

TARGETED MAPPING TO SDTM ENABLES LEGACY DATA REUSE WITHOUT FULL CONVERSION

Bereket Tesfaldet, Helena Sviglin, Tejas Patel, Eileen Navarro Almario, for MATIG
CDER/OTS/OCS and DCVS/NHLBI/NIH



Introduction

The Meta-Analytical Interagency Group (MATIG) is composed of multidisciplinary scientists from the National Heart, Lung, and Blood Institute (NHLBI), Food and Drug Administration (FDA), and other academic institutions. Our research interest is to advance public health outcomes in cardio-metabolic disorders using existing data from publicly available sources such as NHLBI data repository (BioLINCC). Disparate formats, inadequate metadata, and limited traceability from data-as-collected to analysis data are practical challenges to this reuse. We outline the approach developed to overcome these challenges by building a common data model (CDM) based on the essential CDISC standard domains to answer a specific research question, enabling data reuse with limited legacy transformation.

1. State Research Hypothesis

In subjects with T2DM, serum HDL-cholesterol (HDL-C) raising and/or triglycerides (TG) lowering is associated with a reduced risk of heart failure (HF).

2. Develop a Statistical Analysis Plan

- Time-to-first-event analyses to compare the occurrence of HF among 4 groups based on % change in HDL-C and TG from baseline to 6 months
 - Achieves both HDL-C and TG targets
 - Achieves neither target
 - Achieves only HDL-C target
 - Achieves only TG target.
- Perform multivariate analysis to adjust for age, gender, body mass index (BMI), waist circumference, established cardiovascular disease (CVD), history of hypertension, diabetes, or metabolic syndrome, and estimated glomerular filtration rate [eGFR].

3. Identify Relevant Data Elements

- Identify the data elements required to answer the research question and build a CDM based on the SDTM standard.
- Identifying the elements needed to perform the analyses facilitates the identification, extraction and harmonization of the core data from individual studies that are in their legacy data model (LDM).

4. Build a Compendium to Identify Relevant Studies

- A compendium is a framework that describes details needed to identify and stack data elements across databases for the specific research purpose.
- It facilitates a priori determination whether all needed data elements that support the meta-analytic research concepts are present in each new data source and whether they are sufficiently uniform to enable extraction into the pooled database.

Research Concept	RCT	Data Element (Definition)	Source/ Type	Stackability* Assessment (Limitations)
Endpoint	Trial 1	HF (FRS criteria)	CRF: binary variable	Good candidate to support the research question. (Adjudication status unknown.)
	Trial 2	HF	Data not available	Study is not able to support the research question.
	Trial 3	HF (FRS criteria)	CRF: binary variable	Good candidate to support the research question. (Adjudication status unknown, but verified by physician's statement in medical record.)
Population	Trial 1	T2D status (based on 1997 ADA criteria)	CRF: binary variable	Good candidate for a rigorous definition of disease status. Data verified by study staff.
	Trial 2	T2D status (patient self-report, medical record, previous treatment w/ insulin or OHAs)	CRF: binary variable	Suboptimal candidate to support the research question. (Disease status less rigorously defined.)
	Trial 3	T2D status (patient self-report, medical record, current treatment w/ insulin or OHAs)	CRF: binary variable	Suboptimal candidate to support the research question. (Disease status less rigorously defined.)

* Stackability refers to qualities resident in separate data sources (like structure and content) that enable data to be 'stacked' or integrated into one database with little transformation or curation. Assessing the appropriateness of data sources in a meta-analysis is the investigator's responsibility and includes an assessment of data attributes that enable stacking for data re-use.
Abbreviations: ADA, American Diabetic Association; CRF, Case Report Form; FRS, Framingham Risk Score; HF, Heart Failure; OHAs, Oral Hypoglycemic Agents; T2D, Type 2 Diabetes Mellitus.

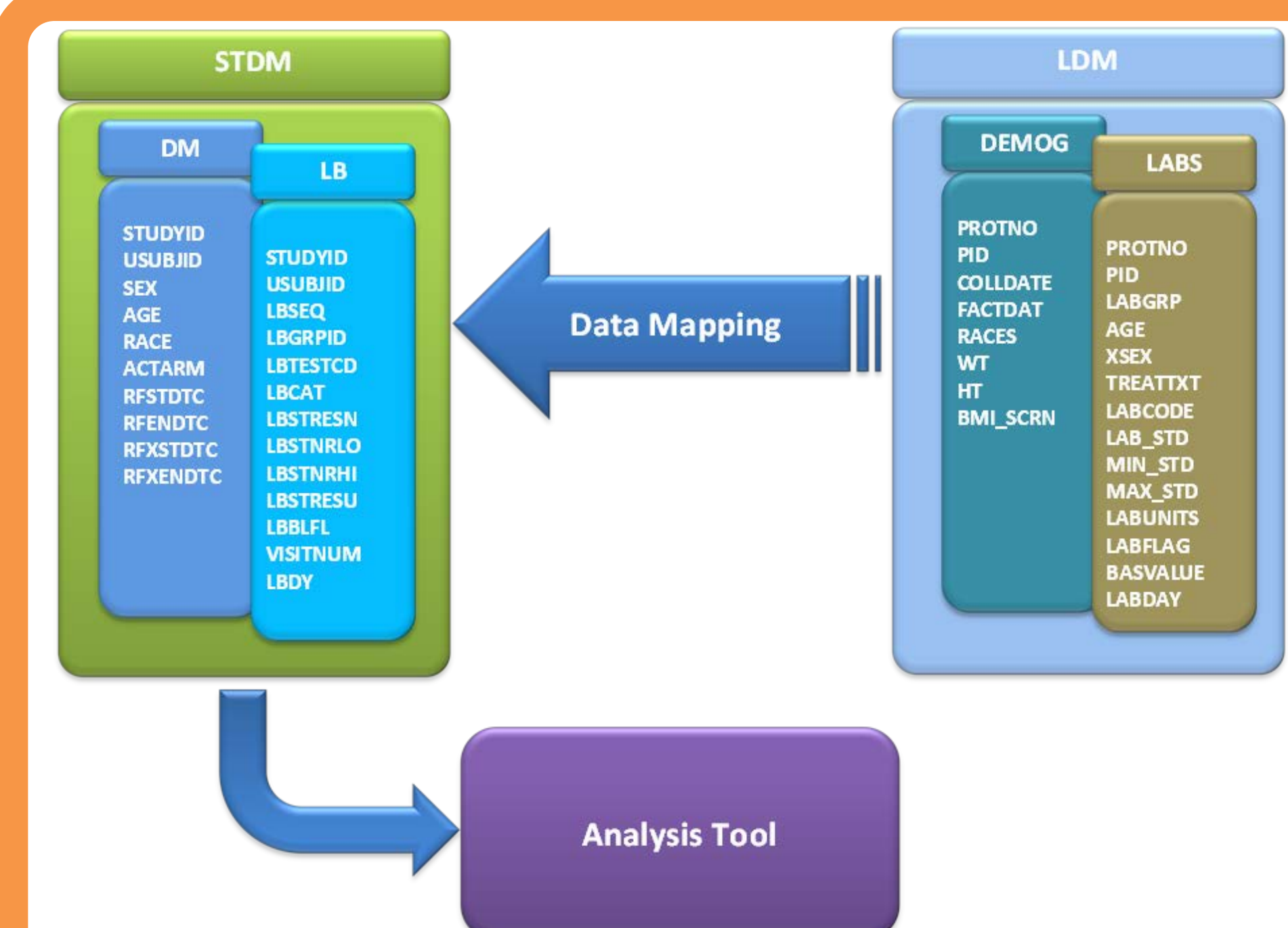
5. Build a Data Mapping Scheme

- For each data source, create a data mapping spreadsheet that maps legacy data elements to CDM data elements.

CDM Variables			LDM Variables			
Name	Type	Label	Study	Dataset	Variable	Derivation
STUDYID	Char	Study Identifier	TRIAL1	Protocol		TRIAL1
USUBJID	Char	Unique Subject ID	TRIAL1	TRIAL1_key	MASKID	
AGE	Num	Age	TRIAL1	TRIAL1_key	BASELINE_AGE	
SEX	Char	Sex	TRIAL1	TRIAL1_key	FEMALE	0 = 'M', 1 = 'F'
RACE	Char	Race	TRIAL1	TRIAL1_key	RACECLASS	
BMI	Num	Baseline Body Mass Index	TRIAL1	baselinehist	WT_KG, HT_CM	10000*WT_KG/(HT_CM*H
ARM	Char	Treatment Arm	TRIAL1	oryphysical	TREATMENT	
STUDYID	Char	Study Identifier	TRIAL3	Protocol		TRIAL3
USUBJID	Char	Unique Subject ID	TRIAL3	TRIAL3_bl	ID	
AGE	Num	Age	TRIAL3	TRIAL3_bl	AGE	
SEX	Char	Sex	TRIAL3	TRIAL3_bl	SEX	1 = 'M', 2 = 'F'
RACE	Char	Race	TRIAL3	TRIAL3_bl	RACE	
BMI	Num	Baseline Body Mass Index	TRIAL3	TRIAL3_bl		
ARM	Char	Treatment Arm	TRIAL3	TRIAL3_bl		

CDM Variables			LDM Variables			
Name	Type	Label	Study	Dataset	Variable	Derivation
STUDYID	Char	Study Identifier	TRIAL1	Protocol		TRIAL1
USUBJID	Char	Unique Subject ID	TRIAL1	TRIAL1_key	MASKID	
PARAMCD	Char	Parameter Code	TRIAL1	CVDOutcomes		'CHF'
EVNTDSC	Char	Event or Censor	TRIAL1	CVDOutcomes		'Chf'
AVAL	Num	Analysis Value	TRIAL1	CVDOutcomes	FUYRS_CHF	
STUDYID	Char	Study Identifier	TRIAL3	Protocol		TRIAL3
USUBJID	Char	Unique Subject ID	TRIAL3	TRIAL3_bl	ID	
PARAMCD	Char	Parameter Code	TRIAL3	TRIAL3_pe	ETYPE	26 = 'CHF'
EVNTDSC	Char	Event or Censor	TRIAL3	Derived		Chf
AVAL	Num	Analysis Value	TRIAL3	TRIAL3_pe, TRIAL3_endpts	DAYEVENT, DEATHFU	if CNSR=0, min(DAYEVENT), DEATHFU

6. Implement the Data Mapping and Run the Analyses



Conclusion

- A targeted mapping of data elements from clinical trials identified in a compendium to the SDTM format enables patient-level meta-analysis without full conversion of trial data.
- This approach can save the cost and time needed to do full conversion of disparate clinical trials data into common data format.

Disclaimer: The views expressed in this presentation are those of the authors and do not represent FDA policy.