MACHINE LEARNING FOR HEALTH

Clinical Data: What Does the Model Really See?

Heterogeneity, quality, missingness, and information leakage

TECHNICAL HANDOUT · SESSION 02

Professor Flávio Luiz Seixas
Graduate Program in Computing

Academic teaching material based exclusively on the PDFs available in the project

Table of Contents

1. Introduction
2. 1. The clinical data ecosystem
3. 2. Quality as fitness for use
4. 3. Time, selection, and leakage
5. 4. Governance and provenance
6. Session summary
7. References

Introduction

Clinical data are not a continuous, neutral observation of the patient. They result from encounters, instruments, documentation practices, and prior decisions. The matrix given to an algorithm therefore contains both biological signals and traces of the care process. [1]

The goal of this session is to recognize data types, structure, temporality, missing-data mechanisms, selection biases, and leakage. These properties determine what a model can learn and where its performance may fail. [1]

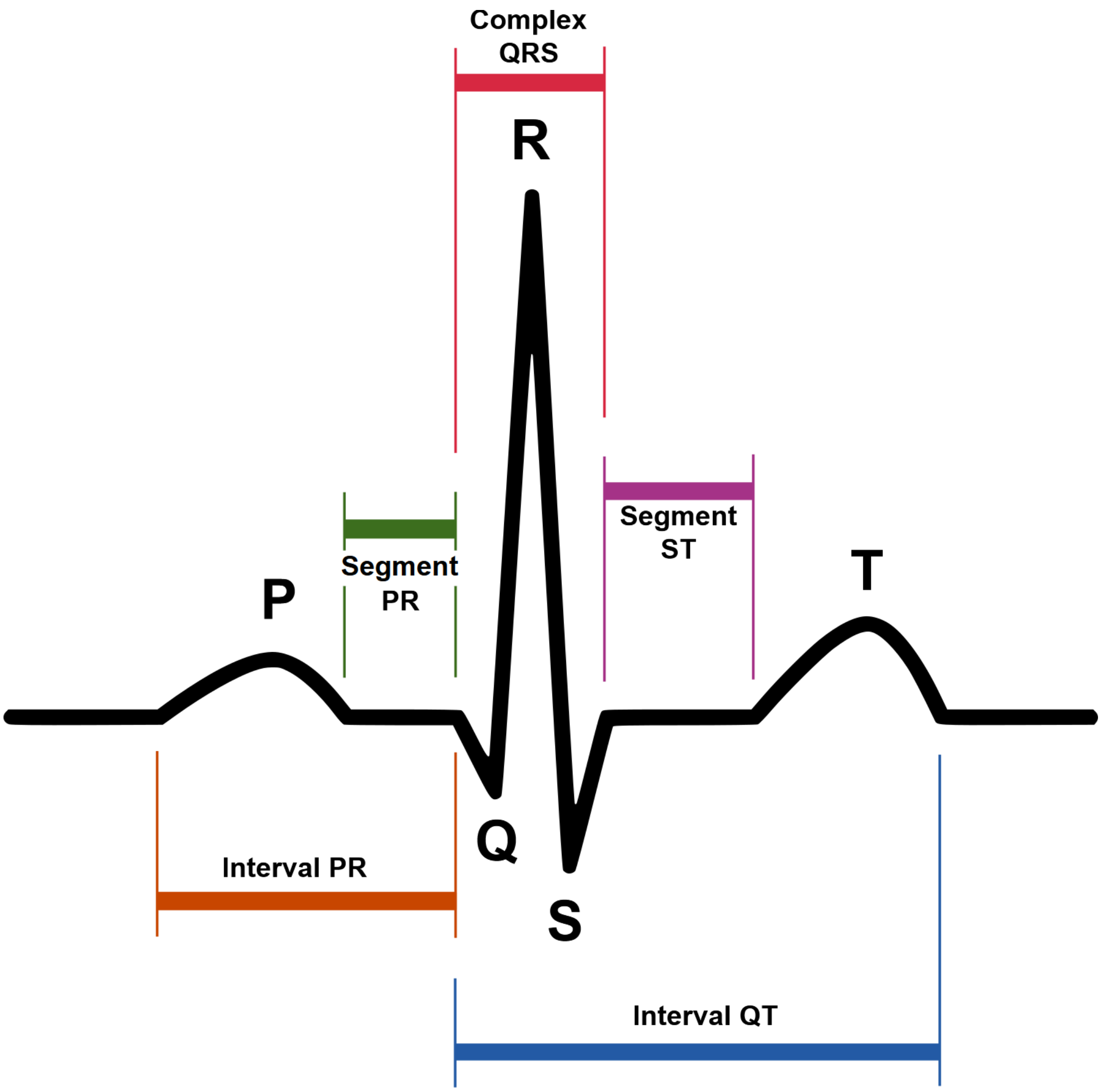

Figure 1 — Electrocardiographic signal as an example of temporal data. The visual source is identified in the session README; the discussion is based on [1] and [2].

1. The clinical data ecosystem

The health record brings together observations produced by professionals, patients, and devices. Structured data facilitate querying and aggregation, whereas narratives preserve context and relationships that are difficult to anticipate in fixed fields. Neither form is automatically complete or correct [1].

Images, signals, molecular data, and patient-generated information have different structures and scales. Digital images contain pixels and metadata; signals represent samples over time; omics data often combine high dimensionality with smaller samples. Integration requires preserving provenance and meaning [2].

> **KEY CONCEPT**
> **The health record brings together observations produced by professionals, patients, and devices.**

2. Quality as fitness for use

Quality must be judged in relation to purpose. A code suitable for billing may be insufficient to represent a phenotype; a value collected during care may reflect clinical selection rather than a measurement planned for research. Before modeling, each variable should be described by its origin, unit, time, transformation, and reason for collection [1].

Missing data may be completely random, may depend on observed information, or may be related to unobserved factors. These situations correspond to MCAR, MAR, and MNAR, respectively. Excluding incomplete records or imputing values without considering the mechanism can change the population analyzed and the signal available [3].

3. Time, selection, and leakage

In clinical prediction, time defines what would be available at the moment of use. Variables recorded after the decision point may reveal the outcome or an action caused by it. This leakage produces optimistic estimates that are not reproduced when the system is used prospectively [4].

It is also necessary to ask who enters the dataset. The probability of hospitalization, testing, follow-up, or documentation may depend on access, severity, and local practice. The hospital, equipment, or protocol may become a shortcut that the model learns instead of the intended phenomenon [4].

4. Governance and provenance

Governance should record who may access the data, for what purpose, and under which controls. Provenance makes it possible to trace a variable from acquisition to the representation used by the model. This documentation is essential for auditing, reproducibility, and the evaluation of change [5].

Session summary

The clinical data ecosystem: the health record brings together observations produced by professionals, patients, and devices.
Quality as fitness for use: quality must be judged in relation to purpose.
Time, selection, and leakage: in clinical prediction, time defines what would be available at the moment of use.
Governance and provenance: governance should record who may access the data, for what purpose, and under which controls.

References

[1] Shortliffe, E. H.; Chiang, M. F. Biomedical Data: Their Acquisition, Storage, and Use. In: Biomedical Informatics. Springer, 2021. External link

Local file: pdfs/2-Biomedical Informatics/II-Biomedical Data- Their Acquisition, Storage, and Use.pdf · PDF pages used: 3, 5, 9, 18, 22.

[2] Sim, I.; Sirota, M. Data and Computation: A Contemporary Landscape. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/III-Data and Computation- A Contemporary Landscape.pdf · PDF pages used: 1, 3, 7, 14.

[3] Bellazzi, R.; Dagliati, A.; Nicora, G. Predicting Medical Outcomes. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XI-Predicting Medical Outcomes.pdf · PDF pages used: 6, 7, 8.

[4] Subramanian, D.; Cohen, T. A. Machine Learning Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/VI-Machine Learning Systems.pdf · PDF pages used: 12, 16, 22, 52, 67.

[5] Goodman, K. W.; Miller, R. A. Ethics in Biomedical and Health Informatics: Users, Standards, and Outcomes. In: Biomedical Informatics. Springer, 2021. External link

Local file: pdfs/2-Biomedical Informatics/XII-Ethics in Biomedical and Health Informatics- Users, Standards, and Outcomes.pdf · PDF pages used: 8, 9, 10, 24.