



CDISC India Day

Keynote Presentation

Patti Compton, Pfizer Inc., CDISC Board of Directors





Meet the Speaker

Patti Compton

Title: Vice President and Global Head for Biometrics and Data Management, Statistical Data Sciences, and Analytics organization

Organization: Pfizer

With more than 25 years of experience across the pharmaceutical and academic sectors, Patti is recognized for advancing the value of data through responsible data sharing, analytics, and operational excellence. She leads initiatives focused on modern data delivery, automation, AI, real-world evidence, and global capability development across Asia. A strong advocate for collaboration and innovation. Patti serves on multiple industry advisory groups, including the boards of CDISC and the Society for Women's Health Research.

CDISC INDIA DAY KEYNOTE

From Standards to World Models

Art of Possibility

August 2026

CDISC · Domain knowledge · World models

Art of Possibility

- 01 The paradox**

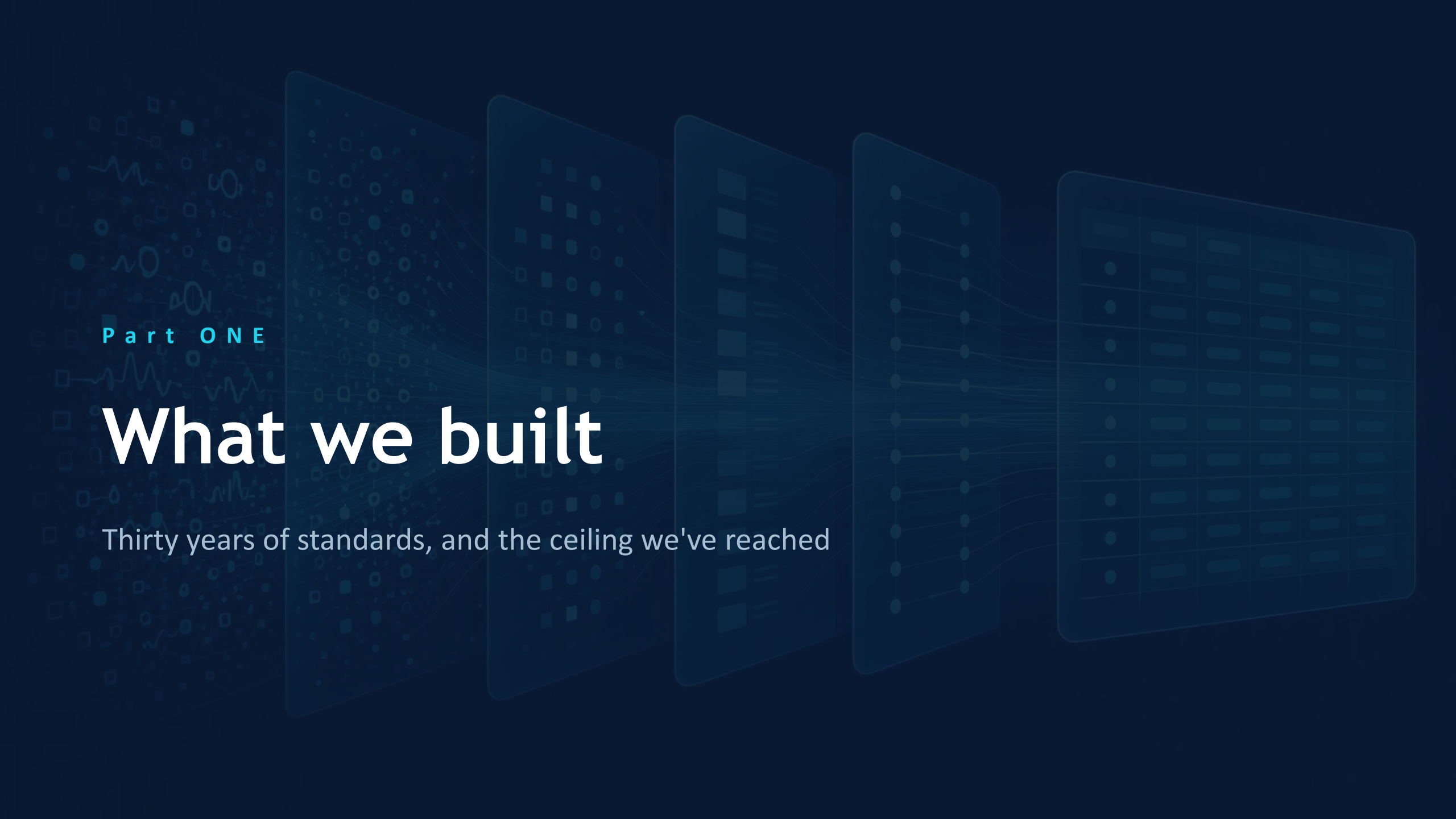
The most standardized data in science is still hard for machines to reason about.
- 02 What CDISC gives us**

Structure, terminology, traceability — and where the model stops.
- 03 Domain knowledge**

The expertise living in reviewers' heads, not in the datasets.
- 04 World models**

What they are, and why clinical research is a strong candidate domain.
- 05 The roadmap**

Five horizons from metadata to grounded reasoning.



Part ONE

What we built

Thirty years of standards, and the ceiling we've reached

The paradox at the heart of our field

We are the best-standardized data domain in science — and among the hardest to automate.

What we have

- Common structures across sponsors and studies
- Controlled terminology for most collected concepts
- Machine-readable metadata in Define-XML
- Decades of curated, regulated, auditable records

What we still don't have

- Explicit meaning — why a variable exists, what it implies
- Encoded causal and temporal relationships
- Rules and conventions that live in people's heads
- A representation a model can reason over, not just read

What the CDISC stack actually gives us

Four kinds of agreement, each solving a different problem.

COLLECTION

CDASH

Agreement on how a question is asked at the site.

TABULATION

SDTM

Agreement on how observations are organized for review.

ANALYSIS

ADaM

Agreement on how analysis-ready data is derived and traced.

MEANING

Define-XML + CT + 360

Machine-readable metadata and controlled vocabulary.

Each layer is an agreement about form. Only the last one begins to be an agreement about meaning.

Traceability is the asset we undersell

The chain from protocol to result is one of the few end-to-end provenance systems in applied science.



Why this matters for AI

A model trained on clinical data can, uniquely, be asked to show its work — every derived value points back to a documented rule and a source observation. Almost no other domain can audit its own training data this way.

Where the model stops

Standards describe the shape of the data. Four things they were never designed to carry.

Intent

Why this variable was collected, and what question it answers.

Causality

Which observations plausibly explain which others.

Time

Ordering, windows, and what counts as 'before' in a clinical sense.

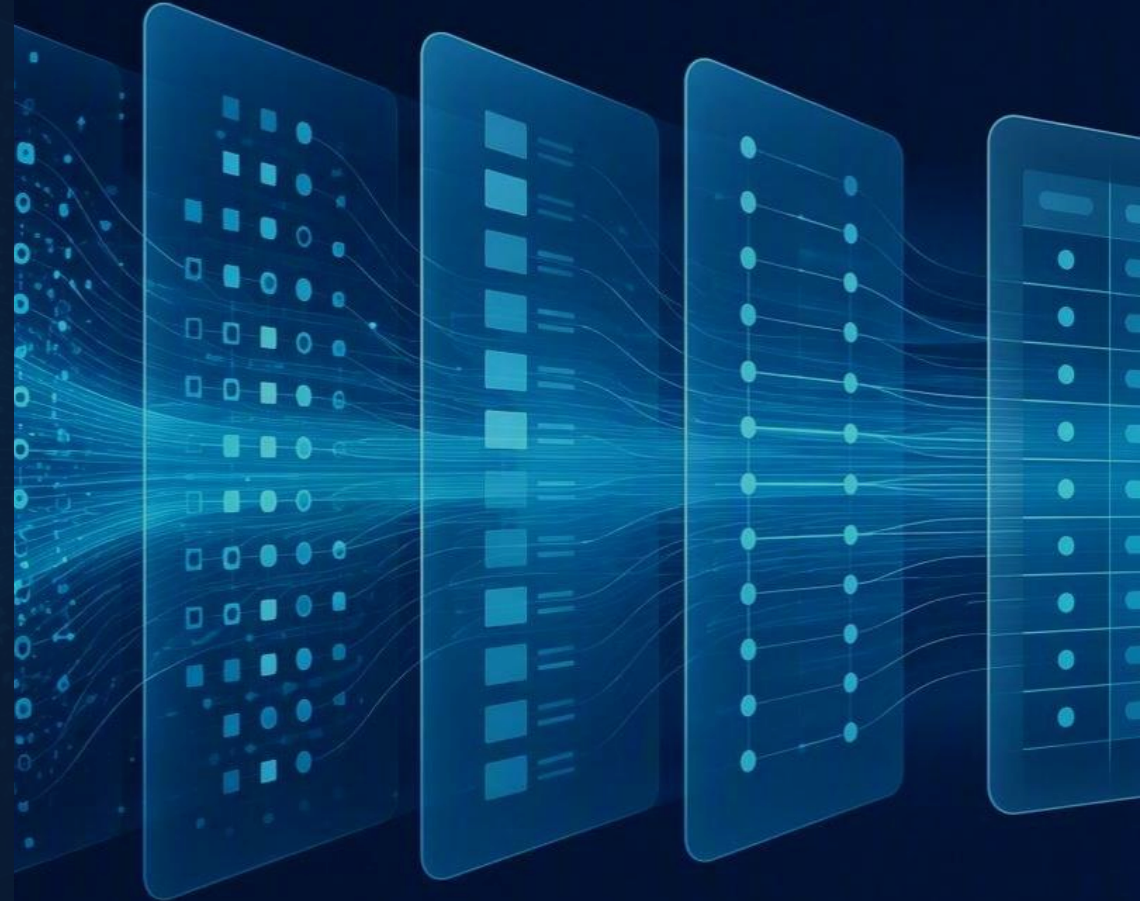
Constraint

What combinations of values are impossible, not merely absent.

Part TWO

What we know

Domain knowledge, and why it never made it into the file



What a reviewer knows that the dataset doesn't say

The dataset states

- A term, a date, a severity grade
- A flag indicating seriousness
- A treatment assignment
- A set of concomitant medications

The reviewer infers

- Whether the timing is plausible for this mechanism
- Whether the pattern resembles a known class effect
- Which concomitant medication is the likelier explanation
- Whether this looks like a data problem or a safety signal

Domain knowledge comes in three layers

They fail differently, and they have to be captured differently.

Structural

How our data is shaped

Which domain holds what, how keys join, what a record means. Largely already encoded — this is the CDISC layer.

Clinical

How biology behaves

Mechanism, plausible timing, expected ranges, what co-occurs. Lives in medical monitors and experienced reviewers.

Regulatory

How judgment is made

What counts as evidence, what a reviewer will question, precedent from prior submissions. Almost entirely tacit.

Tacit knowledge is expensive in hidden ways

Re-derived

Every study rebuilds the same conventions from scratch.

Unverifiable

If the rule isn't written down, it can't be checked or tested.

Perishable

It leaves when the person leaves.

The reframe

Every convention you explain twice is a candidate for encoding. Knowledge capture isn't a documentation project — it's the training data problem for our domain.

The opening: metadata is knowledge

Three assets we treat as deliverables that are actually machine-readable knowledge.

Controlled terminology

Shared concepts with stable identifiers — the closest thing we have to a vocabulary a machine can rely on.

Define-XML

Derivation logic, origins, and value-level detail, already expressed in a structured, parseable form.

Biomedical concepts

Concept-level definitions that tie a clinical idea to its representation across collection, tabulation, and analysis.

Treated as submission artifacts, these are overhead. Treated as a knowledge base, they are a head start.

Moving toward Biomedical Concepts



Biomedical concepts serve as the essential semantic framework that connects and aligns various standards by linking*

Common language

Biomedical concepts provide a common language for representing critical clinical data, ensuring consistence, accuracy and interoperability across systems.

Linking key standards

Biomedical concepts enable data flow and support regulatory compliance.

* Smiley, J. & LeRoy, B, (2026, March 17) *How AI and Standards Work Together: Insights from the CDISC AI Innovation Biomedical Concepts Acceleration Challenge* [Webinar]. [cdisc How AI and Standards Work Together: Insights from the CDISC AI Innovation Biomedical Concepts Acceleration Challenge | CDISC](#)

From documents about data to models of data

FROM

- Conventions in a PDF specification
- Rules enforced by review
- Knowledge that lives with individuals
- Standards checked at the end

TO

- Conventions expressed as machine-readable rules
- Rules enforced by execution
- Knowledge held in a shared, versioned model
- Standards asserted at the point of authoring



Part THREE

What we could build

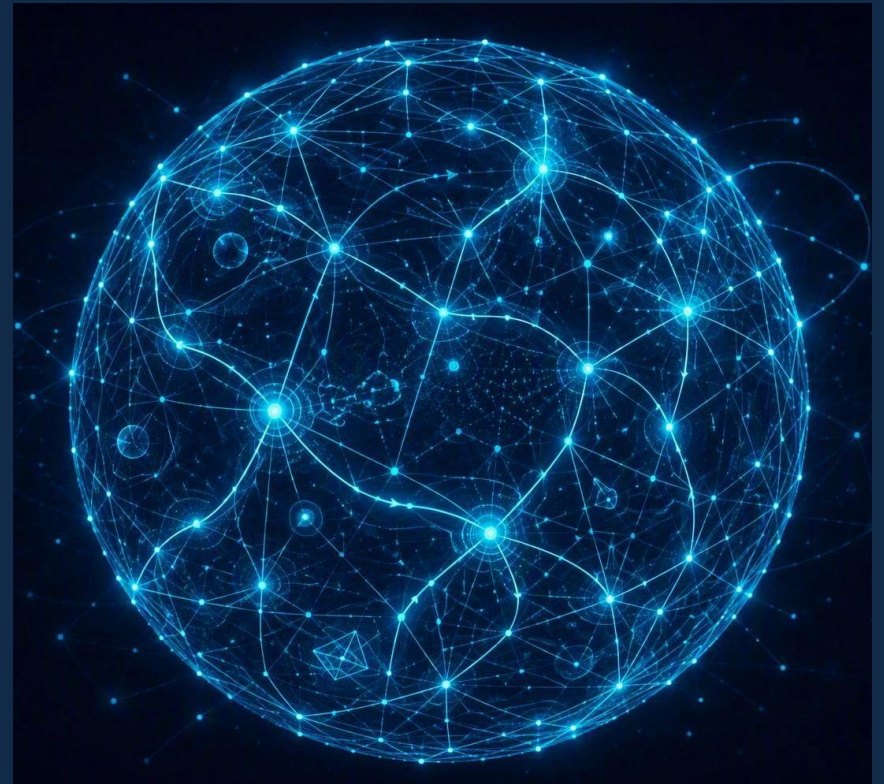
World models, and why this domain is a strong candidate

What a world model is

An internal representation of how a domain works — its entities, the rules they obey, and how a state changes when something acts on it.

The practical test

- It can predict what happens next, not just describe what happened
- It answers questions it was never explicitly shown
- It knows what is impossible, not only what is unobserved



Pattern matching is not world modeling

	Pattern-matching system	World model
Learns from	Correlations in text and tables	Structure, constraints, and dynamics
Answers	What usually follows what	What would happen if
Fails by	Producing fluent, plausible wrong answers	Refusing or flagging when state is inconsistent
Needs	Volume	Grounding

Why clinical research is a strong candidate

Most domains lack at least one of these. We have all four.

Bounded entities

A finite, agreed set of things: subjects, visits, exposures, events, results.

Explicit rules

Protocols, edit checks, and derivations already state what is allowed.

Native time

Everything we record is anchored to a study day and a reference start.

Auditable ground truth

Traceability lets us check a model's claim against a documented source.

What a clinical world model must know

- | | | |
|---|--------------------|---|
| 1 | Entities | Subject, arm, exposure, event, assessment — and how each relates to the others. |
| 2 | States | Baseline, on-treatment, follow-up, and what a transition between them implies. |
| 3 | Dynamics | What plausibly changes a state, over what interval, and in which direction. |
| 4 | Constraints | Combinations that cannot occur — impossible ranges, sequences, and overlaps. |
| 5 | Provenance | Where each assertion came from and how confident we should be in it. |

Note how much of rows one and five we can already produce from metadata we generate today.

What ungrounded models get wrong here

Illustrative failure modes — the reason domain grounding isn't optional.

Confident derivation

Produces an analysis variable that looks correct and follows no documented rule.

Time blindness

Treats an event before first dose as treatment-emergent because the text reads that way.

Terminology drift

Silently maps a term to a similar-sounding concept from a different codelist.

Unsourced certainty

States a relationship with no traceable basis, phrased exactly like a sourced one.

Spectrum of AI Agency



Key Insight: Most current pharma AI solutions operates at 1st–2nd order agency

THE ROADMAP

Five horizons to AI-ready clinical data

H1

Clean metadata

H2

Executable rules

H3

Linked concepts

H4

Simulated studies

H5

Grounded answers

Horizons 1 and 2: In progress

H1

Clean, complete metadata

- Generate define metadata from the same source that drives programming
- Retire study-specific variants of standard structures
- Treat controlled terminology as a governed asset, not a per-study lookup

H2

Executable rules

- Express conventions as checks that run, not prose that is reviewed
- Version and test rules the way you version and test code
- Capture the rationale beside the rule, not in a separate memo

Horizons 3 to 5: where it becomes research

H3

Linked concepts

Connect study variables to concept-level definitions and biomedical vocabularies, so meaning travels with the data across studies.

H4

Simulated studies

Use the encoded structure and constraints to generate plausible synthetic study data — and use it to stress-test analyses and pipelines.

H5

Grounded reasoning

Systems that answer clinical data questions by reasoning over the model, and cite the rule and source behind every claim.

Horizons 4 and 5 are open research. Say so, and fund them as research.

What happens

Data managers

Stop resolving the same convention twice. Write it down as a rule the moment you explain it a second time.

Statistical programmers

Treat derivations as versioned, testable logic with the reasoning attached, not as study-local code.

Standards leads

Govern terminology and concept definitions as a shared asset with an owner, a version, and a change process.

Leaders

Fund metadata quality as infrastructure. It is the prerequisite for everything else on the roadmap.

What to watch for

Three things worth watching, stated plainly.

Scale may surprise us

General models keep absorbing capabilities we assumed needed explicit structure. Some of this roadmap may become unnecessary.

Encoding is harder than it sounds

Much clinical judgment resists formalization. Writing it down can flatten nuance that mattered.

Governance is the bottleneck

The technical work is tractable. Agreeing across sponsors, vendors, and regulators is the slow part.

The takeaway

**We didn't spend thirty years writing rules.
We spent thirty years writing the first draft
of a model...**

Encode what you explain twice.

Thank You!

Acknowledgements:

CDISC: Julie Smiley, Bess LeRoy

Pfizer: Jeba Kumar, Priya Iyer, Subha Madhavan