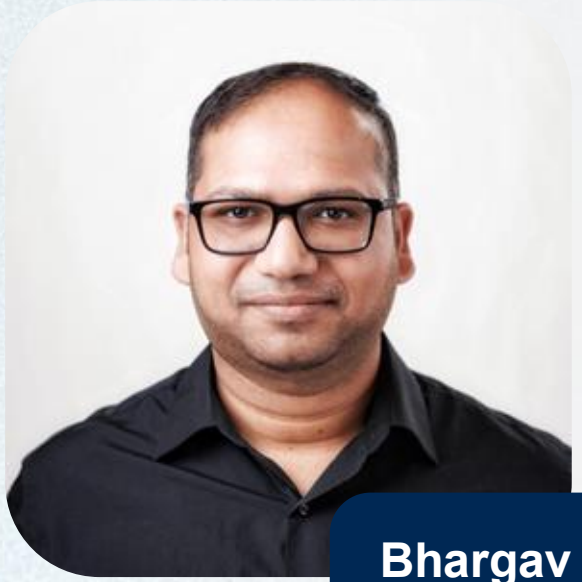
Balancing Privacy and Traceability:

A Multi-Step Methodology for De-Identifying Clinical Trial Data

Bhargav Koduru, Luke Morgan

2026-07-15



Bhargav Koduru
Director, Data Curation
Medidata Solutions

Luke Morgan
Senior Manager, Clinical Data Standardization
Medidata Solutions

Secondary Data Demand Clashes with Regulatory and Privacy Standards

THE CORE CONFLICT

Exponential AI & Research Demand

- Secondary analysis of clinical trial datasets is crucial for academic research, healthcare insights, and training advanced AI/ML algorithms.
- Clinical databases from Electronic Data Capture (EDC), Electronic Health Records (EHRs), and Patient Reported Outcomes (PROs) contain immense scientific utility.
- However, reusing person-specific health information is strictly governed by global legal frameworks to prevent identity theft and privacy loss.

The Friction: Privacy vs. Traceability

- Privacy Regulations (GDPR, HIPAA, PHIPA) legally restrict secondary data use without explicit consent unless data is successfully anonymized.
- Health Authorities (FDA, EMA, Health Canada) demand strict data traceability, data integrity, audit trails, and data provenance.
- The Dilemma: Anonymization aims to protect privacy by minimizing traceability, whereas health regulators demand total traceability back to source records.

GDPR Recital 26 Exempts Successful Anonymization from Data Protection Rules

LEGAL LANDSCAPE

GDPR Recital 26:

"The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

- Once clinical data is sufficiently and successfully anonymized, it falls completely outside GDPR restrictions and can be reused for secondary analysis.
- SUCCESS does not require proving that re-identification is mathematically impossible. Irish DPC (2017) and international frameworks state that data is anonymous if it can be shown that re-identification is highly unlikely given current technology and specific case circumstances.

GDPR Treats Pseudonymous Data as Personal Data with Key Relaxations

THE DEFINITIONAL RIFT

Truly Anonymous Data

- Complete and irreversible separation between the data and the identified individuals.
- No direct or indirect translation mechanism exists that allows secondary recipients to map anonymous records back to their original subjects.
- Exempt from oversight: Completely escapes GDPR restrictions, offering full freedom of secondary use.

Pseudonymous Data

- Reversible Masking: Direct identifiers are replaced with tracking values (keys). The original controller can re-identify, but secondary recipients cannot.
- Legally Personal: Pseudonymous data remains within the scope of the GDPR.
- Relaxed Provisions: GDPR relaxes some provisions (data by design, security) to encourage its use for scientific and historical research, provided an independent re-identification risk analysis is performed.

Study Variables Categorized into Three Risk Profiles Under EMA Policy 0070

VARIABLE CLASSIFICATION

Direct Identifiers

Variables that explicitly and immediately point to a specific individual.

Key Elements:

- Sponsor Identifiers
- Participant IDs
- Investigator Details
- Site Details
- Country / Region / Address
- Date of Birth

Quasi-Identifiers

Nuanced variables that do not directly identify but can lead to re-identification when combined with external databases.

Key Elements:

- Age Outliers (e.g., >89 years)
- Extreme Height / Weight values
- Rare clinical parameters

Non-Identifiers

Variables carrying minimal to no risk of re-identification.

Key Elements:

- Standard clinical outcomes
- Aggregated lab values
- Common adverse events
- No changes or masking required
- Preserves baseline utility

Multi-Layered Pipeline Systematically Handles Privacy at Three Distinct Levels

THE THREE-PHASE PIPELINE

PHASE 1

Raw/Source Data

Masking directly at extraction:

- Replace direct identifiers with masked/null values
- Apply systematic date offsets to protect timeline dates
- Cap demographic outliers (age >89, height, weight)

PHASE 2

SDTM/Data Model

Standardization and terminological de-identification:

- Convert trial data into standard CDISC SDTM structures
- Harmonize source-specific terminologies into standards
- Erase unique trial design metadata signatures ('footprints')

PHASE 3

Database Integration

Addressing integrated cross-study risks:

- Merge clinical trials into combined project databases
- Harmonize potential quasi-identifiers such as visit and dose values across stacked trials
- Split studies into randomized 'pseudo studies' to obscure subject counts

Randomized Split into Pseudo Studies Prevents Re-Identification via Subject Counts

PHASE 3: THE PSEUDO STUDY STRATEGY

The Study-Size Vulnerability

- Uniqueness Hazard: A clinical trial with a small or highly specific subject count (e.g., exactly 24 subjects) is highly vulnerable. An adversary matching external trial listings (like ClinicalTrials.gov) can re-identify study participants.
- The Randomized Split: To neutralize this risk, our methodology splits original studies into randomized, non-overlapping 'pseudo studies'.
- Obscured Patient Counts: Each pseudo study is assigned a newly generated 32-character randomized hexadecimal ID, ensuring the final patient counts differ significantly from original trials while preserving standard analytic utility.

Original Study ID	Randomized Pseudo Study ID
MDS001 (Study 1)	9AS80D7FA0SD7FA0S7JK434H
MDS001 (Study 1)	24L36B52K43JBO3G6OKYI4K5
MDS002 (Study 2)	B57H352VJK453O0SG7FCS09
MDS003 (Study 3)	31284KJ7B659A8SDF74VD947

Phase 1 Protects Core Identity While Preserving Clinical Utility & Intervals

PHASE 1: RAW/SOURCE DATA

Direct Masking & Blinding

- Masking Names and Identifiers: Replacing subject, investigator, and site identifiers with randomized, non-reversible keys.
- Sponsor & Product Blinding: Masking specific investigational product names and compounds which would immediately reveal the identity of the sponsor.

Interval Integrity & Outliers

- Systematic Date Offsets: Rather than deleting dates (which ruins time-to-event analysis), all dates are shifted back or forward by a randomized number of days per subject. This preserves exact clinical visit intervals and timelines while masking actual dates.
- Demographic Outlier Capping: High-risk outliers (like age >89, extreme weight or height) are mathematically capped to standard maximums, neutralizing identification via rare demographic combinations.

Phase 2 Eliminates the Source Footprint by Harmonizing Terminology

PHASE 2: SDTM STANDARDIZATION

Natural SDTM Anonymization

- Data Model Conversion: Raw heterogeneous EDC, EHR, and Patient-Reported Outcomes (PROs) are structured into standardized CDISC SDTM domains.
- Harmonization of Values: Custom, source-specific terminologies are mapped directly to industry-standard controlled terminology.
- Structure Alignment: Restructuring the database architecture ensures that no trace of the original trial design remains visible.

Erasing the Source Footprint

- Eliminating Ad-Hoc Names: Mapping highly customized source/protocol fields to a unified public terminology.
- Metadata Masking: Blinding variables that indicate unique administrative protocols or proprietary electronic data capture structures.
- Regulatory Integration: Standardization ensures the final dataset remains fully traceable for regulatory compliance audits while being safely anonymized.

Stacked Integrated Databases Harmonize Diverse Trial Visit Structures

PHASE 3: CROSS-STUDY INTEGRATION

Source Clinical Trial	Original Variable / Visit Value	Harmonized Terminology (Clean Stacked Domain)
Study 1 (Source A)	Screening / Visit 1 / Visit 2 / Visit 3 / Follow-up	Screening and Baseline / Visit 1 / Visit 2 / Visit 3 / Follow-up
Study 2 (Source B)	Screening / Day 1 / Week 1 / Week 2 / Follow up	Screening and Baseline / Visit 1 / Visit 2 / Visit 3 / Follow-up
Study 3 (Source C)	Baseline / Visit 1 - Day 1 / Visit 2 - Day 7 / Visit 3 - Day 14 / Follow-up Visit	Screening and Baseline / Visit 1 / Visit 2 / Visit 3 / Follow-up
Study 1 (Dosing)	Pre-Dose / Post-Dose	Pre-Dose / Post-Dose
Study 2 (Dosing)	PreDose / During Dose / Post Dose	Pre-Dose / Concurrent Dose / Post-Dose
Study 3 (Dosing)	Pre-Dose / Concurrent Dose / Post-Dose	Pre-Dose / Concurrent Dose / Post-Dose

Conclusion

The growing demand for secondary use of clinical trial data, particularly for research, regulatory insights, and AI/ML development, requires a careful balance between patient privacy and regulatory expectations. Although anonymization and de-identification are essential for protecting participant confidentiality, health authorities continue to require data traceability, integrity, provenance, and auditability.

Our patent-pending, standards-based approach demonstrates that these seemingly conflicting objectives can coexist. By leveraging CDISC SDTM principles, systematic management of direct and quasi-identifiers, value harmonization, data aggregation, and study-splitting methodologies, clinical datasets can be effectively anonymized while preserving their scientific value and regulatory alignment.

Ultimately, this framework enables the responsible reuse of clinical data, protecting patient privacy without compromising data quality, transparency, or the ability to generate meaningful insights that advance research, innovation, and AI-driven healthcare.