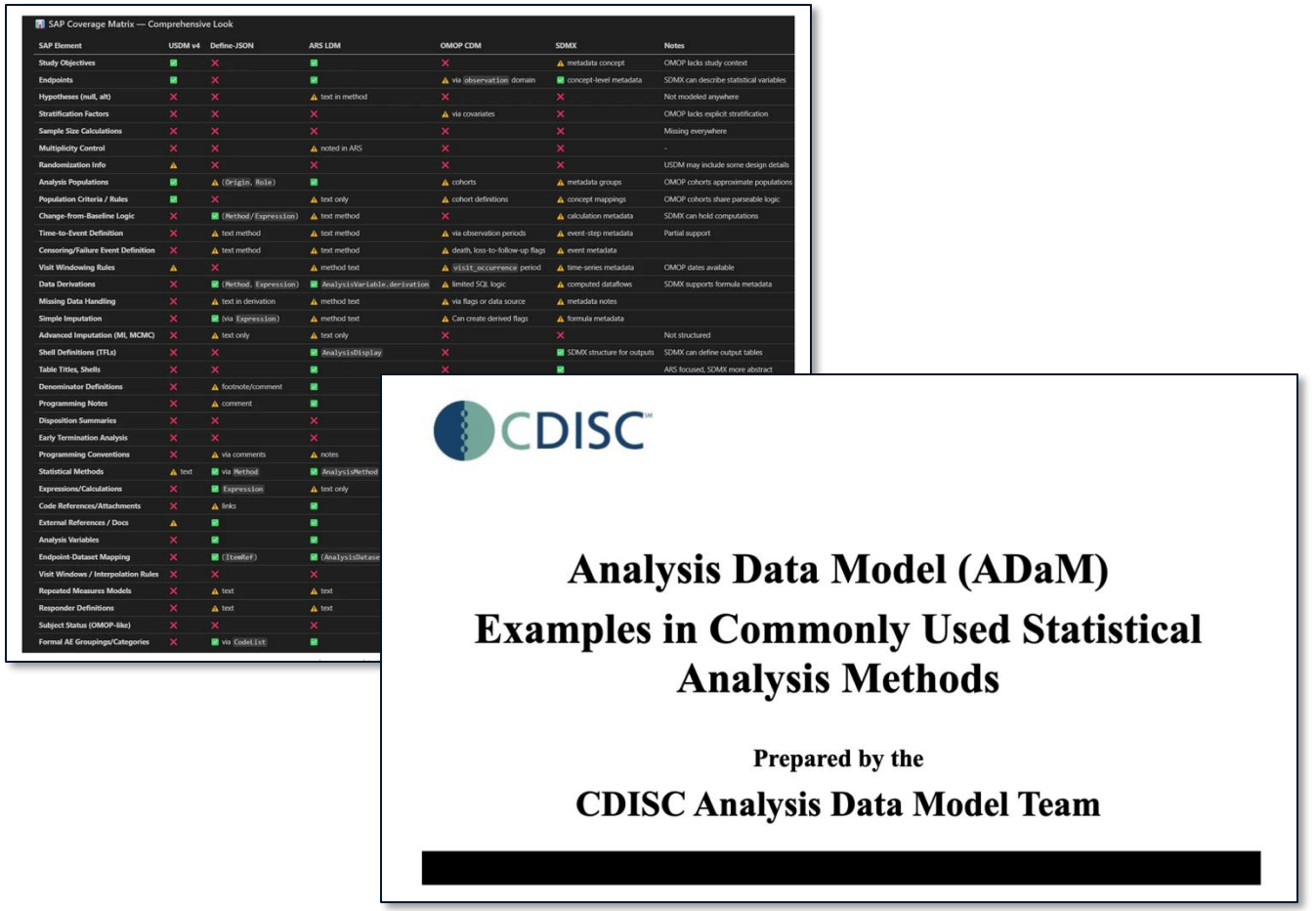


# AC DC

| | Analysis Concept (AC) | Derivation Concept (DC) |
|---|---|---|
| Purpose | Examines existing data or information to draw conclusions, identify patterns, or test hypotheses | Generates new derived values from raw or derived data |
| Direction | Typically works with completed measurements or observations to extract meaning | Transforms or processes data to create new representations |
| Process | Involves applying statistical methods, critical thinking, and interpretative frameworks to understand data | Uses mathematical operations, formulas, or algorithms to calculate new quantities |
| Outcome | Produces insights, conclusions, or evaluations based on the data - aggregated data (not subject-level) | Creates derived data (subject level) that serve as inputs for subsequent analysis or derivations |
| Example 1 | The p-value (from Type III Sums of Squares for treatment dose), based on linear model analysis of CHG for dose response; using randomized dose and site group in model. | CHG: Change from Baseline to Week 24 in ADAS Cog (11). Use LOCF is missing value at week 24. |
| Example 2 | Mean value of CHG by visit | CHG: Change from BASELINE in ADAS Cog (11) by visit<br>BASELINE = 'Y' if ADAS Cog (11) at visit 2 |

# Approaches Considered

- Top-Down
  - What are the element covered across sample SAPs?
  - What element are covered across standards?
- Bottom-up
  - Build a picture of an analysis concept using the most common statistical analyses



# Development Plan

## Subset SAPs

- Use a script to download and parse SAPs
- Extract the primary analysis description

## Tokenization and Pattern Recognition

- Tokenize text into meaningful components
- Identify common patterns
- Visualize relationships

## Iterative Modeling

- Build generalized model of analysis concepts using tokenized elements
- Refine based on new feedback and examples
- Compare and validate against existing models (USDM, ARS)

## Template Development

- Common analyses
- Software agnostic

# Important that the statistical community see the vision and benefits

- ‘Automation’ stakeholders
  - CDISC
  - PHUSE
  - USDM dev team
  - TCB / Pharma-Biotech
  - CRO’s
  - Vendors

- ‘Scientific’ stakeholders
  - Regulators
    - FDA, EMA, PMDA
  - Statistical communities
    - United States
      - Society for Clinical Trials (SCT) - https://www.sctweb.org/
      - Statistical Analysis and Modelling in Clinical Development (SAMCD) - https://community.amstat.org/biop/home (Part of the ASA Biopharmaceutical Section)
      - Drug Information Association (DIA) - https://www.diaglobal.org/
    - Europe
      - European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) - https://www.efspi.org/
      - Statisticians in the Pharmaceutical Industry (PSI) - https://www.psiweb.org/
      - European Clinical Research Infrastructure Network (ECRIN) - https://ecrin.org/

# Challenges

- Defining the 80 percent (80/20 rule)
- Defining the boundaries between USDM, Analysis Results Standard, Biomedical Concepts, and Analysis Concepts
- How and if to leverage existing ontologies such as STATO
- How and if to leverage Statistical Data and Metadata eXchange (SDMX)
  - Set of technical standards designed to describe statistical data and metadata (ISO 17369)
- How do we make this relatable to the CDISC community?
  - Creation of an eSAP?
- Limited volunteer time

# Thank You!

Contact:

Bess LeRoy bleroy@cdisc.org