MACHINE LEARNING FOR HEALTH

Ethics, Explainability, and Privacy in Health AI

How to audit models, reduce inequities, and protect clinical data

TECHNICAL HANDOUT · SESSION 14

Professor Flávio Luiz Seixas
Graduate Program in Computing

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

Table of Contents

Introduction

The evaluation of AI in health does not end with accuracy. It is necessary to understand for whom a system works, what explanation it provides, how it handles sensitive data, and which harms may emerge from the interaction between model, institution, and user. [1]

This session connects explainability, equity, privacy, and security as dimensions of governance throughout the life cycle. [1]

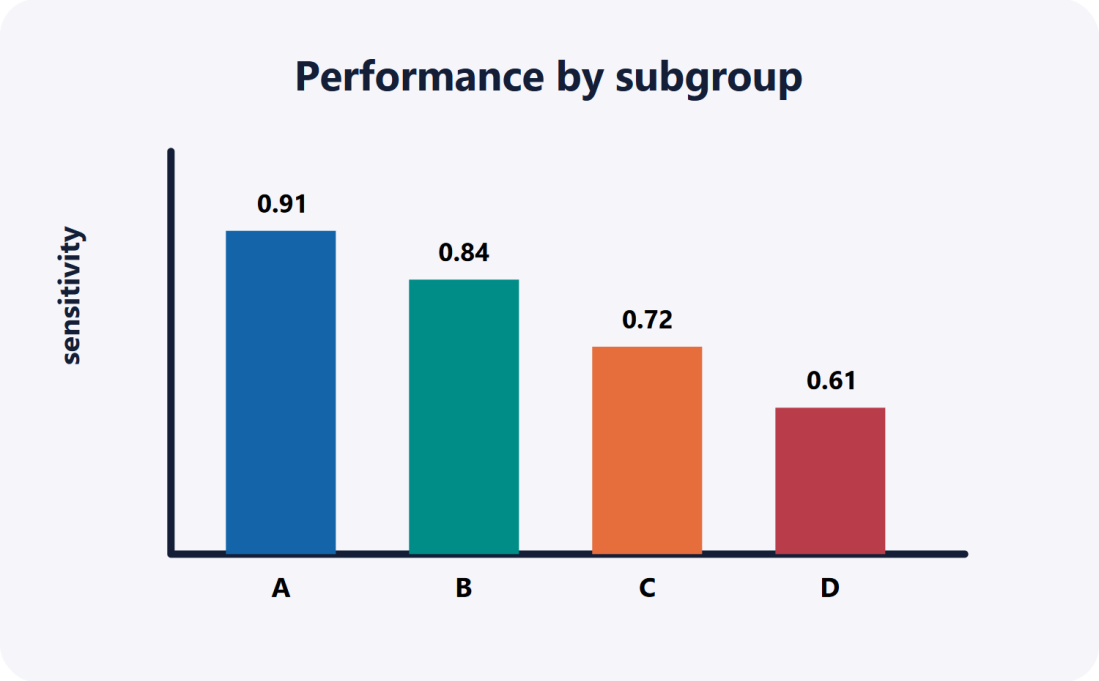

Figure 1 — Performance auditing across subgroups. Source: original illustration based on [2].

1. Interpretability and explainability

Complex models map predictors to outputs through relationships that are difficult to express in human reasoning. Explainability encompasses methods that make aspects of this process accessible. The purpose must be defined: global understanding, local justification, debugging, auditing, and clinical communication require different evidence [1].

Local methods attribute contributions from variables to a prediction; saliency maps highlight image regions associated with the output. These representations do not prove causality or guarantee that the model uses valid clinical mechanisms. They must be evaluated for faithfulness, stability, and utility [1].

KEY CONCEPT
Complex models map predictors to outputs through relationships that are difficult to express in human reasoning.

2. Equity and sources of bias

Bias can be introduced during formulation, data selection, construction, deployment, and human interpretation. Equity is not equivalent to distributing errors in a mechanically equal way: it requires making benefits, harms, affected groups, and context explicit [2].

Auditing should disaggregate performance and calibration and examine coverage, label quality, and proxies. A seemingly neutral variable may represent access, utilization, or institutional practice and reproduce inequity when treated as clinical need [2].

3. Privacy, confidentiality, and security

Privacy concerns control over and appropriate use of personal information; confidentiality concerns the obligations of those who receive the data; security comprises controls that protect confidentiality, integrity, and availability [3].

Removing direct identifiers does not eliminate every reidentification risk. Governance should combine minimization, authorization, access control, logging, retention, and incident response. Cybersecurity affects patient safety when the availability or integrity of clinical systems is compromised [3].

4. Distributed collaboration and surveillance

Federated learning keeps data local while coordinating updates to a shared model. This architecture reduces centralization but still requires evaluation of leakage through updates, heterogeneity across institutions, and governance [4].

Post-deployment surveillance should observe performance, subgroups, incidents, and changes in use. Ethical effects are not fixed properties of the model file; they emerge in the sociotechnical system and can change over time [2].

5. Governance of generative AI

Generative AI adds risks of hallucination, loss of context, and more autonomous behavior. A recent checklist organizes safeguards into predevelopment, development, postdeployment, and cross-cutting themes, including access control, expert verification, robustness to prompts, decision boundaries, a human in the loop for high-risk uses, and a continuous monitoring plan [5].

Responsibility must be distributed before an incident occurs: institutions, professionals, researchers, and developers should have documented duties. The consensus study found greater agreement on institutional and professional duties than on assigning responsibility to developers, showing that a lack of consensus does not justify leaving governance undefined [5].

For LLMs, RAG can anchor responses in literature or guidelines, but explanation quality depends on retrieval, faithfulness, and user understanding. A plausible or lengthy explanation does not substitute for prospective validation, workflow integration, and outcome evaluation [6].

SOURCE LIMITATION
Detailed legal coverage of Brazil's LGPD and the GDPR remains limited; this revision updates generative-AI governance but does not replace legal interpretation of those regulations.

Session summary

Interpretability and explainability: complex models map predictors to outputs through relationships that are difficult to express in human reasoning.
Equity and sources of bias: bias can be introduced during formulation, data selection, construction, deployment, and human interpretation.
Privacy, confidentiality, and security: privacy concerns control over and appropriate use of personal information; confidentiality concerns the obligations of those who receive the data; security comprises controls that protect confidentiality, integrity, and availability [3].
Distributed collaboration and surveillance: federated learning keeps data local while coordinating updates to a shared model.
Governance of generative AI: generative ai adds risks of hallucination, loss of context, and more autonomous behavior.

References

[1] Li, R. C. et al. Explainability in Medical AI. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/VIII-Explainability in Medical AI.pdf · PDF pages used: 1, 2, 4, 8, 13.

[2] Korngiebel, D. M.; Solomonides, A.; Goodman, K. W. Ethical and Policy Issues. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XVIII-Ethical and Policy Issues.pdf · PDF pages used: 1, 7, 8, 9, 14.

[3] DeMuro, P.; Norwood, H. Privacy and Security. In: Health Informatics. 2021.

Local file: pdfs/4-Health Informatics/XXVIII-Privacy and Security.pdf · PDF pages used: 1, 2, 5, 32.

[4] Cohen, T. A.; Patel, V. L.; Shortliffe, E. H. Reflections and Projections. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XX-Reflections and Projections.pdf · PDF pages used: 14, 26, 30.

[5] Cha, H. et al. Development of research ethics guidelines for healthcare generative artificial intelligence: deriving expert consensus through a Delphi study. BMC Medical Ethics, 2026. External link

Local file: pdfs/1-Machine Learning/s12910-026-01513-4.pdf · PDF pages used: 1, 2, 6, 7, 9, 13, 14, 15.

[6] Zamanifar, A.; Taherkordi, A.; Farhadi, A. (eds.). Explainable Large Language Models in Healthcare Applications. Springer, 2026. External link

Local file: pdfs/1-Machine Learning/Explainable Large Language Models in Healthcare Appliadeh Zamanifar & Amir Taherkordi & Amirfarhad Farhadi.pdf · PDF pages used: 84, 143, 158, 160, 162, 207, 283.