2026 Korea CDISC Day

29 October 2026 · Seoul, Korea

Define-XML Origin

Lessons Learned Before Automating CDISC

Miae Lee · Product Manager · JNPMEDI

Speaker



Miae Lee

Product Manager
JNPMEDI

Miae Lee brings more than fifteen years across clinical practice, clinical trials, systems development and clinical data work to building digital data solutions. At JNPMEDI, Miae plans clinical data standards software, leading the product planning and management of Maven Converter, which converts clinical data into the CDISC SDTM and ADaM models and authors and validates Define-XML.

Before moving to industry, Miae spent nearly eight years at Seoul National University Hospital as a clinical data manager and research coordinator in the Division of Colorectal Surgery, building the division's own data programs and writing their specifications. An earlier career covered hospital information systems and anaesthesia nursing.

Miae holds a BS in Nursing and a BS in Information Science, and is completing an MS in Health Informatics and Biostatistics. Credentials include CDISC SDTM and ARS certificates, Certified MedDRA Coder, and SAS Certified Specialist.

This talk comes out of reading the standards closely enough to plan software against them.

Disclaimer and Disclosures

> *The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the official policy or position of CDISC.*

Agenda

1. **What validation does not catch**
   It checks that an Origin is there. It almost never asks whether the one you chose is right.
2. **What Origin records, and what it commits you to**
   Type, Source, the values each model allows, and the element each Type must point to
3. **Why Predecessor is built differently**
   It drops Source, and names the source variable inside Description instead
4. **When one variable has two origins**
   Value-level metadata, or two Origin elements — and on what basis?
5. **The decision is a person's to make**
   How to choose a Type, where it gets difficult, and what a system can and cannot do for you

What the rules actually check

2,790 rules check that an Origin is there — almost none ask whether it is right

They do check

- That an Origin is there
- That Type and Source are controlled terms
- Derived → MethodOID; Collected from an Investigator or Subject → aCRF page

They do not check

- Which Type is right for this value
- Whether a Predecessor target exists
- Whether the data agree with what was declared

Where a Type is checked	What it fixes	Scope
CDISC TIG0048 (ancestor SEND111)	BRTHDTC should be Collected	Nonclinical
Pinnacle 21 DD0105–DD0108	--DY Derived · DOMAIN, RDOMAIN Assigned · STUDYID Protocol	Warning (PMDA)

Counted: Define-XML v2.1 CR, SDTM CR v2.0, ADaM CR v5.0, TIG CR v1.0 — 2,790 unique rules, 4,744 entries · P21 DD0148 checks only that a Predecessor reads DATASET.VARIABLE · DD0109/DD0110 check only that USUBJID and VISITNUM keep one Origin across datasets

Why Origin matters — traceability in the reviewer's words

Origin is where the trace is written down — one value per variable, chosen by a person

What the reviewer has to do

- Understand how the analysis datasets were built
- Determine which observations and algorithms derived a variable
- Understand how a p-value was calculated
- Relate report counts back to the data

> *“Based upon reviewer experience, establishing traceability is one of the most problematic issues associated with any data conversion.”*
>
> FDA sdTCG v6.1, §8.1.3, p.68

> *“If the reviewer is unable to trace study data from the data collection of subjects participating in a study to the analysis of the overall study data, then the regulatory review of a submission may be compromised.”*
>
> FDA sdTCG v6.1, §8.1.3, p.68

FDA Study Data Technical Conformance Guide v6.1 (December 2025), §8.1.3 Overview, p.68 — reviewer tasks paraphrased from its four bullets

Purpose of Define-XML

The define.xml is the reviewer's map of what was submitted

SDTM Define-XML	ADaM Define-XML
Datasets and variables	Datasets and variables
Controlled terminology	Controlled terminology
Value-level metadata	Parameter-level metadata
Data origin	Derivations
	Traceability to SDTM

> *"… where the approval process for regulated human and animal health products may require the submission of data from clinical trials or other studies, Define-XML is among the standards that may be required as part of a complete data submission."*

Define-XML v2.1 §1.2, p.7

> *"One of the key benefits to FDA reviewers is that this format provides both a machine-readable format … and … a browser-based report describing the contents of a study."*

Define-XML v2.1 §1.1, p.5

Origin: what it records

Type says how the value came to be. Source says who is responsible.

Type C170449 · non-extensible

Term	NCI code
Collected	C170548
Derived	C170549
Assigned	C170547
Protocol	C170551
Predecessor	C170550
Not Available	C126101
Other	C17649

> *"Representation of SDTM Origin metadata has been enhanced to identify the source in addition to the origin details and to support multiple origins."*
>
> Define-XML v2.1 §1.1.3

Source C170450 · non-extensible

Term	NCI code
Subject	C41189
Investigator	C25936
Vendor	C68608
Sponsor	C70793

Each model allows fewer

Model	Allowed
SDTM	5 — all but Not Available
ADaM	3 — Derived, Assigned, Predecessor
SEND	7 — Not Available for legacy only

Define-XML CT 2025-09-26 · Define-XML v2.1 §4.3.2.1–3, §5.3.12.3 · ADaM's three are enforced by P21 DD0073, not by CT — the text only advises

The Origin Types, in the specification's own words

Collected

“A value that is actually observed and recorded by a person or obtained by an instrument. Note that a collected entry translated to a synonymous controlled term still has a type Collected.”

Derived

“A value that is calculated by an algorithm or reproducible rule, and which is dependent upon other data values, including data values available within the dataset or externally provided data values. MethodDef must be used to document the algorithm or rule used for a derived value.”

Assigned

“Data that is either: determined by individual judgment as provided by an evaluator, or coded terms supplied as part of a coding process, or values set independently of any subject-related data value in order to complete a dataset.”

Protocol

“Data that is defined as part of the study protocol, investigator instructions, standard operating procedures or trial design preparation.”

Predecessor

“An entry that is copied from a variable in another dataset. The Description child element identifies the dataset and variable that is copied.”

Not Available

“Used when the origin is not available and cannot be determined. Sponsors should specify additional details that may be helpful to the reviewer in the Comments section of the data definition file.”

Your model allows fewer — SDTM five, ADaM three, SEND keeps Not Available for legacy data

Define-XML v2.1 §4.3.2, p.30–31 — the specification defines these six; CT adds Other

Valid combinations — which Type may be used with which Source

Source depends on Type — and some pairs are simply not allowed

SDTM

Type	Subject	Investigator	Vendor	Sponsor
Collected	ePRO	CRF	Lab data, ECG	X
Derived	X	X	Lab data, ECG	SDTM
Assigned	X	X	Adjudicator	SDTM
Protocol	X	X	X	SDTM
Predecessor	X	X	X	X

ADaM

- Derived, Assigned → Source is always Sponsor
- Predecessor → no Source

SEND

- Type only — no Source
- Not Available: legacy data only

X = not valid · Define-XML v2.1 §4.3.2.1 (SDTM table), §4.3.2.2 (ADaM), §4.3.2.3 (SEND) · Collected “should not be used for data that have been interpreted, calculated, or derived from other information”

5.3.12.3 def:Origin Element

The def:Origin element is intended to be define the Origin metadata for variables in SDTM or SEND datasets or source metadata for ADaM datasets. The use of the ODM Origin attribute within Define-XML has been deprecated and is not valid in Define-XML v2.1.

Element Name	def:Origin
Element XPath(s)	/ODM/Study/MetaDataVersion/ItemDef/def:Origin
Element Textual Value	None

Usage	• Requirement: Conditional
	  ○ For regulatory submissions, def:Origin metadata must be provided for all SDTM, ADaM, or SEND variables.
	  ○ In some cases, def:Origin will be provided only at the Value level.
	• Cardinality: Zero or more

	○ Multiple def:Origins may be provided when there are multiple sources for a single Variable but no way to construct a ValueList that clearly identifies the cases where each def:Origin applies.
	• Business Rules:
	  ○ If all data rows for the variable represented by the parent ItemDef have the same origin, def:Origin should be provided at the variable level.
	  ○ If the parent ItemDef includes a def:ValueListRef and each ItemDef element referenced in the corresponding def:ValueListDef includes a def:Origin element, there should be no def:Origin on the parent ItemDef.
	  ○ If the variable or value is derived, the corresponding ItemRef must include a MethodOID attribute that references the corresponding MethodDef.
	  ○ When def:Origin/@Type="Collected" and def:Origin/@Source in ("Investigator", "Subject"), there must be a def:DocumentRef child element; in addition, def:DocumentRef/@leafID must match the ID attribute of a def:leaf element corresponding to the annotated CRF document.

Attributes	Type, Source
Child Elements	def:DocumentRef, Description

Attribute	Usage	Allowable Values	Description
Type	Required	Value Description: Origin type • The list of allowable values is not extensible and is expected to be available as CDISC Controlled Terminology in the near future. Once available, the list of allowable values in the "Allowable Values" column will reference the C-Code and name of the "Origin Type" codelist. • See Section 4.3.2, Origin/Source/Traceability Considerations	Business Rule: If the variable is derived, a MethodDef must be provided.
Source	Conditional	Value Description: Data source • Source is required for SDTM and ADaM datasets when Origin Type is not Predecessor. Source is not required for SEND datasets. • See Section 4.3.2, Origin/Source/Traceability Considerations	Business Rule: The list of allowable values for each def:Origin/@Source value is not extensible and is expected to be available as CDISC Controlled Terminology in the near future. Once available, the list of allowable values in the "Allowable Values" column will reference the C-Code and name of the "Origin Source" codelist subsets.

def:Origin — what §5.3.12.3 says

What it is

- Origin of SDTM and SEND variables; source of ADaM variables
- The ODM Origin attribute is deprecated — not valid in v2.1

When it is needed

- Regulatory submissions: every SDTM, ADaM and SEND variable **(must)**
- Sometimes given only at the value level
- Several per variable only when no ValueList can show where each applies

Business rules

- All rows share one origin → give it at the variable level **(should)**
- Every value-level ItemDef has one → none on the parent **(should)**
- Derived → the ItemRef carries a MethodOID to its MethodDef **(must)**
- Collected from an Investigator or Subject → a DocumentRef to the annotated CRF **(must)**

Attributes

- Type — Required; a non-extensible list now Define-XML CT C170449
- Source — required for SDTM and ADaM unless Predecessor; not for SEND C170450

Define-XML v2.1 §5.3.12.3 def:Origin Element, p.82–83 · must / should as written in the specification

Origin is traceability metadata — what each Type must carry

Choose a Type, and the document owes the reviewer a link

If Type is	The document must carry	Which links to
Derived	ItemRef/@MethodOID	MethodDef
Collected · Investigator or Subject	def:DocumentRef + def:PDFPageRef	aCRF page
Predecessor	A Description child	dataset.variable — as text
Assigned · Protocol	Nothing further	—

> *"For SDTM data collected using a CRF, the variable definition must include a reference to pages in the aCRF to provide traceability information for reviewers."*
>
> Define-XML v2.1 §4.6, p.53

> *"The algorithm is linked to a variable or value using the MethodOID attribute in the corresponding ItemRef element."*
>
> Define-XML v2.1 §4.7, p.54

Define-XML v2.1 §4.3.2 p.30 · §5.3.7.1.1 · §5.3.12.3 — Collected · Vendor needs no page reference; Assigned and Protocol carry no stated requirement

Predecessor — Description replaces Source

A copied value has no Source — the document says where it was copied from instead

> *“When Origin is Predecessor, the Source attribute is not used and the def:Origin has a child element Description that contains a reference to the dataset variable from which the value was copied, in the form <dataset name>.<variable name>.”*
>
> Define-XML v2.1 §4.3.2.2, p.31

```
<ItemDef OID="IT.ADAE.STUDYID" Name="STUDYID" DataType="text" Length="12">
   <Description><TranslatedText>Study Identifier</TranslatedText></Description>
   <def:Origin Type="Predecessor">
      <Description><TranslatedText>ADSL.STUDYID</TranslatedText></Description>
   </def:Origin>
</ItemDef>
```

- No Source attribute
- Two Descriptions — the label, and the source reference

ADaM-MSG v1.0 sample define.xml (Define-XML v2.1) — 366 variables carry Predecessor

In the sample: Predecessor

Predecessor names the source variable — so the trace can continue

ADaM — ADADAS.TRTSDT, TRTEDT

Variable	Where Condition	Label / Description	Type	Length or Display	Controlled Terms or ISO Format	Origin / Source / Method / Comment
TRTSDT		Date of First Exposure to Treatment	integer	DATE9.		Predecessor: ADSL.TRTSDT
TRTEDT		Date of Last Exposure to Treatment	integer	DATE9.		Predecessor: ADSL.TRTEDT

No Source — the Description carries dataset.variable instead

SDTM — EX.SPDEVID, EPOCH

Variable	Label / Description	Type	Role	Length or Display	Controlled Terms or ISO Format	Origin / Source / Method / Comment
SPDEVID	Sponsor Device Identifier	text	Identifier	200		Predecessor: EC.SPDEVID
EPOCH	Epoch	text	Timing	9	Epoch • "SCREENING" = "Screening" • "TREATMENT" = "Treatment"	Predecessor: EC.EPOCH

CDISC ADaM-MSG v1.0 and SDTM-MSG v2.0 sample packages — define.xml rendered with the packaged stylesheet

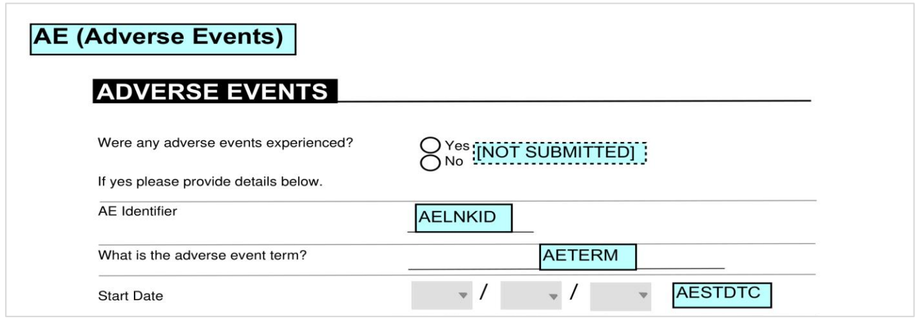
In the sample: Collected

Collected from an Investigator — the define.xml points to the aCRF page

define.xml — AE.AELNKID

Variable	Origin / Source / Method / Comment
AELNKID	[Origin] Collected (Source: Investigator) Annotated CRF [22 23]

aCRF page 22 — where that reference lands



A Vendor source would carry no page — the value never passed through the CRF

CDISC SDTM-MSG v2.0 sample submission package — define.xml rendered with define2-1.xsl, acrf.pdf

In the sample: Derived

Choose Derived, and the algorithm comes with it

define.xml — EX.EXDOSE

Variable	Label / Description	Type	Role	Length or Display	Controlled Terms or ISO Format	Origin / Source / Method / Comment
EXDOSE	Dose	integer	Record Qualifier	8		Derived (Source: Sponsor) EXDOSE = ECDOSE * ECPSTRG expressed in mg.

… and its MethodDef in the Methods table

Method	Type	Description
Algorithm to derive EXDOSE	Computation	EXDOSE = ECDOSE * ECPSTRG expressed in mg.

> *"This attribute and the associated MethodDef are Required when the Type attribute for the def:Origin child element of the referenced ItemDef is "Derived"."*
>
> Define-XML v2.1 §5.3.9.2 — ItemRef/@MethodOID, p.72

CDISC SDTM-MSG v2.0 sample submission package — define.xml rendered with define2-1.xsl, acrf.pdf

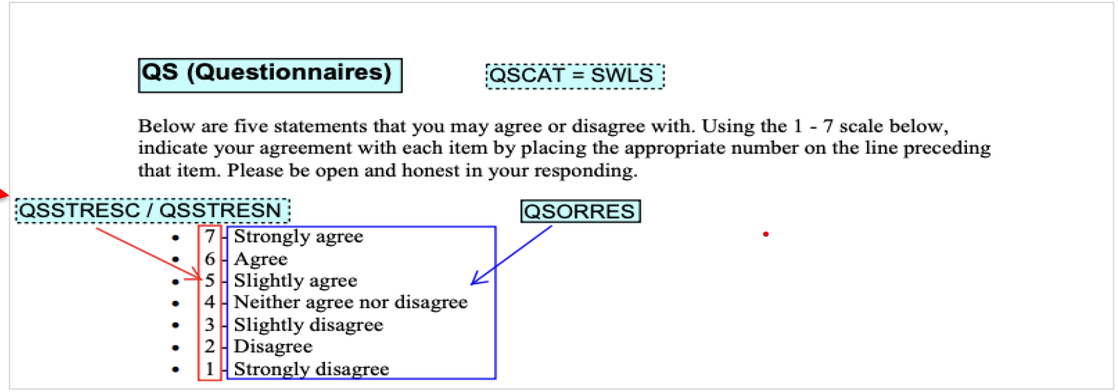
In the sample: Derived, with an aCRF page

Derived can carry an aCRF page too — here it shows the scale the Method maps to

define.xml — QSSL.QSSTRESC

QSSTRESC	[Origin] Derived (Source: Sponsor) Annotated CRF [16] [Method]If QSORRES="Strongly disagree" then 1 If QSORRES="Disagree" then 2 If QSORRES="Slightly disagree" then 3 If QSORRES="Neither agree nor disagree" then 4 If QSORRES="Slightly agree" then 5 If QSORRES="Agree" then 6 If QSORRES="Strongly agree" then 7

aCRF page 16



The page reference is optional for Derived. The sample adds it so the reviewer can see the printed 1–7 scale the Method converts.

> *"Required for def:Origin/@Type="Collected" and def:Origin/@Source in ("Investigator","Subject") — Optional in all other cases"*
>
> Define-XML v2.1 §5.3.7.1.1 — def:PDFPageRef, p.68

CDISC SDTM-MSG v2.0 sample — define.xml (speaker's capture; Variable and Origin columns) and acrf.pdf p.16 · 6 items in the sample are Derived with a page reference

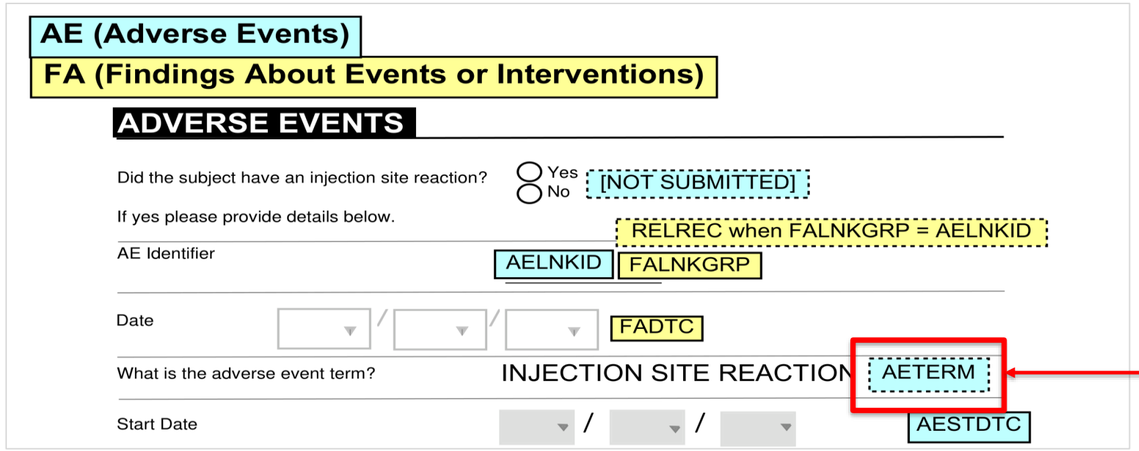
In the sample: Assigned

Assigned needs no page — the sample added one to explain the value

define.xml — AE.DOMAIN, USUBJID, AETERM

Variable	Where Condition	Label / Description	Type	Role	Length or Display	Controlled Terms or ISO Format	Origin / Source / Method / Comment
DOMAIN		Domain Abbreviation	text	Identifier	2	SDTM Domain Abbreviation, subset used for Adverse Events • "AE" = "Adverse Events"	Assigned (Source: Sponsor)
USUBJID		Unique Subject Identifier	text	Identifier	8		Assigned (Source: Sponsor)
AETERM VLM		Reported Term for the Adverse Event	text	Topic	200		
	AETERM = "INJECTION SITE REACTION"	Reported Term for the Adverse Event	text		200		Assigned (Source: Sponsor) Annotated CRF [23 Even though the variable is 'Assigned' an annotation has been added to page 23 to clarify the assignment.

aCRF page 23



CDISC SDTM-MSG v2.0 sample submission package — define.xml rendered with define2-1.xsl, acrf.pdf

In the sample: Protocol

Protocol — a value the protocol itself fixed

define.xml — EX.STUDYID and EX.EXDOSU

Variable	Label / Description	Type	Role	Length or Display	Controlled Terms or ISO Format	Origin / Source / Method / Comment
STUDYID	Study Identifier	text	Identifier	12		Protocol (Source: Sponsor)
EXDOSU	Dose Units	text	Variable Qualifier	2	Unit, subset to be used for EXDOSU • "mg" = "Milligram"	Protocol (Source: Sponsor)

> *"Data that is defined as part of the study protocol, investigator instructions, standard operating procedures or trial design preparation"*
>
> Define-XML v2.1 §4.3.2, p.30

CDISC SDTM-MSG v2.0 sample submission package — define.xml rendered with define2-1.xsl, acrf.pdf

When Origin is not uniform

If the origin changes from row to row, the document has to say where

> *“Sponsors are cautioned to recognize that a derived origin means that all values for that variable were derived ... When both derived and collected values are reported in a variable, the origin is to be described using value-level metadata in the Define-XML document.”*
>
> SDTMIG v3.4 §4.1.8.2, p.26

The situation	What to write
All rows share one origin	def:Origin on the variable
The condition can be written	ValueList + Where Clause, def:Origin per value
The condition can not be written	Several def:Origin on one ItemDef — last resort

Define-XML v2.1 §5.3.12.3: several def:Origins only when there is “no way to construct a ValueList that clearly identifies the cases where each def:Origin applies”; with a ValueList, “there should be no def:Origin on the parent ItemDef”

Two origins in one variable — how CDISC handled it

CDISC's own sample split DM.RACE by value — Collected and Assigned

> *“For Demographics (DM), RACE had multiple origins, ‘Collected’ for the 5 individual races on the CRF, and ‘Assigned’ for the value of MULTIPLE ... Origin was handled in the value-level metadata where different codelists were added to each with only the value(s) possible within that instance.”*
>
> SDTM-MSG v2.0, §2 Define-XML Document, p.5

Where it is written	What it says
Define-XML §5.3.12.3	Two def:Origin only if no ValueList is possible
SDTMIG §4.1.8.2	Mixed derived and collected → value level
SDTM-MSG v2.0	RACE handled in value-level metadata
FDA sdTCG · PMDA	Nothing on this choice

The standard never says “recommended” — the alternative opens only when no ValueList can be built

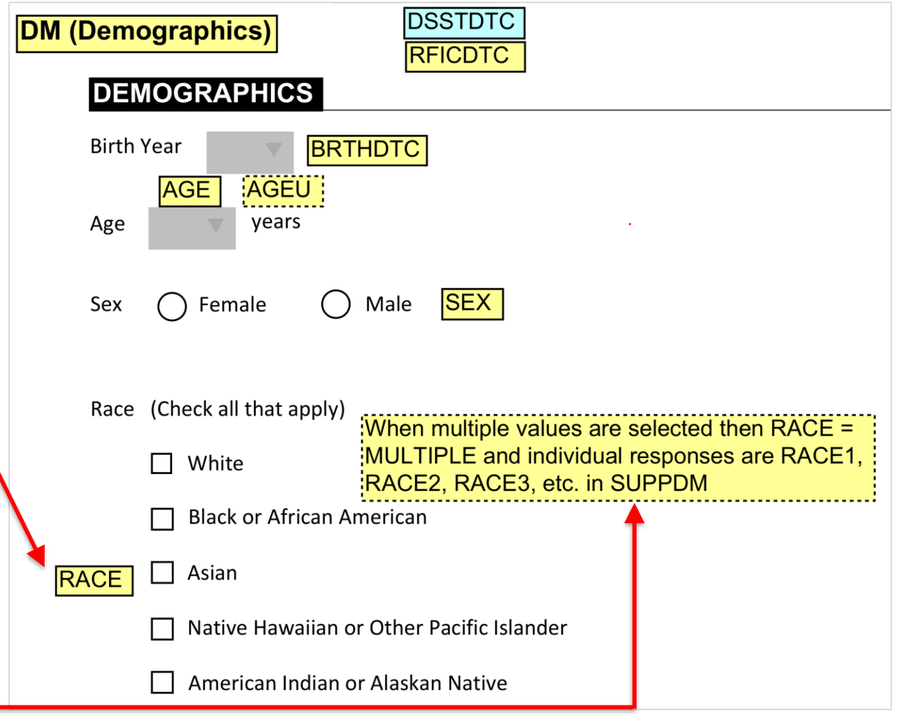
In the sample: one variable, two origins

DM.RACE — Collected for the five races, Assigned when the answer is MULTIPLE

define.xml — DM.RACE value-level metadata

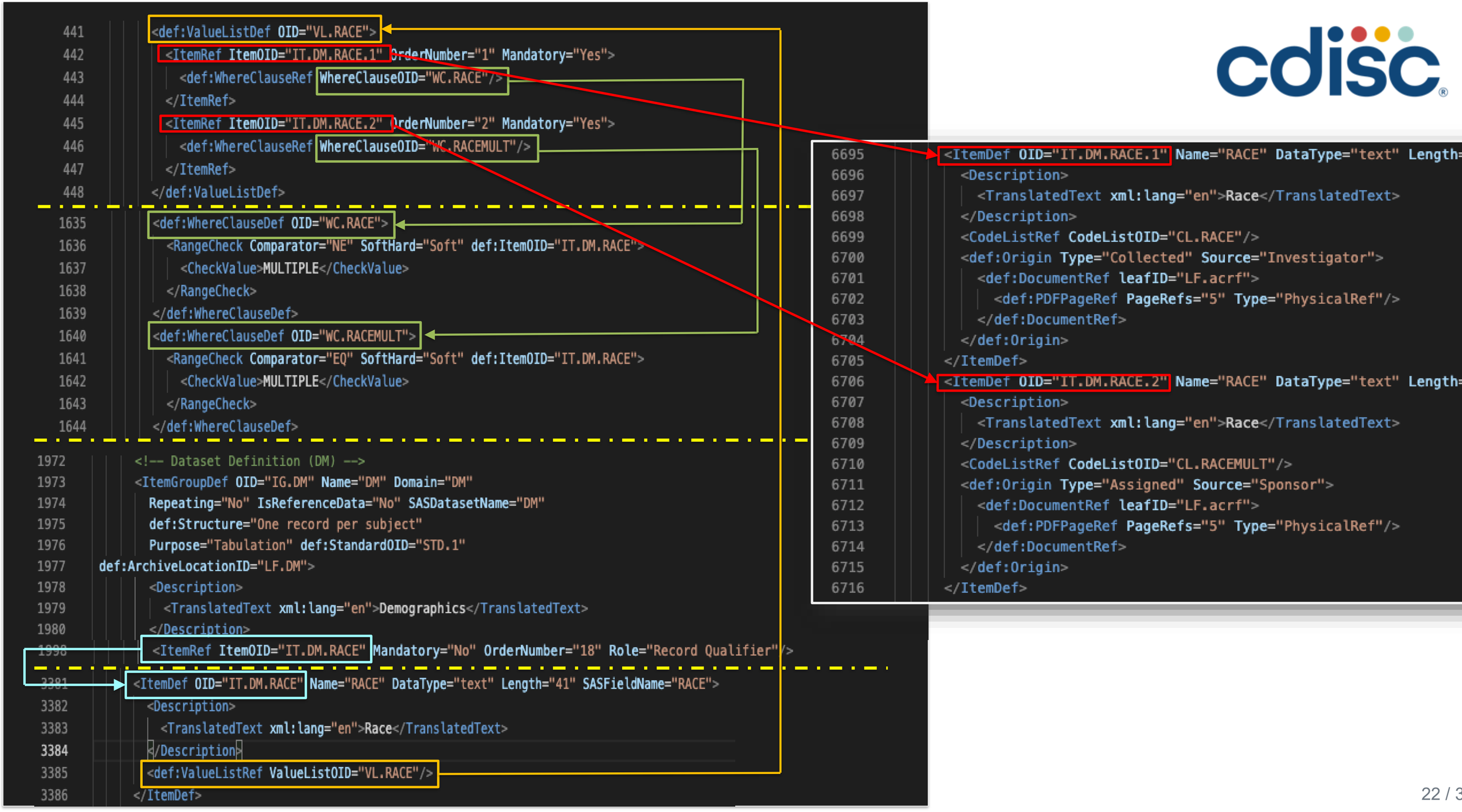
RACE VLM		Race	text	Record Qualifier	41		
	RACE ≠ "MULTIPLE"	Race	text		41	Race • "AMERICAN INDIAN OR ALASKA NATIVE" = "American Indian Or Alaska Native" • "ASIAN" = "Asian" • "BLACK OR AFRICAN AMERICAN" = "Black Or African American" • "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER" = "Native Hawaiian Or Other Pacific Islander" • "WHITE" = "White"	[Origin] Collected (Source: Investigator) Annotated CRF [5]
	RACE = "MULTIPLE"	Race	text		8	Race Multiple Selections • "MULTIPLE" = "Multiple"	[Origin] Assigned (Source: Sponsor) Annotated CRF [5]

aCRF page 5 — where both rows point



Same page, two origins — the checkboxes are collected; “MULTIPLE” is the sponsor's rule written next to them

CDISC SDTM-MSG v2.0 sample submission package — define.xml (speaker's capture; Label, Type, Role and Length columns omitted) and acrf.pdf p.5


441 <def:ValueListDef OID="VL.RACE">
442 <ItemRef ItemOID="IT.DM.RACE.1" OrderNumber="1" Mandatory="Yes">
443 <def:WhereClauseRef WhereClauseOID="WC.RACE"/>
444 </ItemRef>
445 <ItemRef ItemOID="IT.DM.RACE.2" OrderNumber="2" Mandatory="Yes">
446 <def:WhereClauseRef WhereClauseOID="WC.RACEMULT"/>
447 </ItemRef>
448 </def:ValueListDef>
1635 <def:WhereClauseDef OID="WC.RACE">
1636 <RangeCheck Comparator="NE" SoftHard="Soft" def:ItemOID="IT.DM.RACE">
1637 <CheckValue>MULTIPLE</CheckValue>
1638 </RangeCheck>
1639 </def:WhereClauseDef>
1640 <def:WhereClauseDef OID="WC.RACEMULT">
1641 <RangeCheck Comparator="EQ" SoftHard="Soft" def:ItemOID="IT.DM.RACE">
1642 <CheckValue>MULTIPLE</CheckValue>
1643 </RangeCheck>
1644 </def:WhereClauseDef>
1972 <!-- Dataset Definition (DM) -->
1973 <ItemGroupDef OID="IG.DM" Name="DM" Domain="DM"
1974 Repeating="No" IsReferenceData="No" SASDatasetName="DM"
1975 def:Structure="One record per subject"
1976 Purpose="Tabulation" def:StandardOID="STD.1"
1977 def:ArchiveLocationID="LF.DM">
1978 <Description>
1979 <TranslatedText xml:lang="en">Demographics</TranslatedText>
1980 </Description>
1998 <ItemRef ItemOID="IT.DM.RACE" Mandatory="No" OrderNumber="18" Role="Record Qualifier"/>
3381 <ItemDef OID="IT.DM.RACE" Name="RACE" DataType="text" Length="41" SASFieldName="RACE">
3382 <Description>
3383 <TranslatedText xml:lang="en">Race</TranslatedText>
3384 </Description>
3385 <def:ValueListRef ValueListOID="VL.RACE"/>
3386 </ItemDef>
6695 <ItemDef OID="IT.DM.RACE.1" Name="RACE" DataType="text" Length="41"
6696 <Description>
6697 <TranslatedText xml:lang="en">Race</TranslatedText>
6698 </Description>
6699 <CodeListRef CodeListOID="CL.RACE"/>
6700 <def:Origin Type="Collected" Source="Investigator">
6701 <def:DocumentRef leafID="LF.acrf">
6702 <def:PDFPageRef PageRefs="5" Type="PhysicalRef"/>
6703 </def:DocumentRef>
6704 </def:Origin>
6705 </ItemDef>
6706 <ItemDef OID="IT.DM.RACE.2" Name="RACE" DataType="text" Length="8"
6707 <Description>
6708 <TranslatedText xml:lang="en">Race</TranslatedText>
6709 </Description>
6710 <CodeListRef CodeListOID="CL.RACEMULT"/>
6711 <def:Origin Type="Assigned" Source="Sponsor">
6712 <def:DocumentRef leafID="LF.acrf">
6713 <def:PDFPageRef PageRefs="5" Type="PhysicalRef"/>
6714 </def:DocumentRef>
6715 </def:Origin>
6716 </ItemDef>

Which one to use — the test is traceability

Both are valid XML. Only one can be followed back to a value.

Value-level metadata

```
<def:ValueListDef OID="VL.RACE">
  <ItemRef ItemOID="IT.DM.RACE.1">
    <def:WhereClauseRef WhereClauseOID="WC.RACE"/>
  <ItemRef ItemOID="IT.DM.RACE.2">
    <def:WhereClauseRef WhereClauseOID="WC.RACEMULT"/>
<ItemDef OID="IT.DM.RACE.1" Name="RACE">
  <def:Origin Type="Collected" Source="Investigator"/>
<ItemDef OID="IT.DM.RACE.2" Name="RACE">
  <def:Origin Type="Assigned" Source="Sponsor"/>
```

- Each value has its own Origin
- The rows are named in a RangeCheck

Two def:Origin on one variable

```
<ItemDef OID="IT.DM.RACE" Name="RACE">
  <def:Origin Type="Collected" Source="Investigator"/>
  <def:Origin Type="Assigned" Source="Sponsor"/>
```

- Both happened
- Nothing says which rows

Ask: reading only this document, could a reviewer tell where each origin applies?

Define-XML v2.1 §4.3.2.2, §4.5.1, §5.3.12.3 · SDTMIG v3.4 §4.1.8.2

Where Clause across datasets

A where clause may reach into DM — and then the join must be written down

> *“Where Clauses can only contain references to: 1. Variables in the current dataset 2. Variables in other subject-level datasets within the same metadata version in a Define-XML document. In this case, the implied dataset join should be documented in a comment attached to the WhereClauseDef.”*

Define-XML v2.1 §4.5.1, p.46

> *“Required when RangeCheck includes def:ItemOID values that belong to different ItemGroupDef elements”*

Define-XML v2.1 §5.3.10 — def:CommentOID

```
<def:WhereClauseDef OID="WC.VS.VSTESTCD.WEIGHT.[DM].COUNTRY.CMETRIC"
      def:CommentOID="COM.SUBJECTDATA-JOIN-DM">
  <RangeCheck def:ItemOID="IT.VS.VSTESTCD" Comparator="EQ"><CheckValue>WEIGHT</CheckValue></RangeCheck>
  <RangeCheck def:ItemOID="IT.DM.COUNTRY" Comparator="IN">
      <CheckValue>CAN</CheckValue><CheckValue>MEX</CheckValue></RangeCheck>
</def:WhereClauseDef>
```

§4.5.1 says “should”; §5.3.10 says “Required” — and Section 5 is normative

Example 4.5.1.1, p.46 · its CommentDef is Example 4.5.2.2.1, p.50 · §1.1.1: “The content of this section is normative” · §4.5.1 Note: “there is no mechanism for machine-readable join specifications”

Written in the protocol — does that make it Protocol?

A formula written in the protocol is still Derived

> *"Other derivation examples: calculations within ePRO (e.g., questionnaire section scores) and calculations within EDC (e.g., BMI, BSA)."*

Define-XML v2.1 §4.3.2.1, p.31

> *"An example would be VSPOS (Vital Signs Position), which could be specified in the protocol and be provided by other means (e.g. CRF, eDT)."*

Define-XML v2.1 §4.3.2.1, p.31

The formula goes into the MethodDef

"Per the formula in protocol §7.2, BMI = WEIGHT / (HEIGHT/100)²"

Derived	X	X	Lab data, ECG	SDTM	Derivation examples include calculations performed during data collection (e.g., --DY). Other derivation examples: calculations within ePRO (e.g., questionnaire section scores) and calculations within EDC (e.g., BMI, BSA).

The only Protocol example in the spec

```
<ItemDef OID="IT.STUDYID" Name="STUDYID" ...>
  <def:Origin Type="Protocol"/>
```

No Source — though §4.3.2.1 requires one

The test is not where the rule came from — it is what the value depends on

Define-XML v2.1 §4.3.2 p.30, §4.3.2.1 p.31, Example 4.3.3.2.1 p.34

Where the judgment gets hard — seven cases side by side

A stored value that differs from the source is not the test

The case	Origin
SEX collected as 1 / 2, stored as M / F	Collected · Investigator
--DY, computed from RFSTDTC	Derived · Sponsor
A result from a central laboratory	Collected · Vendor
A questionnaire total score	Derived · Sponsor
AEDECOD, from MedDRA coding	Assigned · Sponsor
BMI, computed by the sponsor	Derived · Sponsor
EX values carried over from EC	Predecessor

Derived asks one thing: does this value depend on any data value other than itself?

Define-XML v2.1 §4.3.2 p.30, §4.3.2.1 p.31 · SDTMIG v3.4 §4.1.8, §4.4.4, §6.3.5.6 LB Assumption 6 · SDTM-MSG v2.0 §5.3.1 — a device or lab that computes a value by its own procedure reports a Collected value

How to decide — six questions, and six to try

1. Copied unchanged from another CDISC dataset?
2. Did the protocol fix the value itself — not the formula, and not the instruction to collect it?
3. Observed or measured — by a person, an instrument, or a lab using its own procedure?
4. Computed by us, or by EDC / ePRO, from data already collected?
5. Set without looking at subject data — a look-up, a code?
6. Do the answers differ inside one variable?



Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	
DOMAIN	
VSPOS — the protocol says sitting	
ADSL.TRTSDT carried into ADAE unchanged	
SEX collected as 1 / 2, stored as M / F	
VSORRES — results collected, BMI computed	

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

How to decide — six questions, and six to try

1. Copied unchanged from another CDISC dataset?
2. Did the protocol fix the value itself — not the formula, and not the instruction to collect it?
3. Observed or measured — by a person, an instrument, or a lab using its own procedure?
4. Computed by us, or by EDC / ePRO, from data already collected?
5. Set without looking at subject data — a look-up, a code?
6. Do the answers differ inside one variable?

Predecessor

Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	
DOMAIN	
VSPOS — the protocol says sitting	
ADSL.TRTSDT carried into ADAE unchanged	Predecessor
SEX collected as 1 / 2, stored as M / F	
VSORRES — results collected, BMI computed	

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

How to decide — six questions, and six to try

1 Copied unchanged from another CDISC dataset? **Predecessor**

2 Did the protocol fix the value itself — not the formula, and not the instruction to collect it? **Protocol**

3 Observed or measured — by a person, an instrument, or a lab using its own procedure?

4 Computed by us, or by EDC / ePRO, from data already collected?

5 Set without looking at subject data — a look-up, a code?

6 Do the answers differ inside one variable?



Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	
DOMAIN	
VSPOS — the protocol says sitting	Protocol · Sponsor
ADSL.TRTSDT carried into ADAE unchanged	Predecessor
SEX collected as 1 / 2, stored as M / F	
VSORRES — results collected, BMI computed	

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

How to decide — six questions, and six to try

1	Copied unchanged from another CDISC dataset?	**Predecessor**
2	Did the protocol fix the value itself — not the formula, and not the instruction to collect it?	**Protocol**
3	Observed or measured — by a person, an instrument, or a lab using its own procedure?	**Collected**
4	Computed by us, or by EDC / ePRO, from data already collected?	
5	Set without looking at subject data — a look-up, a code?	
6	Do the answers differ inside one variable?	

Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	
DOMAIN	
VSPOS — the protocol says sitting	Protocol · Sponsor
ADSL.TRTSDT carried into ADAE unchanged	Predecessor
SEX collected as 1 / 2, stored as M / F	Collected · Investigator
VSORRES — results collected, BMI computed	

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

How to decide — six questions, and six to try

1	Copied unchanged from another CDISC dataset?	**Predecessor**
2	Did the protocol fix the value itself — not the formula, and not the instruction to collect it?	**Protocol**
3	Observed or measured — by a person, an instrument, or a lab using its own procedure?	**Collected**
4	Computed by us, or by EDC / ePRO, from data already collected?	**Derived**
5	Set without looking at subject data — a look-up, a code?	
6	Do the answers differ inside one variable?	

Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	Derived · Sponsor
DOMAIN	
VSPOS — the protocol says sitting	Protocol · Sponsor
ADSL.TRTSDT carried into ADAE unchanged	Predecessor
SEX collected as 1 / 2, stored as M / F	Collected · Investigator
VSORRES — results collected, BMI computed	

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

How to decide — six questions, and six to try

1	Copied unchanged from another CDISC dataset?	**Predecessor**
2	Did the protocol fix the value itself — not the formula, and not the instruction to collect it?	**Protocol**
3	Observed or measured — by a person, an instrument, or a lab using its own procedure?	**Collected**
4	Computed by us, or by EDC / ePRO, from data already collected?	**Derived**
5	Set without looking at subject data — a look-up, a code?	**Assigned**
6	Do the answers differ inside one variable?	

Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	Derived · Sponsor
DOMAIN	Assigned · Sponsor
VSPOS — the protocol says sitting	Protocol · Sponsor
ADSL.TRTSDT carried into ADAE unchanged	Predecessor
SEX collected as 1 / 2, stored as M / F	Collected · Investigator
VSORRES — results collected, BMI computed	

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

How to decide — six questions, and six to try

1	Copied unchanged from another CDISC dataset?	**Predecessor**
2	Did the protocol fix the value itself — not the formula, and not the instruction to collect it?	**Protocol**
3	Observed or measured — by a person, an instrument, or a lab using its own procedure?	**Collected**
4	Computed by us, or by EDC / ePRO, from data already collected?	**Derived**
5	Set without looking at subject data — a look-up, a code?	**Assigned**
6	Do the answers differ inside one variable?	**Value-level metadata**

Stop at the first yes — now try these six	Origin
AGE, calculated from BRTHDTC and RFSTDTC	Derived · Sponsor
DOMAIN	Assigned · Sponsor
VSPOS — the protocol says sitting	Protocol · Sponsor
ADSL.TRTSDT carried into ADAE unchanged	Predecessor
SEX collected as 1 / 2, stored as M / F	Collected · Investigator
VSORRES — results collected, BMI computed	Value-level metadata

In ADaM only Derived, Assigned and Predecessor pass P21 DD0073 · Define-XML §4.3.2.1, Ex. 4.3.3.2.1 · SDTMIG §6.3.5.6 · P21 DD0106, DD0108 · ADaM-MSG sample define.xml

Three sentences worth remembering

1 **Calculated inside the measurement: Collected. Calculated on collected data: Derived.**

"… dependent upon other data values, including data values available within the dataset" — Define-XML v2.1 §4.3.2

2 **Filtering rows is not deriving. If the value did not change, it was copied.**

Taking some rows through a join keeps it Predecessor

3 **Predecessor means copied between CDISC datasets. Anything from outside is Collected.**

A value pulled from a raw lab file is not a Predecessor

Define-XML v2.1 §4.3.2, §4.3.2.1, §4.3.3.4 · SDTMIG v3.4 §4.1.8, §6.3.5.6 LB Assumption 6

Source, and part of Type, are outside the pipeline

1 Source is a fact from outside the data

A lab value and a CRF value look identical in the file

2 Protocol and Assigned are the same action

Writing a constant looks the same either way

3 An honest record is not a good design

Recompute what could be copied: Derived, valid, and wrong

> *“AGE ^= null AND AGE def:Origin/@Type = "Derived" → Algorithm is included in the Define-XML document.”*
>
> CDISC TIG0049 (ancestor SEND112.1)

CDISC put the judgment in the condition, not the rule — the standard declines to fix the Type

What the standard leaves unverified

The Predecessor target is free text — and nothing resolves it

> *“OIDs play a critical role in presenting Define-XML as one interconnected document containing a web of multiple parts that are linked together to form a single set of metadata.”*
>
> Define-XML v2.1 §3.5.1

```
<!-- CDISC ARM 1.0 example, ADQSADAS -->
<ItemDef OID="IT.ADQSADAS.EFFFL" ...>
  <def:Origin Type="Predecessor">
    <Description>ADSL.FASFL</Description>
```

- DD0061: something is there
- DD0148: it reads DATASET.VARIABLE
- ADSL has no FASFL — and it passes

- CDISC's SDTM-MSG v2.0 sample: 16 of 18 references are a bare name — ECTRT, not EC.ECTRT

P21 has 12 cross-reference rules. Eleven resolve the target. The Predecessor rule checks only the shape.

And in ADaM, Type="Derived" Source="Vendor" passes too — DD0073 checks Type, nothing checks Source

Define-XML v2.1 §3.5.1, §4.3.2.2 (ADaM Source is always Sponsor), §5.3.9.1 · define-ns.xsd (Description minOccurs=0) · P21 Community 2405.2 DD0061, DD0073, DD0148 (PMDA Warning, no CDISC mapping)

If a system converts, will it fill Origin correctly?

For much of Type, yes — the transformation already knows what it did

What the transformation did	Origin	And it can fill in
Applied a formula	Derived	The MethodDef
Copied a CDISC value	Predecessor	dataset.variable
Took a field on the CRF	Collected	The aCRF page
Set a constant or a code	Assigned	—

None of this needs inference — it is done with rules

- Type decides which Sources are offered
- Predecessor fills Origin and Description
- Candidates: registered SDTM variables only
- A flow derivation becomes the MethodDef
- Required follows the Core attribute
- Cross-dataset where → CommentOID required

How the argument runs

1	**The gap**	Rules check presence, not correctness
2	**What Origin is**	Type, Source, and what each Type owes
3	**Predecessor**	No Source — a reference as free text
4	**Two origins**	Value-level metadata, not two elements
5	**The judgment**	How to choose, and what a pipeline cannot see

A tool settles

- An Origin is there
- Type and Source are controlled terms

Left to you

- Whether the Type is right — five narrow rules aside
- Whether the reference resolves

You can only check what you could have written yourself.

Thank You

Miae Lee · ma.lee@jnpmedi.com

2026 Korea CDISC Day · 29 October 2026

