Data Protection and Ethics in Clinical Research

From ”How I want to do it" to "How I am allowed to do it"

Mariana M S Santos, PhD
Investigator Assistant
ERA Chair iSTARS

From Planning to Practice

Old view: Plan → Execute → "Fix" ethics later.

New view: **Ethics-by-Design** (Integrating GCP/GDPR into the initial roadmap).

The Pillars of Trust

Why Patient Trust is the Most Valuable Asset

The Social Contract

- Patients share sensitive information with the expectation of **confidentiality** and **benefit**.
- Trust is the fuel for Big Data; without it, data sources dry up.

The Risk of "Data Silos"

- When trust is broken (e.g., data leaks or commercial misuse), patients opt-out.
- Lower participation = Biased datasets = Poor Medical AI.

The Pillars of Trust

Why Patient Trust is the Most Valuable Asset

The Three Pillars
Transparency: Clear communication on *how* and *why* data is used.
Beneficence: Ensuring the research aims to improve patient outcomes or public health.
Security: Technical excellence in protecting the digital "body" of the patient.

Data is not just a resource; it is a digital extension of the patient.

Key Regulations Overview

The Two Pillars: ICH-GCP vs. EU-GDPR

ICH-GCP (The Ethical Standard)	EU-GDPR (The Legal Framework)
Focus: Clinical Trial Quality & Ethics	**Focus:** Data Privacy & Digital Rights
Protects: The Rights, Safety, and Wellbeing of human subjects.	**Protects:** The Personal Data of individuals within the EU.
Data Goal: Integrity, Traceability, and Verifiability (ALCOA+).	**Data Goal:** Security, Minimization, and Lawful Processing.
Main Document: Trial Protocol & Informed Consent Form (ICF).	**Main Document:** Privacy Notice & Data Processing Agreement.

GCP tells you how to conduct the research;
GDPR tells you how to handle the data resulting from it.

Faculdade de Medicina

ICH-GCP Principles

The Gold Standard for Clinical Research

Principles for Data Science

Scientific Rigor: Trials should be scientifically sound and described in a clear, detailed protocol.

Ethics Committee Oversight: Mandatory review by an independent board (CEIC/CE in Portugal).

Quality Assurance: Implementing systems with procedures that assure every aspect of the trial is handled correctly.

If a study is not GCP-compliant, its data cannot be used for regulatory submissions (EMA/FDA) or high-impact publications.

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

ICH-GCP: Informed Consent Form (ICF) 2.0

The Ethical Gateway to Data

No data processing can start before a subject provides voluntary, informed, and documented consent.***

"Digital" Challenge

- ✓ **Specificity:** Consent must be clear about *what* data is collected and *how* it will be analyzed (e.g., AI/Machine Learning).
- ✓ **Secondary Use:** Can the data be used for future, unspecified research? (Broad vs. Specific Consent).

Dynamic Consent

The shift from a "one-time paper signature" to digital platforms where patients can manage their permissions.

Faculdade de Medicina

ICH-GCP: Informed Consent Form (ICF) 2.0

The Ethical Gateway to Data

The Data Science Dilemma: Specific vs. Broad Consent

Specific Consent
Consent is tied to a single, well-defined study.
Risk: Data becomes "locked" in a silo; reuse for future AI training is legally restricted.

VS.

Broad Consent
Allows data use for future scientific research within a specific field.
Advantage: Enables Big Data and retrospective Data Science projects.

Modern ICFs must explicitly mention **Automated Processing, Machine Learning,** and potential **International Data Transfers**.

What happens if a patient opts out after the data is integrated into a Neural Network?

U LISBOA UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Faculdade de Medicina

ICH-GCP: Informed Consent Form (ICF) 2.0

The Ethical Gateway to Data

The Right to be Forgotten vs. Machine Learning

The "Data Death" Dilemma

Withdrawal of Consent: Participants have the right to opt-out at any time.

The Technical Conflict: If a patient withdraws, how do we "unlearn" their data from a pre-trained Neural Network?

Planning Ahead: Your data pipeline must account for data deletion/extraction without compromising the entire model.

Irreversibility: Once data is fully anonymized, withdrawal is often no longer possible (or required)—but the transition point must be clear.

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Data Integrity (ALCOA+)

If it’s not Traceable, it’s not Science

ALCOA+

- **Attributable:** Who created the data point? (Audit trails).
- **Legible:** Can it be read and understood years later?
- **Contemporaneous:** Was it recorded at the time of the event?
- **Original:** Is it the source record or a certified copy?
- **Accurate:** Is it free from errors and edited only via audit trails?
- **"+" (The Complete Picture):** Complete, Consistent, Enduring, and Available.

The Golden Rule:

"Data that lacks integrity is ethically compromised. Science built on untraceable data cannot be validated, reproduced, or published."

The Role of the Ethics Committee (CE)

Your Partners in Ethical Data Science

Gatekeepers:

- No clinical data can be accessed or analyzed without a formal **favorable opinion**.
- The CE evaluates the balance between **Scientific Value** and **Participant Risk**.
- CE joins the DPO (Data Protection Officer) to ensure both GCP and GDPR are met.

What the CE looks for in Data Science?

- **Data Flow:** Where will the data be stored? (Institutional servers vs. Cloud).
- **The "Why":** Is the use of AI/Machine Learning justified for this clinical question?
- **Impact:** How will the results affect patient care or privacy?

Approval must be obtained before any data extraction begins.

GDPR Fundamentals

What is the GDPR?

Definition: General Data Protection Regulation

- The most comprehensive data privacy law in the world (EU Regulation 2016/679).
- It grants individuals **control** over their personal data and unifies privacy laws across Europe.
- Allows the free flow of data for scientific progress while ensuring the **fundamental rights** of the patient.

GDPR Fundamentals

Understanding Data Categories under GDPR

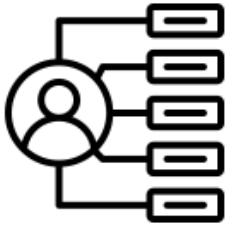
Data Hierarchy: who is the data owner?

- **Standard Personal Data:** Name, ID, Location, IP address.
- **Special Categories (Art. 9):** Health, Genetic, and Biometric data. These are "Protected by Default" and require extra security.

The "identifiability" Rule:
If a person can be identified (directly or indirectly), the GDPR applies.

GDPR Fundamentals

Personal Data vs. Special Categories



What is Personal Data? (The Broad Definition)

Any information relating to an **identified** or **identifiable** person.

- *Direct Identifiers:* Name, Citizen Card Number, Phone, Email.
- *Indirect Identifiers:* IP Address, GPS coordinates, or even a unique combination of demographics.

GDPR Fundamentals

Personal Data vs. Special Categories

Special Categories (Article 9): "Sensitive Data”

- **Health Data:** Medical history, diagnoses, lab results, clinical notes.
- **Genetic Data:** DNA sequences, inherited characteristics.
- **Biometric Data:** Fingerprints, facial recognition, iris scans.

Rule: Processing is *prohibited* unless a specific derogation (like Scientific Research) applies.

GDPR Fundamentals

Personal Data vs. Special Categories

How do I know?

If you can link a data point back to a person using "all means reasonably likely to be used," it is Personal Data.

GDPR Fundamentals

The "Scientific Research" Exception (Art. 9, 2j):
GDPR explicitly recognizes that scientific research is a public good.
It allows the processing of sensitive data without needing the same level of granularity as commercial data use.

"Friend" Benefits	"Price" of Friendship
Secondary Use: Eases the process of using data for purposes other than the original collection (if safeguards are met). **Longer Retention:** Allows researchers to keep data for longer periods if necessary for validation.	You must implement **Safeguards**: Technical and organizational measures (Encryption, Pseudonymization, Access Control).

GDPR: Institutional Actors

Who is Who? Roles and Responsibilities

Data Controller (The Institution)	• The Hospital, University, or Research Center. • They "own" the legal responsibility and decide **why** and **how** data is processed.
Data Processor (The Researcher/Lab)	• The person or entity that processes data on behalf of the Controller (e.g., a Data Scientist or an external CRO).
Data Protection Officer (DPO)	• The mandatory supervisor within the hospital. • *Your Strategy:* Engage with the DPO early to validate your Data Science pipeline.
Data Subject	• The patient. The center of all protection efforts.

Compliance in Portugal

Navigating the Portuguese Legal Landscape

Regulatory Authority:

CNPD (Comissão Nacional de Proteção de Dados): The Portuguese supervisory authority that enforces GDPR.

The Legal Pillars:

- **Law 58/2019 (The GDPR "Adaptation"):** Clarifies how health data can be reused for research in Portugal.
- **Law 21/2014 (Clinical Research Law):** Regulates how trials and studies are conducted (linked to GCP).

Compliance in Portugal

Navigating the Portuguese Legal Landscape

Key Requirements for Researchers:

- ❑ **DPIA (Data Protection Impact Assessment):** Mandatory for high-risk processing (e.g., large-scale health data or AI).
- ❑ **Security Standards:** Specific mandates for database encryption and access logging in public hospitals.

Faculdade de Medicina

Data Science Safeguards

Anonymization vs. Pseudonymization

Pseudonymization (The "Reversible" Shield)	Anonymization (The "Point of No Return")
• **Definition:** Replacing identifiers (names, IDs) with a code or "Token." • **Status:** It is still **Personal Data** because a "key" exists that can relink the data. • **GDPR Impact:** Full GDPR rules apply, but it counts as a security "safeguard."	• **Definition:** Processing data so it is **irreversibly** impossible to identify the person. • **Status:** It is **Non-Personal Data**. • **GDPR Impact:** GDPR no longer applies. This is the "safe zone" for Open Data.

Can someone identify the patient by cross-referencing this dataset with other available data?

If yes → It is **Pseudonymized**. | If no → It is **Anonymized**.

U LISBOA UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Data Science Safeguards

Data Minimization in Practice

From "Collect Everything" to "Extract what Matters"

The Principle of Minimization (Art. 5): Data must be adequate, relevant, and limited to what is necessary for the research goal.

The Technical Filter:

- **Feature Selection:** Do you really need the patient's exact birth date, or just their age bracket?
- **Vertical Filtering:** Never SELECT *. Explicitly list only the columns required for your hypothesis.
- **Horizontal Filtering:** Only extract records that meet your study's inclusion criteria (e.g., specific diagnosis codes).

Data Science Safeguards

Data Minimization in Practice

From "Collect Everything" to "Extract what Matters"

Privacy by Design:

Apply filters at the **source** (database level) rather than downloading everything and filtering in your Python environment.

The Goal:

Reduce the "Blast Radius." If your local machine is compromised, the less sensitive data you have, the better.

Security Measures & The "Clean Room"

Protecting Data: Encryption, Storage, and Access

The "Golden Rule" of Storage

Institutional Servers: Always the primary choice. Managed by IT, backed up, and behind hospital firewalls.

Personal Laptops: Forbidden for raw/pseudonymized sensitive data unless encrypted and authorized. Never store data on unencrypted USB sticks or commercial clouds (Dropbox, iCloud).

Security Measures & The "Clean Room"

Protecting Data: Encryption, Storage, and Access

Encryption: The Double Layer

At Rest: Data stored on disks must be encrypted (e.g., BitLocker, FileVault). If the laptop is stolen, the data remains a "black box.”

In Transit: Use secure protocols (HTTPS, SFTP, VPN) when moving data between the database and your analysis environment.

Faculdade de Medicina

Security Measures & The "Clean Room"

Protecting Data: Encryption, Storage, and Access

Access Control (Role-Based)

Need-to-know basis: Only team members involved in the analysis should have access.

Audit Logs: Every "read" or "write" action should be recorded.

The "Clean Room" Concept:

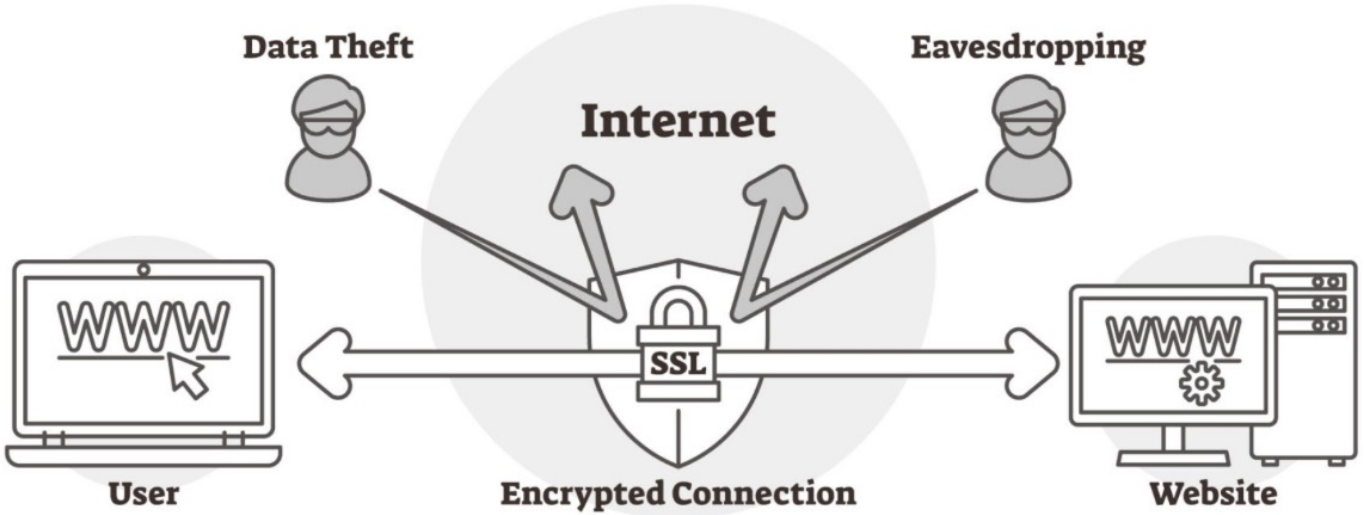
A secure, isolated environment where data is analyzed but cannot be downloaded or exported without oversight.

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Faculdade de Medicina

SSL
Secure Sockets Layer



LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Data Transfer Agreements (DTA)

When Data Leaves the Nest: Managing Multicenter Studies

What is a DTA?

A formal contract that defines the terms under which data is shared between institutions.

It specifies: **Purpose**, **Duration**, **Allowed Users**, and **Security Standards**.

The Legal Transfer Flow:

Sending Institution: Remains the "Controller" (or Joint Controller).

Receiving Institution: Becomes a "Processor" or "Joint Controller".

Beyond Law – Ethics in AI/ML

Fairness, Bias, and the "Black Box" Problem

❖ The Bias Trap (Garbage In → Bias Out)

AI models learn from historical data. If the data contains human prejudices (e.g., under-representation of minorities), the AI will replicate them.

Medical Example: A diagnostic algorithm for melanoma trained only on Caucasian skin types may fail on patients with darker skin tones.

❖ Fairness in Data Science

Ensuring the model performs equally well across different genders, ethnicities, and socio-economic backgrounds.

Faculdade de Medicina

Beyond Law – Ethics in AI/ML

Fairness, Bias, and the "Black Box" Problem

❖ Explainability

A "legally clean" model is useless if a doctor cannot understand **why** it made a prediction.

The "Right to Explanation" under GDPR: Patients may have the right to know the logic behind an automated decision.

❖ Researcher’s Responsibility

Audit your model for bias before deployment. Ethics is a continuous process, not a checkbox.

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Beyond Law – Ethics in AI/ML

Fairness, Bias, and the "Black Box" Problem

❖ Transparency and Reproducibility

Opening the Black Box: Our Ethical Duty to Document

❖ The Reproducibility Crisis

Many published AI models fail when applied to new datasets because the original process was not transparent.

Beyond Law – Ethics in AI/ML

Fairness, Bias, and the "Black Box" Problem

❖ Documentation as an Ethical Act

It is not just about "clean code"; it’s about allowing others to audit your logic.

Data Provenance: Documenting every transformation from raw data to the final model.

Beyond Law – Ethics in AI/ML

Fairness, Bias, and the "Black Box" Problem

❖ **The "Black Box" Challenge**

Deep Learning models are notoriously opaque.

Requirement: Use tools (like SHAP or LIME) to explain which features influenced the model's decision.

❖ **Open Science vs. Privacy:**

✓ Share your **code** (GitHub/GitLab) even if you cannot share the **data**.

✓ Use **Version Control** (Git) to track changes in your hypothesis and methodology.

Faculdade de Medicina

Comments & Questions

Mariana M S Santos
marianamsantos@medicina.ulisboa.pt

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria