2026 Korea CDISC Day

29 October 2026 · Seoul, Korea

AI-Assisted CDISC SDTM Generation

Gihwan Kim, CEO, Clupea

29 October 2026



Speaker



Gihwan Kim

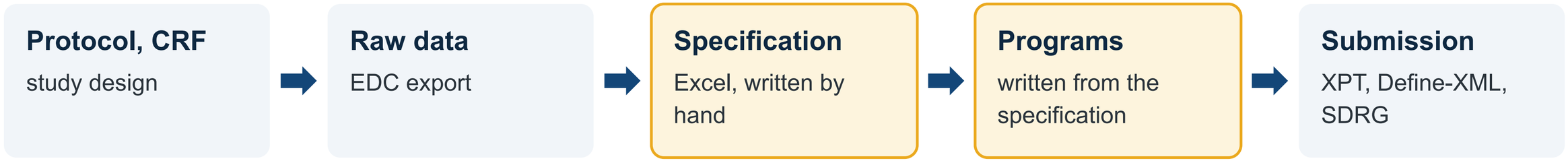
CEO and Co-founder
Clupea Inc. · MediLake

Founder of Clupea, the company behind MediLake, an integrated CDMS, EDC, CTMS and eTMF platform used by CROs, hospitals and pharmaceutical companies in Korea. Clupea also delivers SDTM and ADaM production for sponsors. Involved in CDISC advisory activities since 2017 and founder of the first CDISC Korea Summit. Computer science, Dongguk University; graduate study in physiology at the same university.

Agenda

1. Where AI belongs in SDTM generation
2. Keeping it compliant and reproducible
3. Evidence from two delivered studies
4. Lessons, limits and next steps

SDTM generation is still a manual chain



The specification is prose

People read it; programs do not. The specification and the code can drift apart.

Knowledge lives in people

Conventions, visit logic and recoding decisions are worked out again for each study.

Validation comes last

Findings arrive after the programming is finished.

Why not just ask a language model?

Reproducibility

The same input must give the same datasets on every run. Generation does not guarantee that.

Traceability

21 CFR Part 11 and Annex 11 ask who decided what, when and why. A model output is not a decision.

Data protection

Subject-level clinical data cannot be sent to an external service.

Terminology

A term that sounds right but is not in the codelist is a validation finding.

Principle

1. **AI proposes** — Mappings, terms and translations, as suggestions only
2. **Rules check** — Every suggestion, against the standard pinned for the study
3. **A person decides** — Accept, modify or reject, with an electronic signature
4. **The engine executes** — Deterministically, with no model call at run time

One pipeline, one audit trail

Standards

SDTMIG 3.4
CT 2026-09-25
MedDRA

Study inputs

raw data, protocol, codebook

Mapping

AI proposes
rules check
reviewer signs

the only place a model is used

Transform

deterministic engine
same input, same output

Outputs

XPT v5
Dataset-JSON 1.1
Define-XML 2.1
aCRF, SDRG and ADRG drafts

Validation

CDISC CORE (Open Rules)
each finding gets a disposition;
submission checks for FDA, PMDA, EMA, MFDS

Audit trail · hash-chained events · electronic signatures · standards and AI calls sealed with SHA-256

Everything to the right of the mapping step is deterministic, and ends in an eCTD module 5 package. Everything is recorded underneath.

Seven narrow tasks, each with its own check

The model proposes	Example	Check that needs no model
Trial design from the protocol	arms, visits, epochs	EPOCH term exists in the codelist
Trial summary (TS)	TPHASE = PHASE I TRIAL	parameter and value are in CT
Dataset to domain	lb_central → LB	domain exists in SDTMIG
Column to variable	AESTDAT → AESTDTC	variable is in the IG or the model class
Value to controlled term	"Once daily" → QD	term is in the codelist; extensible or not
Test to test code	Fasting Plasma Glucose → GLUC	TESTCD and TEST share one NCI code
Specification cell to rule	free text → notation	rule parses; items exist in the raw data

Text that is not English is recoded the same way: proposed, checked, signed.

The model selects; it does not invent

1. **Candidates come from the pinned standard**
   Domains, variables and terms are retrieved from the study's snapshot.

2. **The answer is a structured choice**
   The model picks from the candidates. Free-form text never enters the mapping.

3. **Every proposal is checked again**
   Against SDTMIG, the SDTM model and the CT package, by code.

4. **The result decides what a reviewer can do**
   A proposal without grounding cannot be accepted as it is.

GROUNDED

Found in the standard. The reviewer may accept.

PARTLY GROUNDED

For example, the codelist was retired in the pinned CT. Needs a closer look.

UNGROUNDED

Blocked until a person corrects or rejects it.

Grounded is not the same as correct

RAW VALUE
"Once daily"

PROPOSED
~~BID~~

REVIEWER: MODIFIED
QD

Both terms are in the codelist. Only one is true.

RAW VALUE
injection site: abdomen

PROPOSED
~~ABDOMINAL CAVITY~~

REVIEWER: MODIFIED
ABDOMINAL WALL

A subcutaneous injection does not go into the cavity.

RAW COLUMN
RANDDAT

PROPOSED
~~DMDTC~~

REVIEWER: REJECTED
DS milestone

Randomization is a disposition event, not a demographics date.

Synthetic demo study with a rule-based proposer: 118 proposals, 113 fully grounded. The reviewer accepted 115, modified 2 and rejected 1.

What the model is allowed to see

Goes into a prompt

- Names, labels and inferred types of columns
- Candidate lists from the pinned standard
- Distinct values to recode, without the subject
- Protocol and specification text

Never goes into a prompt

- Subject-level records
- Subject identifiers
- Dates of birth

On-premises model only

The endpoint must resolve to a loopback or private address. Anything else is refused.

Every call is logged

The full input and the raw output are kept and sealed with SHA-256.

Local console

The web console answers only on the machine it runs on.

Built for 21 CFR Part 11 and Annex 11

Hash-chained audit trail

Imports, proposals, decisions, runs and exports are events in one chain. Changing a past event breaks it.

Electronic signatures

The signer re-enters the password and states the meaning and the reason. The signature is bound to the signed content.

Version-pinned standards

IG, CT and dictionary versions are fixed for each study and verified by hash when they are loaded.

Lineage for every variable

Source item, transformation, and who wrote the rule: an AI proposal, a person, or the system.

Three more rules: an AI agent’s decision is recorded as an agent decision, never as a signature; the installation produces its own IQ, OQ and PQ evidence; and IQ verifies the signed release file by file.

CDISC Open Rules on every run

CORE 0.17.1 runs on the generated XPT files as part of the pipeline.

A count of findings is not the result. **A disposition for each finding is.**

Each rule is routed to one of three places: the SDRG, a sponsor query, or the reviewer.

STUDY B · VERIFICATION RUN

3,089 findings on the first run → **2,064** after AE terms were coded with MedDRA



about 1,783
explained in the SDRG, including known false positives

about 248
sent to the sponsor as data queries

33
left for the reviewer to check

A benchmark on delivered studies

1. **Take a study that was already delivered**
   Its SDTM was produced by hand and handed over to the sponsor.

2. **Regenerate SDTM from the same raw data**
   Nothing is copied from the delivered datasets.

3. **Align records on natural keys**
   Subject, topic and timing variables. Never --SEQ, which depends on sort order.

4. **Compare every cell and classify each difference**
   Blank, case, whitespace, number format, date precision, or value.

STUDY A

Phase 1, crossover, 29 subjects

31 delivered datasets
17,869 records

STUDY B

Phase 1, oncology

47 delivered datasets
70,965 records

Benchmark reports hold counts only: no subject identifiers and no data values.

The specification makes the difference

Generic rules only

No study-specific mapping

7 of 31 delivered datasets reproduced

about 1% delivered records matched

47.6% cells identical on matched records

Delivered specification, executed

The team's own Excel workbook, run as the mapping

24 of 31 delivered datasets reproduced

67% delivered records matched; 99.2% within those 24 datasets

97.2% cells in agreement: 87.2% identical, the rest differ in format only

Same engine, same raw data. The only change is an explicit mapping.

Free-form specifications are harder

Route	Datasets	Records matched	Cells identical	Incl. format differences
Specification executed as written	31 of 47	14%	80.3%	80.7%
Console workflow: on-premises model and reviewer	26 of 47	43%	86.3%	90.5%

Not there yet

With a reviewer in the loop, fewer than half of the delivered records are reproduced today. Where records do match, nine cells in ten agree.

153

specification cells could not be read as rules

In Study A it was 18.

What the tool did not have

Missing	Consequence
Arm allocation	Kept outside the EDC export, so SE and EX were not built
External laboratory files	PK concentrations arrive separately, so PC was not built
Rules written as free text	18 cells in Study A and 153 in Study B could not be executed
Text that is not ASCII	XPT v5 blocked four Study A datasets until values are recoded
Code not written yet	CO and RELREC are reported, not converted

Most of this is missing input, not engine error

The bottleneck is the specification

The engine agrees with the delivered data on 97% of cells when the mapping is explicit. Prose has to be interpreted first, by a person or by a model.

Write the rule once, for people and the engine

Written for people	Written for people and the engine
Male if 1, female if 2	`"M" if SEX="1", "F" if SEX="2"`
First dosing date and time	`first(EXDAT)+"T"+first(EXTIME)`
Study ID, hyphen, subject number	`SDTM.STUDYID+"-"+SUBJID`
Measurement date at the screening visit	`VSDAT (VISIT="10")`

The same Excel workbook

One sheet per domain, lookup sheets for visits, tests and trial design.

A lint before delivery

Zero open cells means the workbook runs exactly as written.

Nothing is guessed

A rule that cannot be read is reported, and its variable stays empty.

The model translates, the parser judges

Unreadable cell
free text in an existing workbook

→

Model proposes
from the variable, its label, the cell text and raw item names

→

Parser checks
Does the rule parse? Does every item exist in the raw data?

→

Reviewer signs
accept, edit or reject; only checked proposals can be accepted

CELL TEXT	PROPOSED RULE	CHECK
First dosing date	`first(EXDAT)`	**Passed: parses, EXDAT exists**

Status: built and tested end to end with a stub model. Acceptance on Study B's 153 open cells with the on-premises model: **measurement in progress**

What this does not show yet

1	**Two studies**	Both are Phase 1, with specifications from one organization. Other therapeutic areas and writing styles are untested.
2	**Agreement, not correctness**	The delivered data is a reference, not ground truth. Differences have not yet been classified one by one with a data manager.
3	**Review still in progress**	Study B was measured before every review decision had been signed.
4	**Model steps not yet measured**	Acceptance rate and review time for specification translation are pending.
5	**ADaM**	Datasets, ARM and Dataset-JSON are generated, but CORE has no ADaM rules yet: a local validator and Pinnacle 21 remain the check.

From protocol to submission

PLANNED

USDM and ICH M11

Read a digital protocol to pre-fill trial design and TS, instead of parsing a document.

IN PREVIEW

SDTMIG v4.0

A gap report against the public-review draft and a SUPP-- to NS-- conversion, behind a per-study flag until v4.0 is published.

TODAY

Open Rules

CORE on every SDTM build. Each quarterly CT release is impact-analysed per study, applied with a reason, and re-validated.

TODAY

ADaM and Dataset-JSON

ADaM with analysis results metadata in define.xml; Dataset-JSON 1.1; eCTD v4.0 study metadata for FDA and PMDA.

The other talks in this session describe the next inputs to this pipeline.

Three things to take home

1

Put AI where a person can check it

It proposes. The engine that builds the datasets never calls a model.

2

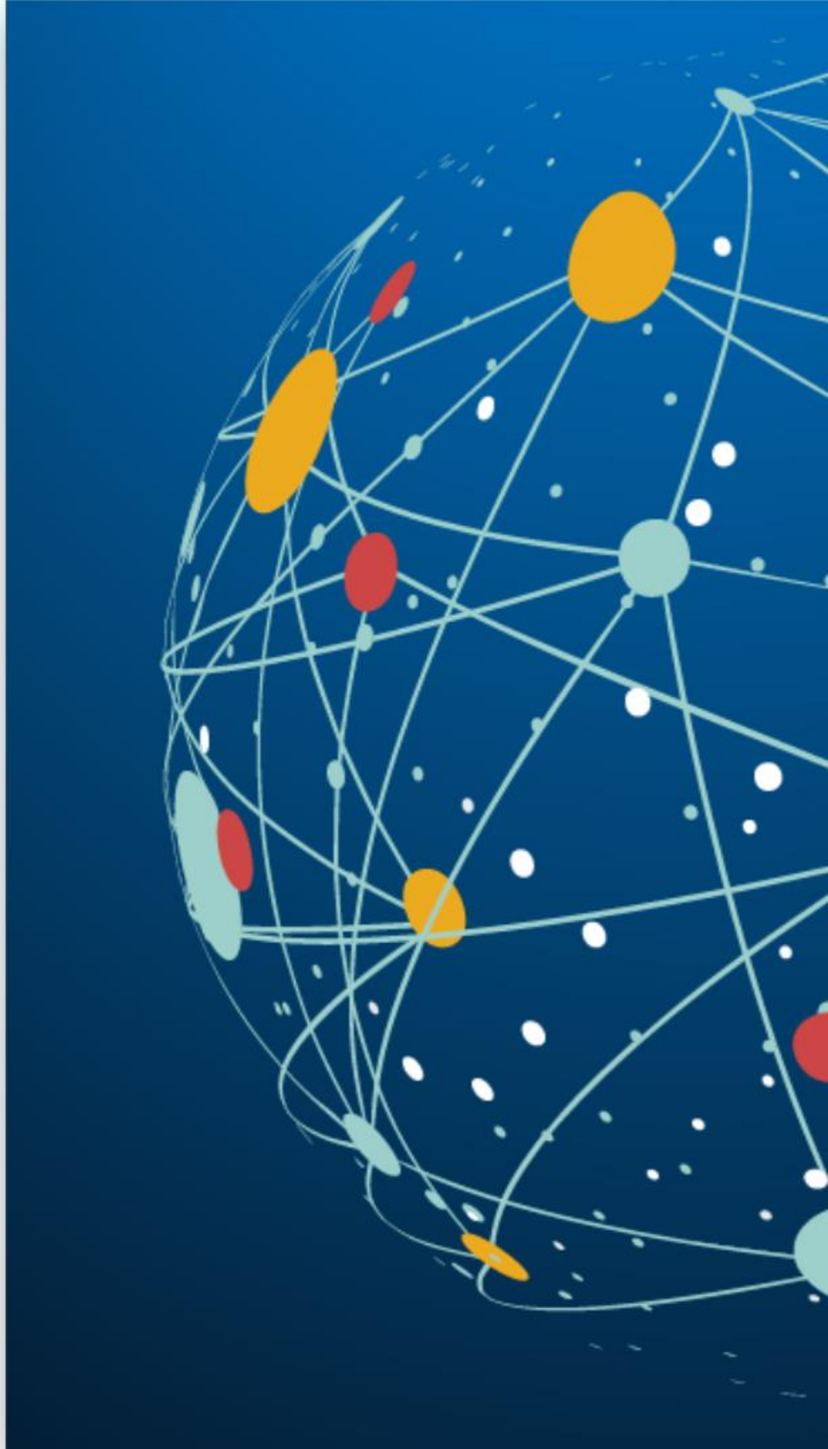
Ground every proposal, and keep the record

Checked against the pinned standard, run on-premises, logged with its full input.

3

Invest in the specification

Explicit rules reproduced 97% of cells in a delivered study. Free text reproduced far less.



Thank You

Gihwan Kim · Clupea · kimgihwan@clupea.co.kr
2026 Korea CDISC Day · 29 October 2026



Questions?

