cdisc C3C

构建临床研究"虚拟数据工厂"

基于CDISC SDTM标准与AI驱动的自动化数据流水线

Building a "Dummy Data Factory" for Clinical Research

AI-Driven Automated Data Pipeline Based on CDISC SDTM Standards

Stanley Wei

Novartis

December 5, 2025

cdisc C3C Disclaimer

免责声明

⚠️ 本演示文稿中表达的观点和意见仅代表作者个人立场，不代表诺华公司的官方政策或立场。

诺华公司不对演示内容的准确性、完整性或可靠性做出任何陈述或保证。

🤖 本演示文稿及内容部分由 AI Agent 协作完成，仅供学习交流使用。

内容可能存在不准确之处，我们不对其正确性、完整性或适用性做出任何保证。 请在实际应用前进行独立验证和评估。

Agenda / 议程

1. **Background of Smart DDF**
   智能虚拟数据工厂背景
   Understanding the foundation and purpose of DDF / 理解DDF的基础和目标

2. **Key Considerations for DDF Design**
   DDF设计的关键考量
   Design principles and data characteristics / 设计原则和数据特征

3. **DDF Workflow & Online Quick Demo**
   DDF工作流与在线快速演示
   Rapid walkthrough of DDF workflow and the interactive experience / 快速浏览DDF工作流和交互体验

4. **Technology Stacks and Architectures**
   技术栈与架构
   Technical implementation and system design / 技术实现和系统设计

5. **Challenges and Limitations**
   挑战与限制
   Technical challenges and current constraints / 技术挑战和当前限制

6. **Q&A**
   问答环节
   Questions and open discussion / 问题与开放讨论

为什么虚拟数据工厂对加速统计编程至关重要

Why DDF Matters in Accelerating Statistical Programming

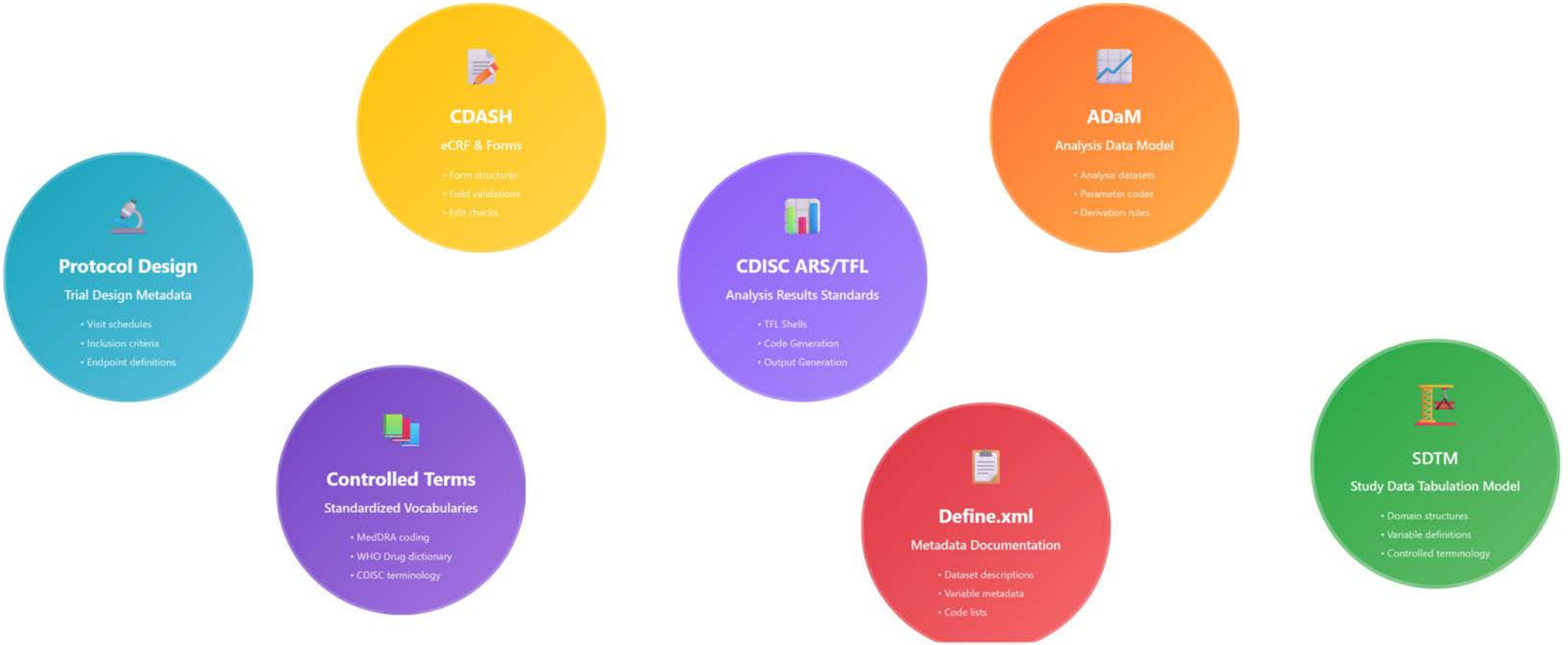
虚拟数据在临床数据运营中发挥着关键作用，包括数据库设置、数据管理和编程开发生命周期，特别是在以下情况下：

Dummy data plays a critical role in clinical data operations, including database setup, data management and programming development lifecycle, especially when:

- 为数据库设置、逻辑检查等提供稳健的用户验收测试（UAT）
  Robust UAT for database setup, edit checks, etc.
- 真实研究数据不可用或需要等待具有有意义的代表性的数据截取
  Real study data is not available or requires waiting for a data cut with meaningful representation.
- 真实研究数据无法覆盖所有场景，包括可能影响宏或自动化过程可靠性的罕见边缘情况
  Real study data cannot cover all scenarios, including rare edge cases that may affect the reliability of macros or automated processes.
- 由于隐私、安全或合规性考虑，无法使用真实研究数据，需要匿名化或

临床试验数据运营中的元数据

Metadata in Clinical Trial Data Operations



Comprehensive metadata ecosystem supporting clinical data operations and analysis

为什么选择创建CDISC SDTM数据集?

Why Choose to Create CDISC SDTM Datasets?

桥梁作用

The Bridge

连接上游源数据和下游ADaM的"原始数据"桥梁

"Raw data" bridge between upstream source data and downstream ADaM

高度结构化

Highly Structured

高度结构化和标准化的数据格式

Highly structured and standardized data format

基于规则的解决方案

Rule-Based Solution

可以建立通用且可靠的基于规则的解决方案

Feasible to establish a general but reliable rule-based solution

灵活场景

Flexible Scenarios

易于生成灵活的适合用途的场景

Easy to generate flexible fit-for-purpose scenarios

DDF相比现有解决方案的优势

Advantages of DDF Over Existing Solutions

现有解决方案
Existing Solutions

虚假数据生成
Fake Data Generation
不适应临床试验场景，生成的数据没有意义
Does not adapt to clinical trial scenarios, generated data not meaningful

无逻辑与关联
No Logic & Relationships
不考虑变量和数据集之间的逻辑和关系
Does not consider the logics and relationships across variables and datasets

手动配置
Manual Configuration
不是元数据驱动的过程，难以配置特定场景的参数
Not a metadata driven process and hard to configure parameters for specific scenarios

学习曲线高
High Learning Curve
需要大量额外努力来生成真实的CDISC数据集，学习曲线高
Need lots of additional efforts to generate realistic CDISC datasets, high learning curve

集成有限
Limited Integration
没有内置API服务，难以集成到应用开发和测试流程中
Not ship with built-in API services and hard to integrate for app development and test process

VS

虚拟数据工厂
Dummy Data Factory

基于规则且可定制
Rule-Based & Customizable
基于规则的解决方案，高度支持用户定制，具有适当的参数设置
Rule-based solution and highly support user customization, with appropriate params setting

场景模拟
Scenario Simulation
支持多样化场景模拟，变量级别生成规则配置，符合CDISC和监管规则
Supports simulation of diversified scenarios, variable level generation rule configuration and in compliance with CDISC and regulatory rules

元数据驱动与API就绪
Metadata-Driven & API Ready
元数据驱动过程以及内置API服务，易于与自动化流程集成
Metadata driven process as well as built-in API services, easy to be implemented with automated processes

端到端解决方案
End-to-End Solution
可以创建高度关联且有意义的SDTM、ADaM、ARDs和TFLs，促进整个流程的开发和验证
Could create highly linked and meaningful SDTM, ADaM, ARDs and TFLs, to facilitate the whole process development and validation

生产就绪

cdisc C3C

Metadata-Driven Dummy Data Factory

A concrete example of how structured metadata enables intelligent dummy data generation

Variable	Label	Type	Length	Controlled Terms
STUDYID	Study Identifier	Char	12	e.g., "ABC-123-001"
DOMAIN	Domain Abbreviation	Char	2	"AE"
USUBJID	Unique Subject Identifier	Char	20	e.g., "ABC-123-001-001"
AESEQ	Sequence Number	Num	8	1, 2, 3...
AETERM	Reported Term for AE	Char	200	Free text (realistic symptoms)
AEDECOD	Dictionary-Derived Term	Char	200	MedDRA PT terms
AESEV	Severity/Intensity	Char	20	MILD, MODERATE, SEVERE
AESER	Serious Event	Char	1	Y, N
AESTDTC	Start Date/Time	Char	19	ISO 8601 format
AEENDTC	End Date/Time	Char	19	ISO 8601 format

AE生成算法逻辑示例

SDTM AE Generation Algorithm Logic

```
# ========================================
## 1. WEIGHTED_CHOICE_MEDICAL Algorithm

    # Apply demographic adjustments
    age = subject_demographics['age']
    sex = subject_demographics['sex']
    medical_history = subject_demographics.get('medical_history', [])

    # Age-based adjustments
    age_adjustments = {
        'common_mild': {
            'young': (18, 35, 1.2),  # Higher rate of mild AEs in young adults
            'middle': (36, 65, 1.0),  # Baseline
            'elderly': (66, 100, 0.8)  # Lower rate of mild AEs, higher serious
        },
        'serious': {
            'young': (18, 35, 0.3),   # Much lower serious AE rate
            'middle': (36, 65, 1.0),  # Baseline
            'elderly': (66, 100, 2.5)  # Much higher serious AE rate
        }
    }

    # Sex-based adjustments for specific AEs
```

元数据驱动的风险因素　MedDRA合规　CDISC SDTM标准　临床相似性

元数据驱动的虚拟数据工厂示例

Metadata-Driven Dummy Data Factory Example

关键洞察：从多个域（DM、EX、MH）的元数据驱动智能、临床相似的数据生成，符合CDISC标准和MedDRA编码

Metadata Repository

Demographics (DM)

- Age: 68 years
- Sex: Male
- Race: Caucasian

Concomitant Medication (CM)

- Drug Class: Cardiovascular
- Duration: 180 days
- Dose: 50 mg QD

Medical History (MH)

- Hypertension
- Type 2 Diabetes

Metadata-Driven Algorithm

1. Age 68 → ↑2.5x Serious AE Risk
2. Male → ↑1.0x MI Risk
3. CV Drug → ↑2.0x Hypertension
4. Duration 180d → Chronic Events
5. MedDRA Mapping Applied

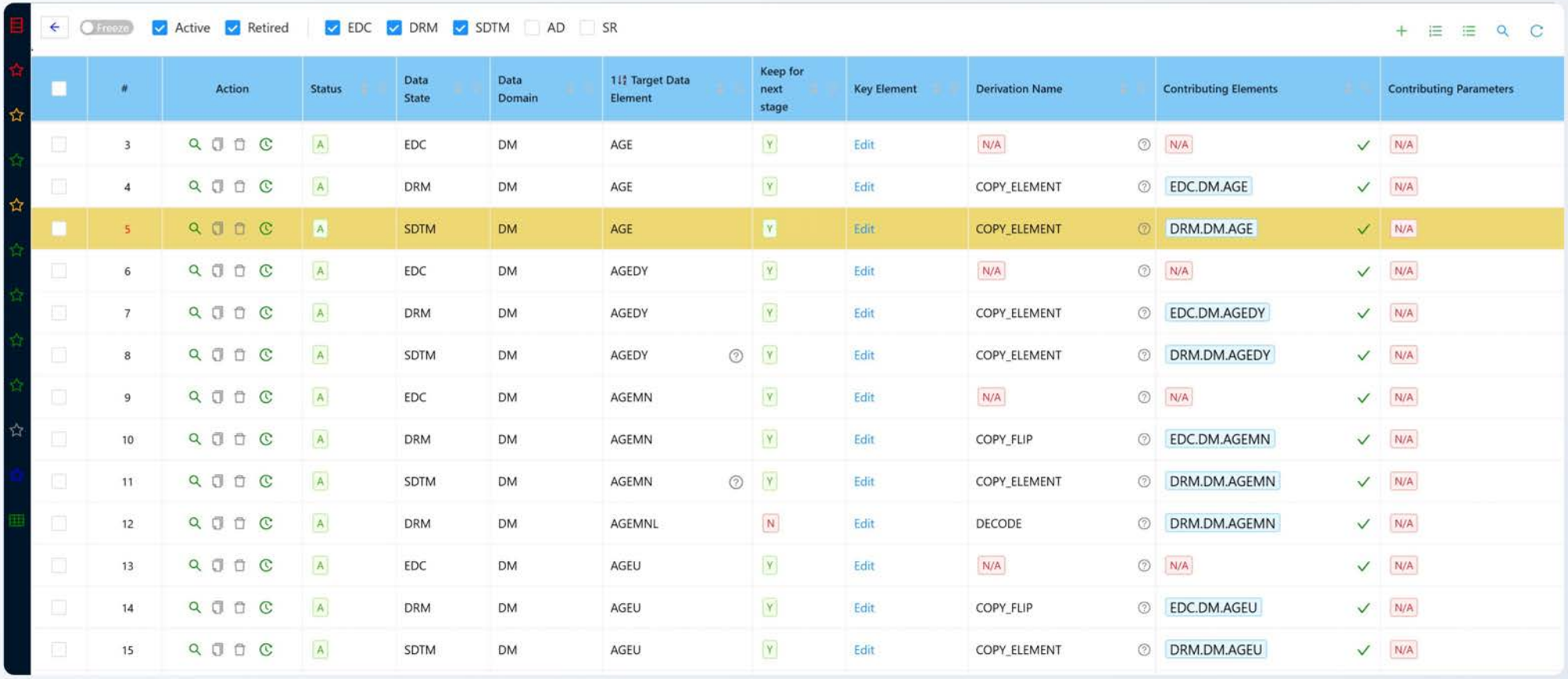
Generated AE Data

SERIOUS AE — Cardiac Disorders

AETERM:	Myocardial Infarction
AESEV:	SEVERE
AESER:	Y (Hospitalization)
AESTDTC:	2024-03-15
AEDECOD:	Myocardial infarction
AEBODSYS:	Cardiac disorders
Duration:	21 days (90% resolution)

cdisc C3C 元数据管理

Centralized Metadata Repository View



#	Action	Status	Data State	Data Domain	Target Data Element	Keep for next stage	Key Element	Derivation Name	Contributing Elements	Contributing Parameters
3		A	EDC	DM	AGE	Y	Edit	N/A	N/A	N/A
4		A	DRM	DM	AGE	Y	Edit	COPY_ELEMENT	EDC.DM.AGE	N/A
5		A	SDTM	DM	AGE	Y	Edit	COPY_ELEMENT	DRM.DM.AGE	N/A
6		A	EDC	DM	AGEDY	Y	Edit	N/A	N/A	N/A
7		A	DRM	DM	AGEDY	Y	Edit	COPY_ELEMENT	EDC.DM.AGEDY	N/A
8		A	SDTM	DM	AGEDY	Y	Edit	COPY_ELEMENT	DRM.DM.AGEDY	N/A
9		A	EDC	DM	AGEMN	Y	Edit	N/A	N/A	N/A
10		A	DRM	DM	AGEMN	Y	Edit	COPY_FLIP	EDC.DM.AGEMN	N/A
11		A	SDTM	DM	AGEMN	Y	Edit	COPY_ELEMENT	DRM.DM.AGEMN	N/A
12		A	DRM	DM	AGEMNL	N	Edit	DECODE	DRM.DM.AGEMN	N/A
13		A	EDC	DM	AGEU	Y	Edit	N/A	N/A	N/A
14		A	DRM	DM	AGEU	Y	Edit	COPY_FLIP	EDC.DM.AGEU	N/A
15		A	SDTM	DM	AGEU	Y	Edit	COPY_ELEMENT	DRM.DM.AGEU	N/A

CDISC SDTM/ADaM Compliant

Interactive Metadata Editor

Automated Validation

Documentation Generator

DDF的数据特征输入

Data Characteristics Inputs for the DDF

虚拟数据工厂 / DUMMY DATA FACTORY

需要准备什么?

What needs to be prepared?

DDF接收精心准备的数据特征，以预填充研究终点指标/实验室检查等研究相关的参数的默认设置，确保生成的数据反映分布特征。

The DDF ingests prepared data characteristics to pre-populate default settings for endpoints, tests, and parameters of interest, ensuring generated data mirrors the distribution profile.

人群人口统计学

Population Demographics

关键队列的地区组合、年龄/性别分布、入组节奏和随机化比例。

Region mix, age/sex distributions, enrollment cadence, and randomization ratios for key cohorts.

研究终点指标/实验室检查库

Endpoint/Test Libraries

参数编码、结果单位、实验室面板、访视窗口和受控术语映射。

Parameter codes, result units, lab panels, visit windows, and controlled terminology mappings.

分布特征

Distribution Profiles

用作合成生成默认值的汇总统计或拟合分布（正态、对数正态、分类概率）。

Summary stats or fitted distributions (normal, log-normal, categorical odds) used as defaults for synthetic generation.

时间与场景规则

Temporal & Scenario Rules

访视频率、非计划访视逻辑、缺失模式、方案偏离和分支逻辑等。

Visit frequency, unscheduled logic, missingness patterns, protocol deviations, and branching triggers.

获取特征 / ACQUIRING THE CHARACTERISTICS

如何获取特征?

How do we acquire characteristics?

结合定量摘要和定性知识，模拟或仿真相关疾病治疗特征场景。

Combine quantitative summaries with qualitative knowledge to simulate or mimic the scenarios of the related disease treatment.

历史研究档案、CT.gov等

Historical Study Archives, CT.gov, etc.

- 从公开文献、数据库或自有研究中提取研究终点指标/实验室检查的统计摘要
  Extract SDTM/ADaM summaries from locked studies
- 按治疗领域或研究药物类型进行版本控制，确保数据特征的准确性和一致性
  Version control per therapeutic area or study drug type

申办方和治疗领域手册

Sponsor & TA Playbooks

- 导入申办方拥有的参考范围和分层研究终点指标/实验室检查
  Import sponsor-owned reference ranges and tiered tests
- 自定义感兴趣研究终点指标/实验室检查值的可接受范围
  Determine acceptable range for endpoints of interest

自动化分析服务

Automated Profiling Services

- 连接到MCP或元数据中心的API，使用LLMs辅助分析数据特征
  APIs to MCP or metadata, with the usage of LLMs
- 机器学习辅助拟合分布默认值
  ML-assisted fitting for distribution defaults
- 持续监控以刷新数据特征

SME和生物统计专家评审

SME & Biometric Peer Review

- 评审研究终点指标/实验室检查的统计摘要
  Peer review of the data characteristics
- 确保数据特征的准确性和一致性
  Ensure the accuracy and consistency of the data characteristics

cdisc C3C 数据特征示例

Data Characteristics Examples

COMMON DEMOGRAPHICS 常见人口统计学

Population Baseline 人口基线

Age
年龄

Mean	54.3 years	SD	12.7
Min	21	Max	82
Median	53		

Sex
性别

Male	124 (53.7%)	Female	107 (46.3%)
Total	231		

Weight (kg)
体重 (kg)

LAB HEMATOLOGY TESTS 实验室血液学检测

Blood Test Parameters 血液检测参数

Hemoglobin
血红蛋白

HGB

Mean	14.2 g/dL	SD	1.6
Min	10.3	Max	18.1

Reference Range

White Blood Cells
白细胞

WBC

Mean	7.2 × 10^9/L	SD	2.1
Min	3.5	Max	14.8

Reference Range

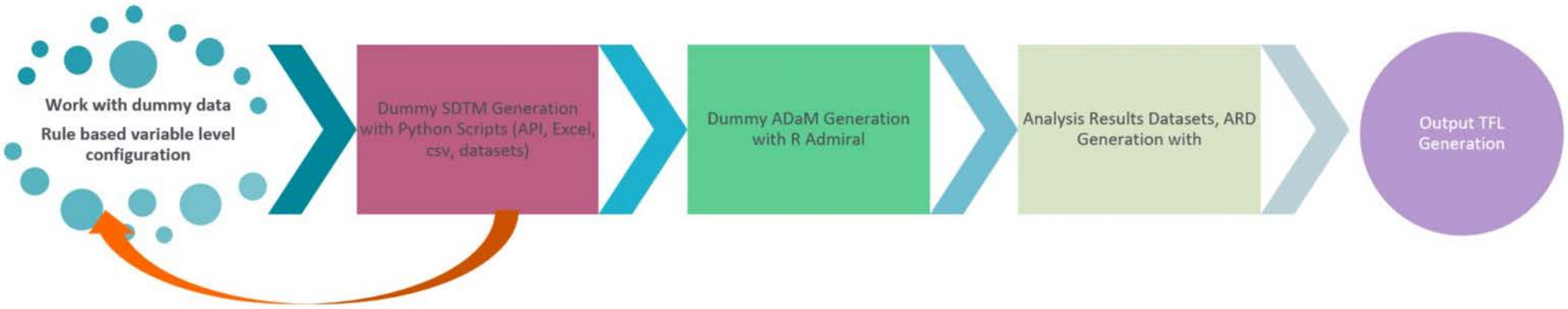
Platelets
血小板

PLT

cdisc C3C

DDF实施流程

DDF Implementation Process



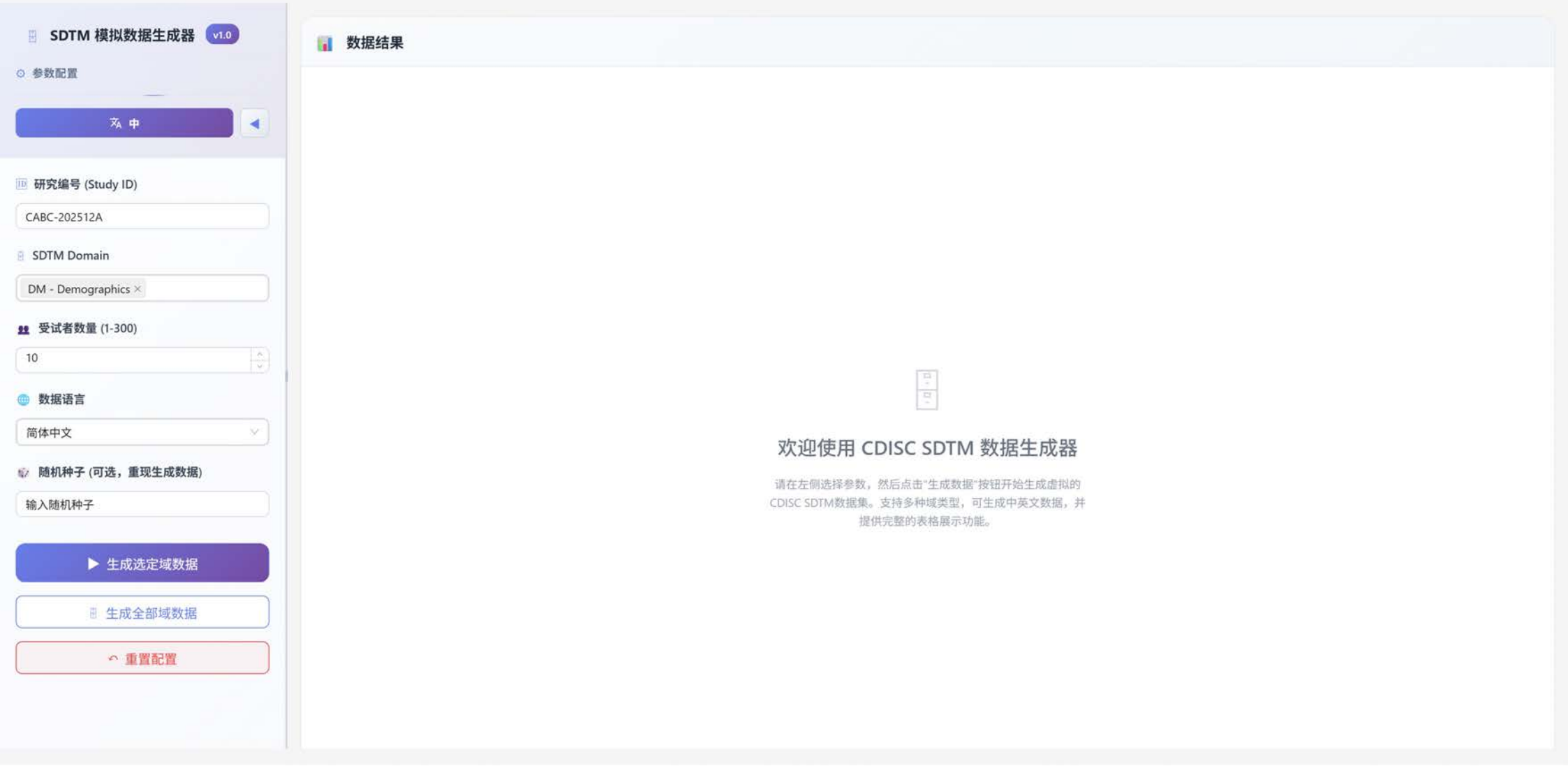
元数据驱动 / Metadata-Driven

自动化工作流 / Automated Workflow

符合CDISC标准 / CDISC Compliant

端到端解决方案 / End-to-End Solution

虚拟数据工厂在线快速演示



Adverse Events Dashboard

受试者数量: 60 加载数据

Interactive analysis dashboard with detailed subject information

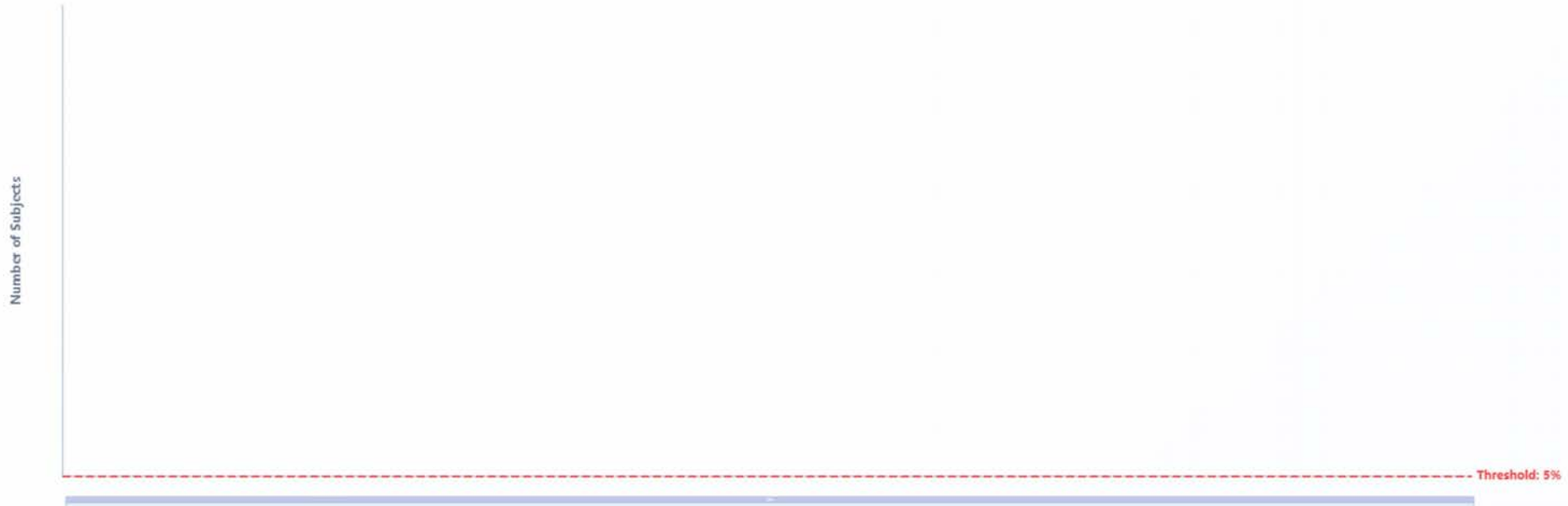
Analysis Controls

仅 TEAE 分组: 亚组: 无 图型: 分组柱状图 统计口径: 按受试者数（去重 USUBJID） 严重度筛选: 全部 排序: 高到低 图表方向: 垂直 显示滑块 阈值线（%）:

5% 10% 自定义

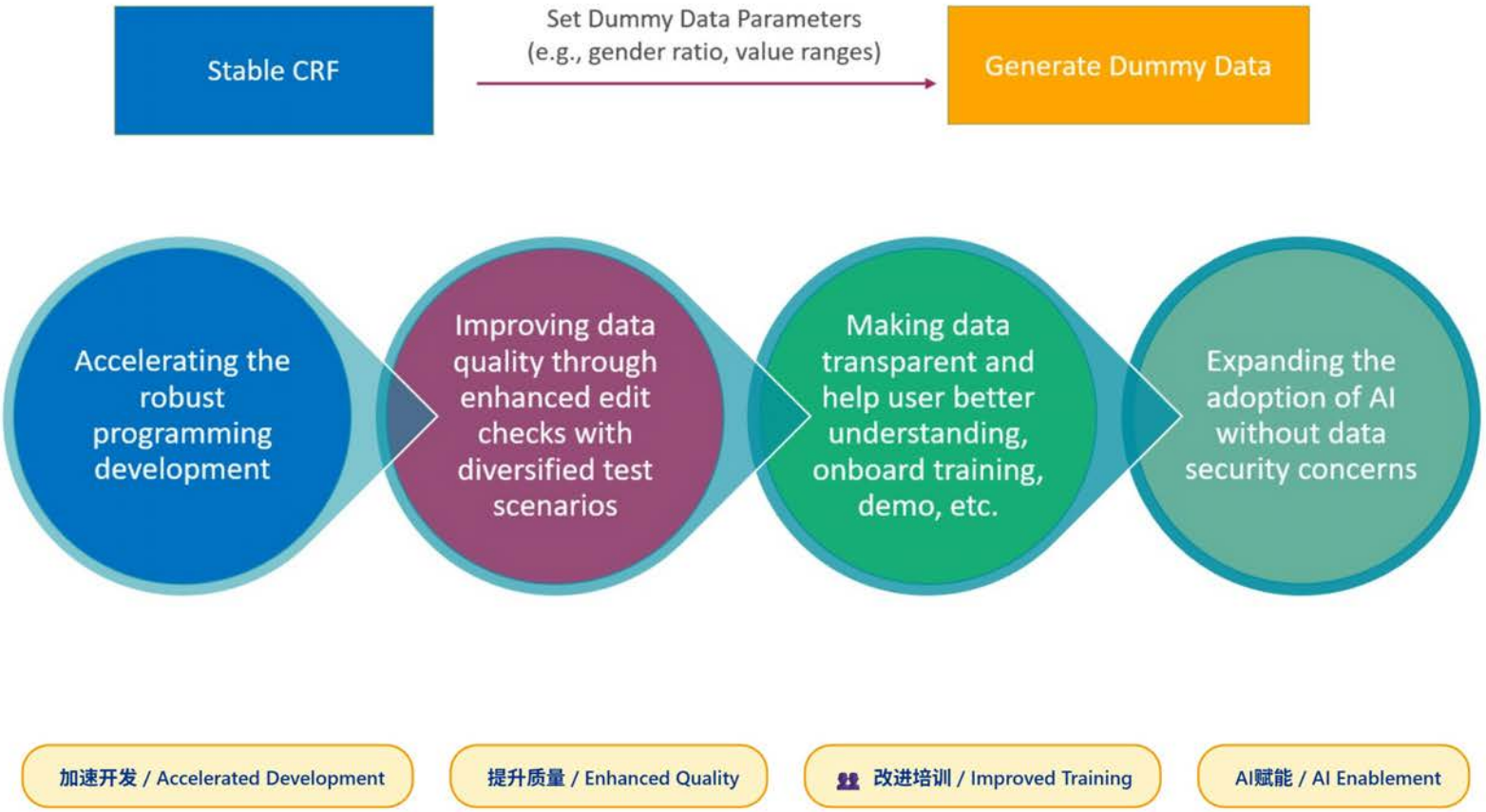
Adverse Event Frequencies by AEDECOD

下载图片



DDF预期成果与收益

DDF Expected Outcomes & Benefits



cdisc C3C

关键技术挑战与解决方案

Key Technical Challenges & Solutions

1 📋 SDTM Standard Compliance

SDTM标准符合性

Challenge / 挑战:

How to ensure generated data strictly adheres to SDTM standards, including data models, variable naming conventions, and coding rules?

如何确保生成的数据严格符合SDTM模型、变量命名、编码规则的精确映射?

Solutions / 解决方案

- ✓ Build comprehensive SDTM metadata repository with domain models, variable dictionaries, and controlled terminologies
  建立完整的SDTM元数据库，包含域模型、变量字典和受控术语
- ✓ Implement automated validation layer using CDISC standards and Pinnacle 21 validation rules
  实现基于CDISC标准和Pinnacle 21验证规则的自动化校验层
- ✓ Create domain-specific generators with built-in SDTM constraints and business rules
  创建内置SDTM约束和业务规则的特定域生成器

Technical Approaches / 技术方法:

Metadata-Driven | Rule-Based Validation | Template Mapping | CDISC CT Integration

2 ✨ Data Realism & Plausibility

数据真实性与合理性

Challenge / 挑战:

How to generate data that is not only structurally correct but also clinically realistic in terms of value ranges, distributions, and logical relationships?

如何让生成的数据不仅格式正确，其内容（数值范围、数据分布、逻辑关系）也贴近真实临床场景?

Solutions / 解决方案

- ✓ Analyze real-world clinical trial data to extract statistical distributions and patterns
  分析真实临床试验数据，提取统计分布和模式特征
- ✓ Apply probabilistic models (Normal, Log-normal, Weibull) for continuous variables
  对连续变量应用概率模型（正态、对数正态、威布尔分布）
- ✓ Implement correlation matrices and copula functions for inter-variable dependencies
  实现相关性矩阵和Copula函数处理变量间依赖关系
- ✓ Define clinical business rules and reference ranges based on therapeutic area expertise
  基于治疗领域专业知识定义临床业务规则和参考范围

Technical Approaches / 技术方法:

Distribution Fitting | Correlation Analysis | Clinical Rules Engine | Statistical Sampling

3 🔗 Complex Logical Relationships

复杂逻辑关联

Challenge / 挑战:

How to efficiently handle complex constraints and maintain consistency across different data domains and within domain variables?

如何高效处理不同数据域之间、同一域内不同变量之间的复杂约束与一致性?

Solutions / 解决方案

- ✓ Design hierarchical data generation pipeline: DM → SV → domain-specific data
  设计分层数据生成管道：DM → SV → 特定域数据
- ✓ Implement dependency graph to manage inter-domain relationships via USUBJID, VISIT, dates
  实现依赖图管理域间关系（通过USUBJID、VISIT、日期关联）
- ✓ Create constraint validation engine for temporal logic, reference ranges, and derived variables
  创建约束验证引擎处理时间逻辑、参考范围和派生变量
- ✓ Apply transaction-based approach to ensure atomicity and rollback on constraint violations
  应用事务性方法确保原子性，约束违反时回滚

Technical Approaches / 技术方法:

Graph-Based Dependencies | Constraint Solver | Temporal Logic Engine | Transaction Management

cdisc C3C

技术栈概览

Technology Stack Overview

前端框架
Frontend Framework

Vue 3.x + Ant Design 响应式设计
Vue 3.x + Ant Design Responsive Design
Composition API

TypeScript
类型安全 / Type Safety

样式与动画
Styling & Animation

CSS3
动画与网格 / Animations & Grid

Canvas API
动态效果 / Dynamic Effects

后端技术
Backend Technology

LLM + Python + R + MySQL/MariaDB
关系型数据库 / Relational Database

Redis
内存缓存 / In-Memory Cache

构建工具
Build Tools

Vite
快速构建工具 / Fast Build Tool

Node.js
运行时环境 / Runtime Environment

cdisc 总结 C3C

Summary

1

加速临床编程

Accelerates Clinical Programming

智能虚拟数据生成显著加快临床编程流程

Smart dummy data generation speeds up clinical programming processes significantly

2

提升数据与代码质量

Improves Data & Code Quality

使用智能虚拟数据有助于通过全面测试提高代码质量

Using smart dummy data helps enhance code quality through comprehensive testing

3

扩展场景覆盖

Expands Scenario Coverage

它能够实现更广泛的场景覆盖，确保更可靠的临床研究工作流

It enables broader scenario coverage, ensuring more reliable clinical research workflows

cdisc 致谢 C3C

Acknowledgements

FY

Fan Yang（杨帆）

Team Contributor for Data Characteristics

CL

Carina Li（李媛）

Team Contributor for Algorithm Design

Thank you for your dedication and valuable contributions!

cdisc C3C

Questions & Answers

问题与解答

Thank you for your attention!

感谢您的关注!

zhijun.wei@novartis.com

Scan to Follow

扫码关注