2026 CDISC China Interchange

# AI-Enabled Synthetic Data Generation for Clinical Trials

Yuting Ji / Platform Business Capability Lead
MSD R&D (China) Co., Ltd
14-Aug-2026

# Agenda

1. Current Situation
2. Opportunity with GenAI
3. Synthetic Participant Data
4. Validation principles

# Speakers

MS of Biomedical Engineering from University of California, Irvine.

10 years of industry experience, PMP certified.

Currently serving as Associate Director with Platform Business Capability Lead role in Data Management and Data Integration fields at MSD, mainly responsible for managing and driving implementation of business systems.

Has solid experience in CSV and end-to-end system tests design and support for various clinical systems such as CDMDB, EDC, IRT, as well as QA/QC process to ensure quality compliance and maintaining the integrity of the system lifecycle/SDLC.


## Yuting Ji

Platform Business Capability Lead, Associate Director

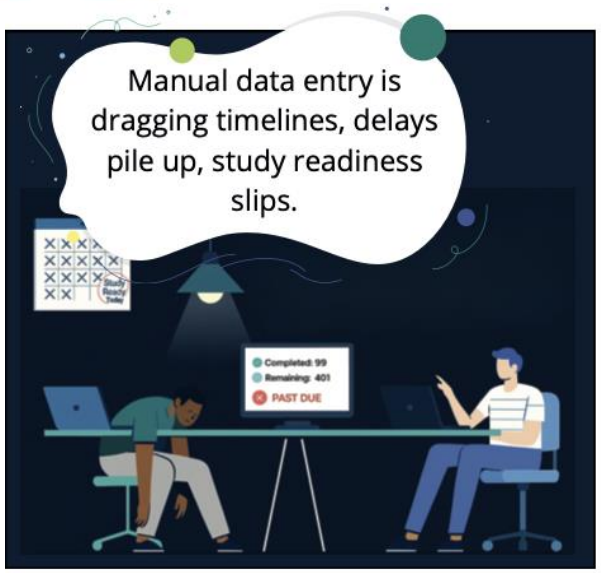
MSD R&D (China) Co., Ltd

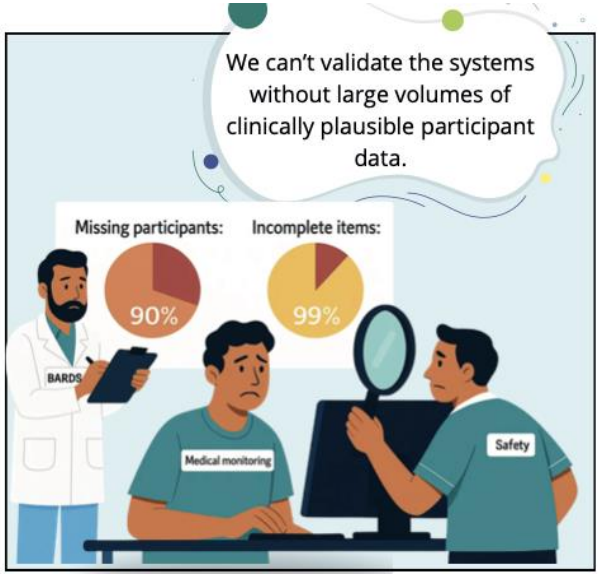
# The Disclaimer

The content presented in this presentation is derived from thorough literature review and personal expertise. It is imperative to acknowledge that the perspectives, opinions, and discoveries expressed herein are solely attributable to the presenter and may not align with the official stance or perspectives of the Company. The information shared is intended for informational purposes exclusively and should not be construed as professional or legal counsel. Any decisions made based on the content of this presentation are at the discretion and accountability of the individual.

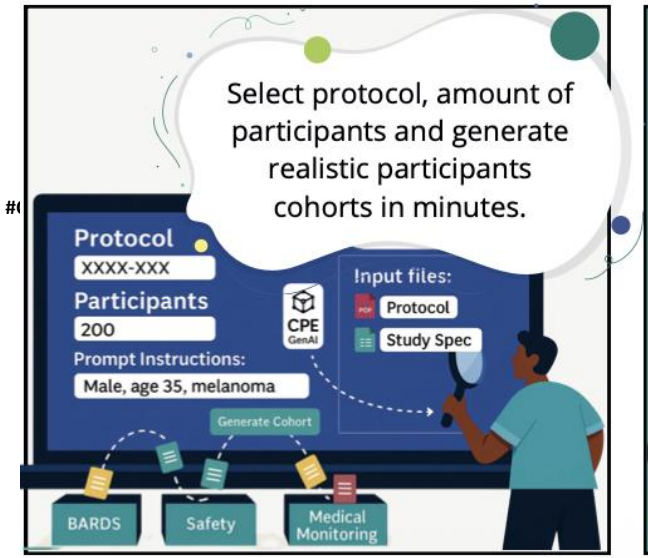
# Current Situation

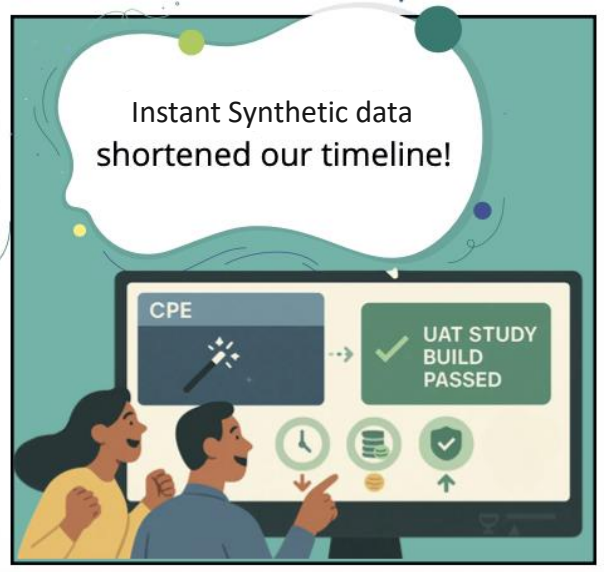
# Clinical Trials



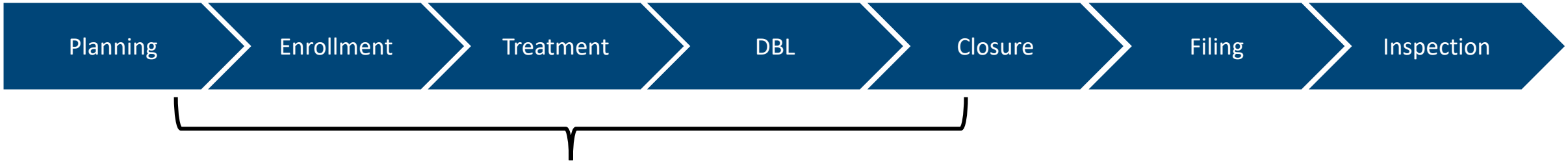





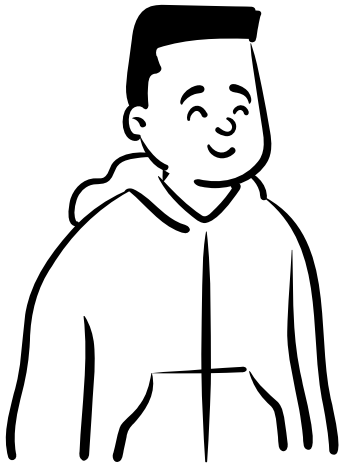



# The Problem




| | |
|---|---|
| **Pain Point 1** | **Manual Efforts**<br>Clinical trial system build and change requests rely heavily on manual data entry. |
| **Pain Point 2** | **Timeline**<br>Current clinical trial activities require meaningful timeline compression. |
| **Pain Point 3** | **Team Capacity and Data Quality**<br>Labor-intensive nature of manual data entry, straining capacity and risking data quality. |

# Study developer

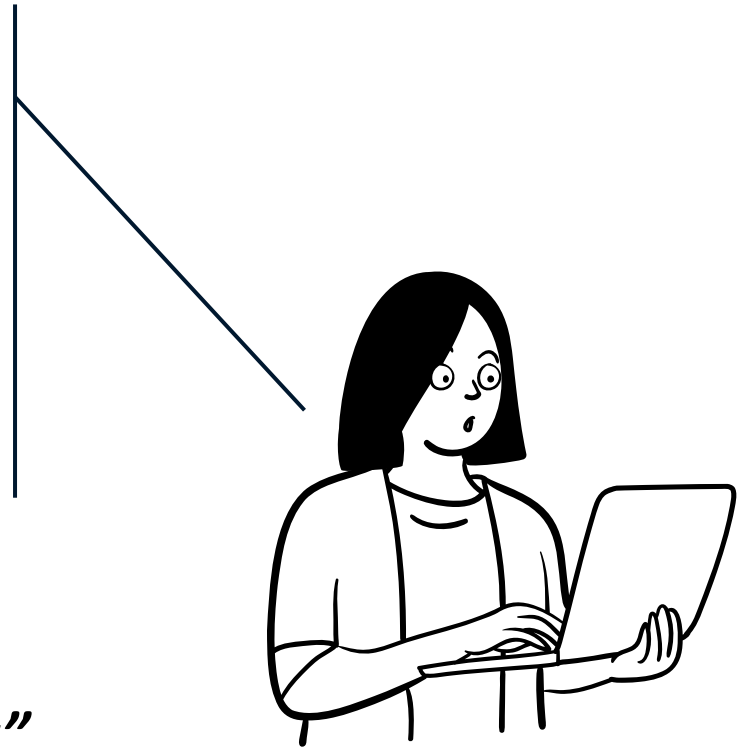


- I need large volumes of EDC data during:
  1) Study build 2) Form/rule changes
- Sometimes changes invalidate previously entered test data
- I want fast feedback during development cycles

***"I am building or changing a study"***

# UAT Lead / Tester

- I want controlled, traceable datasets for UAT
- I need full rule coverage, including edge cases
- Need external data in lower environment
- Repeatable scenarios for re-testing



***"My testing is blocked by data"***

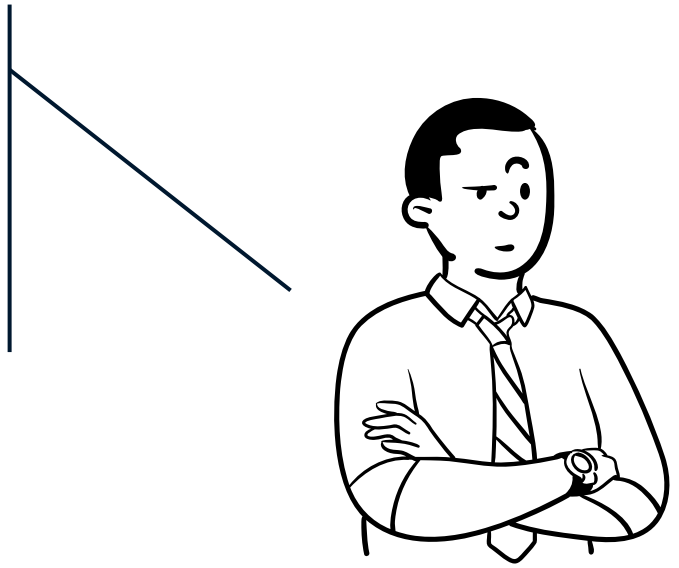
# Clinical Data Manager

- I am asked to enter realistic test data
- I want early detection of data issues
- eCRF fields require precise control or adjustment

***"Test data ≠ manual work by default"***

# Statistical Programmer

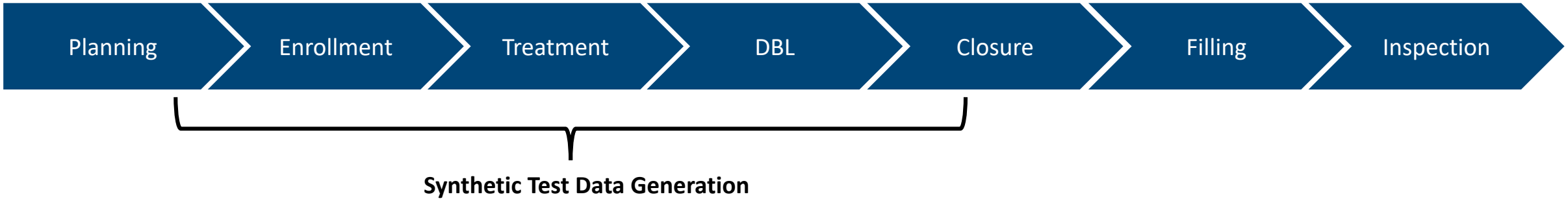
- Analysis depends on late availability of data
- I need large, structured datasets for pipeline and integration testing
- I need consistent data structures across scenarios



***"Code can't run without data"***

# Opportunity with GenAI

# GenAI Solution




Generate clinical plausible data, with programing pre-calculation and LLM

**Synthetic Data**

**GenAI powered**

Utilizing GenAI to fulfill clinical data entry needs

Provide users with comprehensive guidelines and best practices for crafting effective prompts.

**Intuitive User Interface**

**Easy Data Load**

Connect synthetic data with the EDC application through API integration.

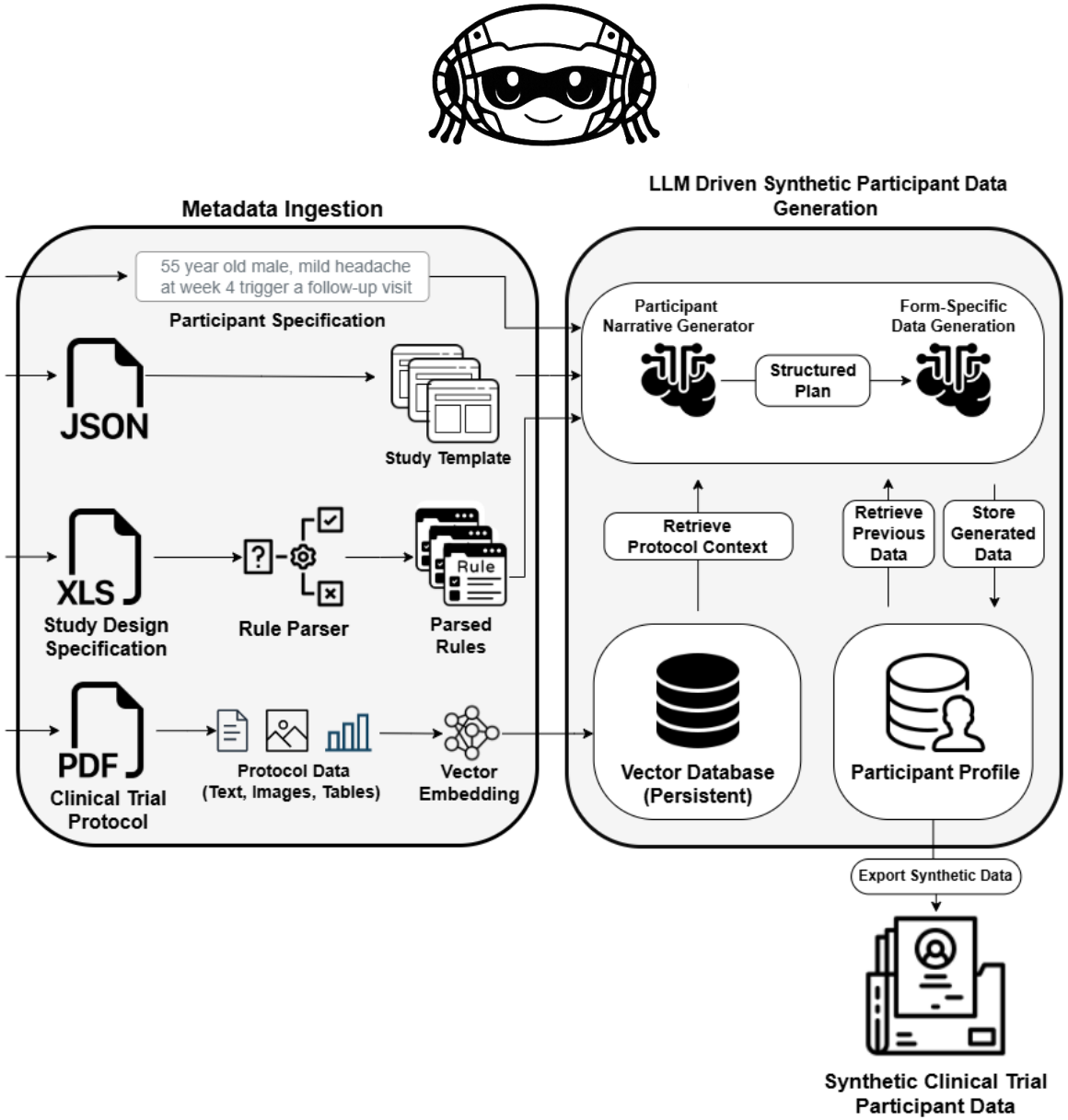
# How we did it?

### Overview

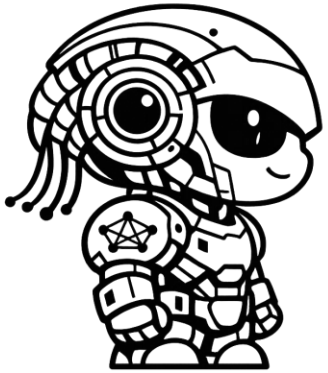
LLM driven system that aims to generate **clinically plausible** structured data for EDC

### Technologies used

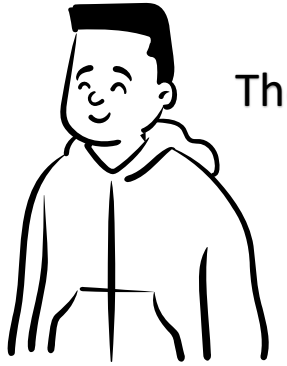
- **Platform**: AWS
- **Frontend**: React
- **Backend:** Python
- **Containerization**: Docker
- **Application code repository** : GitHub – system for data version control – maintaining the system codebase.



# What can it offer?



- Generate synthetic participant data directly from: Casebook design and rule definitions
- Rapid re-generation after study changes
- Shortens build–test–fix cycles

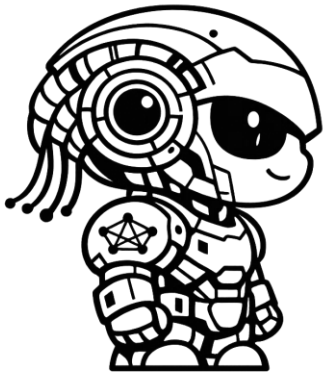


The developer

- Scenario-driven participant specifications
- Rule-aware data generation
- Supports risk-based validation by: 1) Expanding coverage 2) Reducing manual effort

The tester

# What can it offer?



- Clinically plausible synthetic data
- AI generation + manual edit combined
- Reduces manual effort while maintaining realism

The CDM

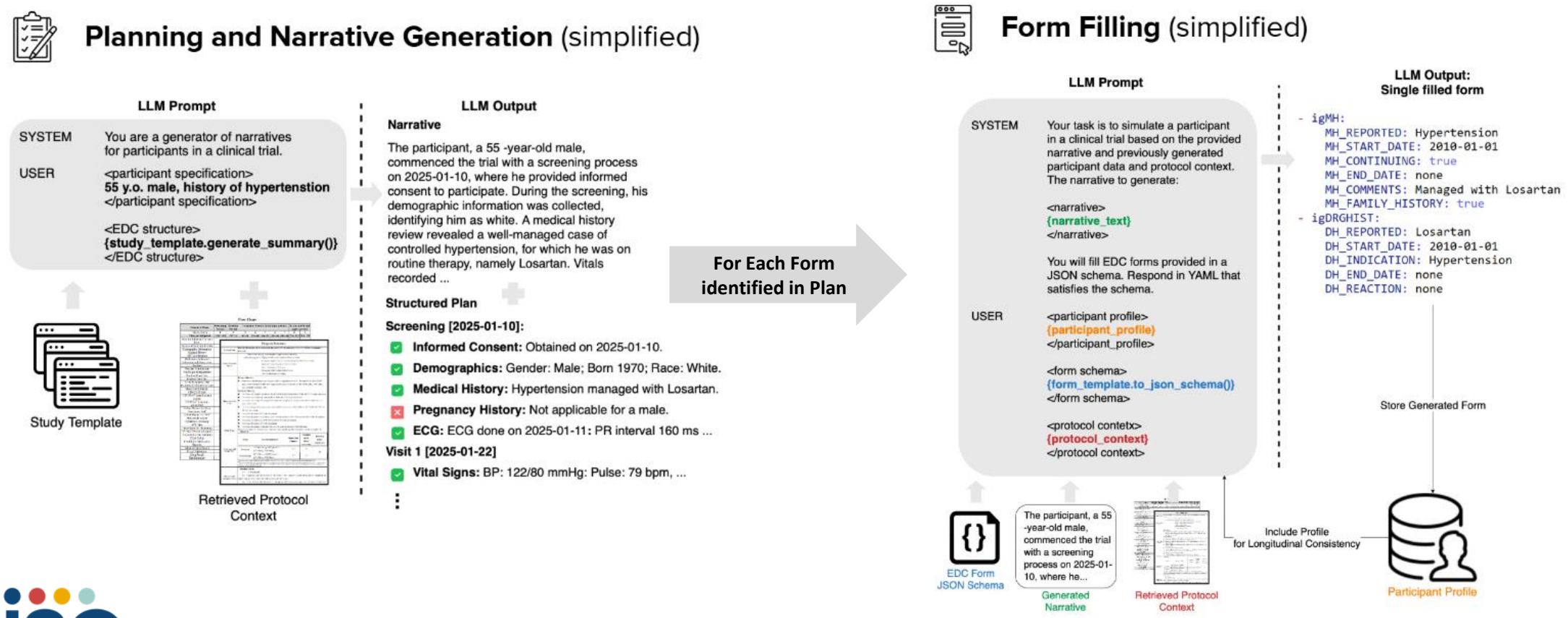
- Standardized, analysis-ready synthetic datasets
- Enables early analytics and pipeline validation
- Improves end-to-end readiness

The programmer

# Synthetic Participant Data

# Mechanism

- **Two stage loop -** A *Planner* first converts study metadata into a timeline of events, a *Filler* then populates each form. Both loop through a validator until every rule passes.
- **"Agentic Design" -** Each stage is a LLM with its own prompt, inspired by agent frameworks for complex reasoning
- **Internal Control Flows -** Structured JSON messages pass plans, form schemas, and validation errors



# Sample Data

**User Request:**
Participant enrolled in cohort A.
Completed the study successfully.

## Study Template & Protocol Context

Table 2 Schedule of Activities: Screening and Rescreening – Cohor

| Study Period | Screening | Rescreening[a] (optional) | Notes: Full screening tests performed within 28 days before randomization (unless stated otherwise) are not required to be repeated. |
|---|---|---|---|
| **Maximum Days** | **28[b]** | **28[b]** | |
| **Administrative and General Procedures** | | | |
| Informed Consent | X | | Obtain before performing any protocol-specific procedures. |
| Inclusion/Exclusion Criteria | X | X | |
| Participant Identification Card | X | X | Provide to participant at time of informed consent. Participant retains card distributed during initial screening. |
| Demographics and Medical History | X | X | |
| Cancer and cancer treatment history | X | X | Collect prior bevacizumab and PARPi treatment along with other cancer treatment history. |
| FIGO Stage | X | | Review staging of ovarian cancer from initial diagnosis. See Appendix 8. |
| Menopausal Status | X | | See Appendix 5 for definition of menopause. |
| Disease-specific Biomarker Data Collection | X | | Record locally available disease-specific biomarker data. |
| History of Blood Transfusions | X | X | Include history of blood transfusions within previous 28 days before the first dose of study intervention and the reasons (eg, bleeding, myelosuppression). |
| Prior Concomitant Medication Review | X | X | See Section 8.1.5. |
| Determine stratification factors | X | X | Complete prior to initiating IRT. See Section 6.3.2 for further information. |

## Study Journey

The participant provided written informed consent on 2025-05-03 and entered the screening period for Cohort A ( ). All screening assessments, including full medical history, menopausal status, ECOG Performance Status (grade 1), laboratory panels, disease-specific ) and locally read 12-lead ECG, were completed without clinically significant abnormalities. Because sex at birth was female, a Menopause Status (RPMPS) form was completed; the participant was post-menopausal. Breast-cancer specific history (MHHDDB), current disease details (MICDDB) and staging classification (RSSTC2) were documented from source records. Inclusion/Exclusion criteria were met and the participant was judged eligible.

On 2025-05-27, the participant was randomized via IRT to Cohort A Arm 1 and assigned open-label Randomization and treatment-trigger forms (DSRND, TRIG) confirmed allocation. No endocrine therapy physician's-choice option (ETPC) was selected.

The participant completed three 28-day treatment cycles without major protocol deviations.
- Cycle 1 Day 1 (2025-05-27): First dose of administered after pre-dose vitals, ECG, ECOG 1, urinalysis, PK sample and PD specimen were obtained.
- Cycle 1 Days 8, 15 and 22 visits were performed within window; Day 15 included additional PK and PD collections per protocol.
- Cycle 2 started on 2025-06-24 and Cycle 3 on 2025-07-22 with identical Day 1 assessments. PK and PD invoices (BSPD) were required on Cycle 2 Day 1 only, per schedule. All ECOG assessments remained grade 1. No clinically significant ECG or laboratory changes were observed. One mild, transient Grade 1 injection-site erythema was recorded and resolved without treatment.

# Sample Data – Cont.

## LLM Output – Structured Data


```
"eventgroups": [
 {
  "name": "eg_SCREENING",
  "sequence": 1,
  "events": [
   {
    "name": "ev_SCREENING",
    "date": "2025-05-10",
    "forms": [
     {
      "name": "DSIC",
      "details": "Signed informed consent obtained 2025-05-03; original filed in site records."
     },
     {
      "name": "IE",
      "details": "All inclusion/exclusion criteria reviewed; participant eligible."
     },
     {
      "name": "DM",
      "details": "Female, 57 y, White; sex at birth = Female, post-menopausal."
     },
     {
      "name": "COHORT",
      "details": "DSRND_DMCOHORT = \"BREAST\"; participant assigned to Cohort A."
     },
     {
      "name": "VS",
      "details": "Screening vitals within normal limits; BP 118/74 mmHg, HR 72 bpm, Temp 36.7 \u00b0C."
     },
     {
      "name": "EGL",
      "details": "12-lead ECG: normal sinus rhythm, QTcF 412 ms."
     },
     {
      "name": "RSECOG",
      "details": "ECOG Performance Status = 1 (restricted in strenuous activity)."
     },
     {
      "name": "LBURIN",
      "details": "Urinalysis unremarkable; negative protein/glucose."
     },
     {
      "name": "MH",
      "details": "History of stage IV HR-positive/HER2-negative breast cancer diagnosed 2023."
     },
     {
      "name": "RPMPS",
      "details": "Post-menopausal >12 months; no menses since 2019."
     },
     {
      "name": "MHHDDB",
      "details": "Documented prior localized disease and prior adjuvant therapy completed 2021."
     },
     {
      "name": "MICDDB",
      "details": "Current metastatic lesions in bone and liver per imaging 2025-04-20."
     },
     {
      "name": "RSSTC2",
      "details": "AJCC 8th edition staging: pT2 pN1 M1."
     }
    ]
   }
  ]
 },
 {
  "name": "eg_RAND",
  "sequence": 1,
  "events": [
   {
    "name": "ev_RAND",
    "date": "2025-05-27",
    "forms": [
     {
      "name": "DSRND",
      "details": "Randomized to Arm 1, Cohort A; [redacted] initiated same day."
     },
     {
      "name": "TRIG",
      "details": "Treatment and observation selection confirmed; no ETPC selected."
     }
    ]
   }
  ]
 },
 {
  "name": "eg_INTERVENTIONA",
  "sequence": 1,
  "events": [
   {
    "name": "ev_DAY_1A",
    "date": "2025-05-27",
    "forms": [
     {
      "name": "VS4",
      "details": "Pre-dose vitals stable; supine after 5 min rest."
     },
     {
      "name": "EGL",
      "details": "ECG normal; no QT prolongation."
     },
     {
      "name": "RSECOG",
      "details": "ECOG 1; unchanged from screening."
     },
     {
      "name": "LBURIN",
      "details": "Cycle 1 required urinalysis; results within normal limits."
     },
     {
      "name": "PC",
      "details": "Pre-dose PK sample drawn 30 min before infusion start."
     },
     {
      "name": "BSPD",
      "details": "PD specimen shipped to central lab; invoice number PD-C1-001."
     }
    ]
   },
   {
    "name": "ev_DAY_8A",
    "date": "2025-06-03",
    "forms": [
     {
      "name": "VISIT",
      "details": "Telephone confirmation; no new symptoms."
     }
    ]
   },
   {
    "name": "ev_DAY_15A",
    "date": "2025-06-10",
    "forms": [
     {
      "name": "VS4",
      "details": "Vitals within normal ranges."
     },
     {
      "name": "RSECOG",
      "details": "ECOG 1."
     },
     {
      "name": "PC",
      "details": "Mid-cycle PK sample collected 2 h post-dose."
     },
     {
      "name": "BSPD",
      "details": "PD specimen sent; invoice PD-C1-015."
     }
    ]
   },
   {
    "name": "ev_DAY_22A",
    "date": "2025-06-17",
    "forms": [
     {
      "name": "VISIT",
      "details": "Site visit; participant fit for next cycle."
     }
    ]
   }
  ]
 },
 {
  "name": "eg_INTERVENTIONA",
  "sequence": 2,
  "events": [
```

# Sample Data vs. CDSIC Standards

**Casebook Design Export** - JSON-based metadata representation of the study design in EDC:

| Term | Description | Data Structure Definitions |
|---|---|---|
| **Event Group** | Collection of related visits, which may repeat (e.g., Screening Period, Treatment Cycle). | (eventgroup_def): Contain the ordered schedule of the study and describe how event groups are organized along with their nested events. |
| **Event** | Scheduled visit where subject data is collected (e.g., Screening, Baseline Visit, Follow-up Visit). | (event_def): Specify individual events, including scheduling constraints, dynamic rules, and offset logic to manage timing. |
| **Form** | Structured data entry interface used at each event (e.g., Vitals, Demographics, Adverse Events). | (form_def): Describe the individual forms used for data capture, detailing aspects such as repeatability and conditional appearance. |
| **Item Group** | Set of related data fields within a form, possibly repeating (e.g., Blood Pressure Measurements, Lab Results, Concomitant Medications). | (itemgroup_def): Organize related data fields within a form into logical groups. |
| **Item** | An individual data point for a trial participant (e.g., Date of Birth, Heart Rate, Medication Dose). | (item_def): Provide metadata for each data field, such as data type, length, validation constraints, and any associated codelists. |

cdisc®

# Characteristics of sample data

- **Comprehensiveness:** The test data should cover all items within the Case Report Form (CRF). It includes data for all scheduled assessments in complete subject profiles and represents various subject scenarios.
- **Realism:** The data reflects real-world patient characteristics and disease progression. It includes logical relationships between different data points, such as consistent vital signs and lab values within expected ranges for the condition being studied.
- **Variability:** High-quality test data incorporates both partial and complete dates to reflect real-world data collection challenges. It includes unscheduled visits and assessments, and represents different treatment arms and randomization scenarios to cover all aspects of the trial design.
- **Protocol Non-Compliance:** The data includes common protocol deviations, such as out-of-window visits and missed assessments. It also represents scenarios like screen failures who receive treatment or subjects randomized despite violating inclusion/exclusion criteria, to test the robustness of conditional forms and fields and analysis processes.
- **Volume and Scalability:** provides a sufficient volume of data to stress-test systems and processes. It scales to represent multi-site, multi-country trials when necessary, ensuring that data management systems can handle the full scope of the planned study.

# Validation principles

# “ Who validates the AI —after deployment? "

***New Era of AI Governance:*** *No fewer than seven major guidance documents and frameworks have been issued or finalized in the past 3 years.*

***Regulators Response -*** *The FDA-EMA joint principles of January 2026 represent the most significant step toward resolving this gap through international harmonization.*

- Regulatory frameworks require human accountability
- Human oversight is necessary (not limited to HITL)
- Focused on Performance Monitoring and Sensitivity/Specificity. Evidence is based on “How does the model perform over time on new data?”
- Models learn and drift in production.
- Validation must be a continuous activity, not a one-time event.

# Human-in-the-Loop (HITL): 3 Layers of Oversight


AUTOMATED

## 1. Machine Validation Layer

Rule-based Planner → Filler → Validator auto-loop

**System-Level**

- Mandatory Data Preview
- Auto-checks every generation cycle

HUMAN-IN-THE-LOOP

## 2. Human Review Layer

Mandatory Data Preview & Approval before EDC load

**Process-Level**

- Reviewer roles defined: i.e. Business Validator
- Change-control gated by human approval

OVERSIGHT

## 3. Governance & Regulatory Layer

EU AI Act · ICH E6(R3) · FDA-EMA 2026 · GAMP 5 / CSA

**Governance-Level**

- EU AI Act (2024) — Human Oversight provisions
- ICH E6 (R3) — Fit-for-purpose validation

# Takeaways & Conclusion

## Synthetic clinical participant data is no longer theoretical

- With GenAI combined with study metadata, rules, and protocols, it is now possible to generate **clinically plausible, rule-aware, and scalable** synthetic data to support study build, UAT, and early analytics.

## Who can use synthetic clinical participant

- Synthetic data is a platform capability—not just a generation tool
- it becomes an **end-to-end enabler** for development, testing, and analytics readiness

## Validation & Governance Conclusions

- Post-deployment validation cannot be left to the AI itself
- Human-in-the-Loop remains a regulatory and ethical requirement

# Thank You!

# &

# Questions?