Etica, intelligenza artificiale e ricerca clinica strumenti per esplorare in sicurezza

Honoria Ocagli

Unit of Biostatistics, Epidemiology and Public Health,
Department of Cardiac Thoracic Vascular Sciences and Public Health
University of Padova

AIEOP
ASSOCIAZIONE ITALIANA EMATOLOGIA ONCOLOGIA PEDIATRICA

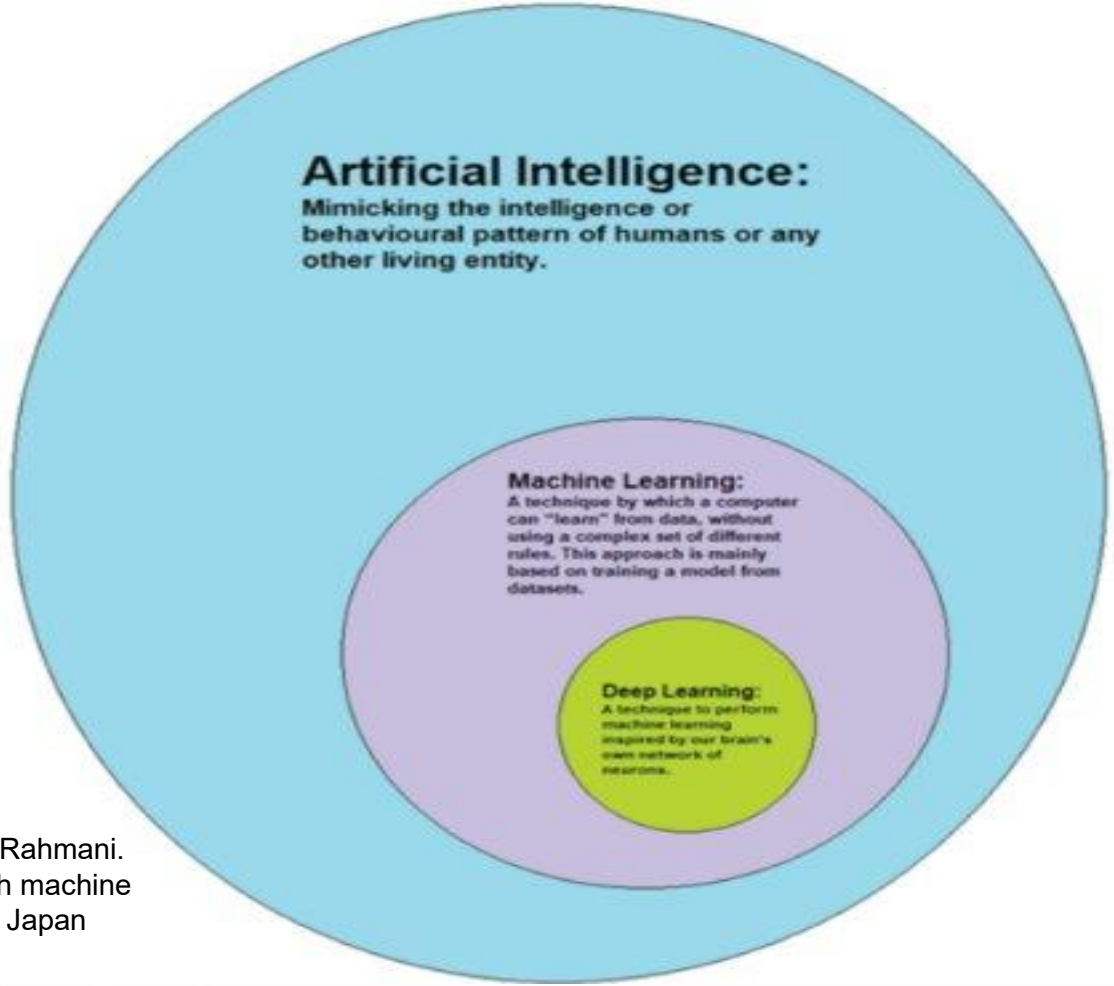
CONGRESSO INFERMIERISTICO AIEOP

ROMA, 23-24 Settembre 2025
CENTRO CONGRESSI
UNIVERSITÀ CATTOLICA
DEL SACRO CUORE

Disclosures of Name Surname

Company name	Research support	Employee	Consultant	Stockholder	Speakers bureau	Advisory board	Other

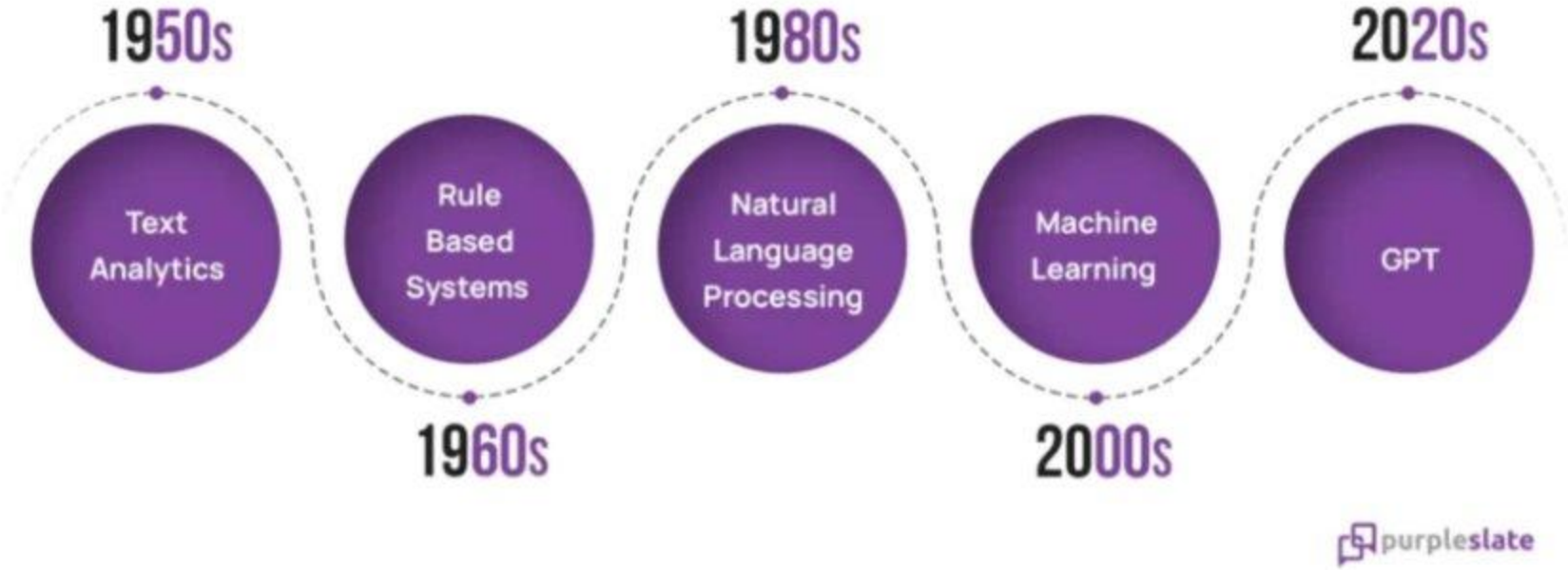
Services	Laboratories	Graduate Masters Programs	PhD programs	Residency Program / Postgraduate School
1. Service for Clinical Trials and Biometrics 2. Service for Systematic Reviews	1. Epidemiological Methods and Biostatistics (LEMBS) 2. Artificial Intelligence in Medicine (LAIMS) 3. Experimental Sciences (LESEXP) 4. Clinical Epidemiology and Digital Health (LACED) 5. Health Service Research (LASER) 6. Environmental Epidemiology and Ecological Statistics (LESE) 7. Nursing Research (NURE) 8. Systematic Reviews and Evidence Medicine (LaSyReM)	1. Biostatistics for Scientific Publication 2. Advanced Biostatistics 3. Experimental Research 4. Research Nursing 5. Big Data, Artificial Intelligence in Clinical Research and Precision Medicine 6. Geographical and Spatial Epidemiological Modeling 7. Systematic Reviews and Reproducible Research 8. Pharmacoepidemiology 9. Omics Data Analysis 10. Radiomics: advanced quantitative image analysis 11. Nursing, clinical research, and artificial intelligence in oncology and palliative care 12. Legal and regulatory aspects of clinical research	1. Clinical Epidemiology and Biostatistics 2. Nursing and Health Sciences	Medical Statistics and Biometrics



Zohuri, Bahman, and Farhang Mossavar Rahmani. "Artificial intelligence driven resiliency with machine learning and deep learning components." Japan Journal of Research 1.1 (2023).

What Are AI and LLMs?

- **Artificial Intelligence (AI)**: Machines performing tasks that typically require human intelligence
- **Natural Language Processing (NLP)**: Understanding and generating human language
- **Machine Learning (ML)**: A subset of AI where systems learn patterns from data instead of following fixed rule
- **Large Language Models (LLMs)**: Trained on vast text data to process free-text input and generate structured outputs
- **GenAI**: A broader term for AI that creates new text (LLMs), images, and more


1950s
Text Analytics
1960s
Rule Based Systems
1980s
Natural Language Processing
2000s
Machine Learning
2020s
GPT
purpleslate

REPORT / STUDY | Publication 08 April 2019

Ethics guidelines for trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

See also

A European approach to artificial intelligence

Related topics

AI Act

The AI Act is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

The AI Act (Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence)is the first-ever comprehensive legal framework on AI worldwide. The aim of the rules is to foster trustworthy AI in Europe.

The AI Act sets out a clear set of risk-based rules for AI developers and deployers regarding specific uses of AI. The AI Act is part of a wider package of policy measures to support the development of trustworthy AI, which also includes the AI Innovation Package, the launch of AI Factories and the Coordinated Plan on AI. Together, these measures guarantee safety, fundamental rights and human-centric AI, and strengthen uptake, investment and innovation in AI across the EU.

To facilitate the transition to the new regulatory framework, the Commission has launched the AI Pact, a voluntary initiative that seeks to support the future implementation, engage with stakeholders and invite AI providers and deployers from Europe and beyond to comply with the key obligations of the AI Act ahead of time.

Why do we need rules on AI?

The AI Act ensures that Europeans can trust what AI has to offer. While most AI systems pose limited to no risk and can contribute to solving many societal challenges, certain AI systems create risks that we must address to avoid undesirable outcomes.

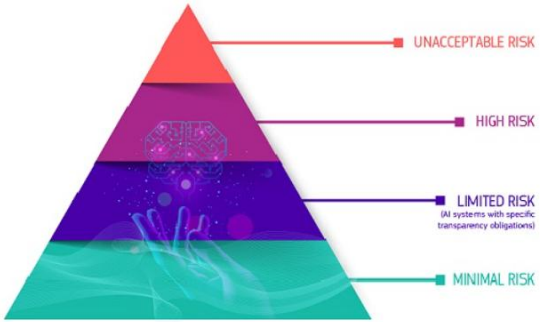
Share

Quick links

- The AI pact
- Text of the AI Act in all EU official languages
- Impact assessment of the regulation
- Study supporting the impact assessment
- FAQs: New rules for Artificial Intelligence
- European AI Office

A risk-based approach

The AI Act defines 4 levels of risk for AI systems:



Unacceptable risk

All AI systems considered a clear threat to the safety, livelihoods and rights of people are banned. The **AI Act prohibits eight practices**, namely:

1. harmful AI-based manipulation and deception
2. harmful AI-based exploitation of vulnerabilities
3. social scoring
4. Individual criminal offence risk assessment or prediction
5. untargeted scraping of the internet or CCTV material to create or expand facial recognition databases
6. emotion recognition in workplaces and education institutions
7. biometric categorisation to deduce certain protected characteristics
8. real-time remote biometric identification for law enforcement purposes in publicly accessible spaces

A Framework for Ethical AI Governance

- Ethical Risk Assessment
  - Integrate ethics review at project inception
  - Conduct a risk assessment to identify potential harms, affected populations, and regulatory compliance requirements
  - Classify projects by risk level to determine the rigor of subsequent reviews and approvals
- Governance Structure
  - Form an AI Ethics Committee comprising cross-functional representatives (engineering, legal, compliance, domain experts, and ethicists)
  - This body reviews high-risk projects, enforces policy adherence, and advises on mitigation strategies
- Lifecycle Integration (next slide)

Lifecycle Integration
Embed ethics checks at key milestones

- Design Phase
  - Define use cases, ethical requirements, and success criteria
- Data Collection & Preprocessing
  - Evaluate data sources for biases and privacy risks
- Model Development
  - Apply fairness constraints and document design decisions
- Validation & Testing
  - Conduct bias audits, adversarial tests, and performance evaluations across subgroups
- Deployment & Monitoring
  - Implement monitoring dashboards for drift detection, fairness metrics, and incident reporting
- Post-Deployment Review
  - Regularly reassess ethical performance and update documentation

Key Ethical Principles

FAIRNESS

TRANSPARENCY

ACCOUNTABILITY

PRIVACY

ROBUSTNESS

Fairness

- Fairness entails ensuring **equitable treatment and outcomes** across demographic groups
- AI systems must be designed to **detect and mitigate biases arising from unrepresentative training data or discriminatory historical patterns**
- Techniques include **pre-processing adjustments** (rebalancing datasets), **in-processing constraints** (fairness-aware learning), and **post-processing corrections** (calibrating outputs to reduce disparate impact)

(Hardt et al., 2016)

1. Data Sourcing

The company obtains data sets from a health care organization to train the MLm component of the system.

- How does the company assess conditions for data use?
- Is the company compliant with data protection rules in its country and in the country of residence of the patients?
- Does the data set include sensitive data that respond to special regulatory provisions, e.g. genetic data?

2. Product Development

The company validates the algorithm against human performance benchmarks.

- Does the training data set accurately represent the source population?
- Is it representative of the target population?
- Can developers explain the logic behind the algorithm?
- What kind of evidence supports the safety and efficacy of the device?

3. Clinical Deployment

The system is adopted in a low-middle income country by a regional health care system that has no clinical genetic services.

- Have doctors been consulted about the introduction of the device in their clinical practice?
- Are patients informed about the use of the system?
- Do patients have equitable access to targeted therapies?
- What monitoring mechanisms are in place to assess the performance of the device in a real-life setting?
- Who is liable in case of a prediction error?

PLOS | MEDICINE

PERSPECTIVE

Machine learning in medicine: Addressing ethical challenges

Effy Vayena[1]*, Alessandro Blasimme[1], I. Glenn Cohen[2]

1 Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland, 2 Harvard Law School, Cambridge, Massachusetts, United States of America

* effy.vayena@hest.ethz.ch

https://doi.org/10.1371/journal.pmed.1002689.g001

Ethical Assessment Tool

UBEP aims to propose an Ethical Assessment Tool (EAT) tailored to evaluate AI systems across key ethical dimensions

Methodology 1/2

- The **UBEP Ethical Assessment Tool (EAT)** is designed to provide a structured and comprehensive framework for evaluating the ethical implications of AI systems
- The tool aims to be practical, scalable, and adaptable to various AI applications
- It incorporates key ethical principles identified in existing frameworks and literature

- The EAT was created as an open access tool
- The tool comprises distinct domains, including **explainability**, **generalizability**, **open data**, **risk of bias**, **impact of wrong predictions**, **transparency**, **data privacy**, **reproducibility**, **fairness**, **accountability**, **stakeholder inclusiveness**, and **financial impact**
- The domains were derived by published guidance for statistics in clinical research
- Each dimension is meticulously defined and operationalized to facilitate a thorough assessment

Development of the EAT

Literature Review

The development process began with a thorough review of existing ethical guidelines and frameworks. This review helped identify the core ethical principles and the gaps that the UBEP tool could address. Key references included guidelines from the European Union, UNESCO, WHO, and recent academic studies on AI ethics.

Stakeholder Engagement

Engaging with stakeholders, including AI developers, ethicists, policymakers, and end-users, was crucial. This engagement ensured that the tool addressed practical concerns and was user-friendly. Workshops and focus groups were conducted to gather input and refine the tool's design.

Framework Development 1/2

The framework comprises several modules, each focusing on a specific ethical dimension:

- Explainability: Assesses the transparency of the model's decision-making process, ranging from complete opacity to full transparency. It focuses on a human-AI interaction, conceptualizing AI mainly as an end-product and humans as being capable of understanding it.
- Generalizability: Measures the model's performance on new data, from poor to excellent adaptation.
- Open Data: Evaluates the accessibility and openness of the data used by the model.
- Risk of Bias: Assesses the fairness and impartiality of the model's predictions.
- Impact of Wrong Predictions: Gauges the consequences of model errors, from severe to negligible.
- Transparency: Evaluates the clarity and openness of the model's design and implementation.
- Data Privacy: Assesses the protection of user data, from insecure to highly secure.

Framework Development 2/2

- Reproducibility: Measures the ease of replicating the model's results, from challenging to straightforward.
- Fairness: Evaluates the model's impartiality across different groups, from consistently discriminatory to completely fair.
- Accountability: Assesses the mechanisms in place to ensure responsible model development and deployment. AI shall be created and operated to provide an unambiguous rationale for all decisions made.
- Stakeholder Inclusiveness: Evaluates the extent to which stakeholders are involved in the model's development, from complete exclusion to full inclusiveness.
- Financial Impact: Assesses the potential financial repercussions of model errors, from severe losses to negligible impact.

Pilot Testing and Validation

The tool was **pilot tested** in several AI projects within UBEP, including applications in healthcare and public health

Feedback from these tests was used to refine the tool further

Validation involved comparing the tool's assessments with those from established frameworks to ensure consistency and reliability

Results 1/3

The tool provides a radar chart with an overall score ranging **from 0 to 1**, depending on the options chosen for the various domains. The radar chart condenses the information derived from each domain into a single polygon, providing a visual representation of the ethical assessment.

- Explainability assesses the transparency of the model's decision-making process. In the pilot tests, models with high explainability scores provided clear and understandable reasons for their decisions, facilitating trust and accountability.
- Generalizability measures the model's performance on new data. Models that scored well in this domain demonstrated excellent adaptation to diverse datasets, ensuring robustness and reliability.

Results 2/3

- Data

Open data evaluates the accessibility and openness of the data used by the model. High-scoring models utilized publicly available datasets, promoting transparency and reproducibility.

- Risk of Bias

The risk of bias assesses the fairness and impartiality of the model's predictions. Models that performed well in this domain had mechanisms in place to detect and mitigate biases, ensuring fairness across different groups.

- Impact of Wrong Predictions

This dimension gauges the consequences of model errors. Models with low impact scores had negligible consequences for wrong predictions, mitigating potential harm to users.

- Transparency

Transparency evaluates the clarity and openness of the model's design and implementation. High-scoring models provided comprehensive documentation and clear explanations of their methodologies.

- Data Privacy

Data privacy assesses the protection of user data. Models that scored well in this domain implemented robust encryption and anonymization protocols, ensuring compliance with privacy regulations like GDPR.

Results 3/3

- Reproducibility

Reproducibility measures the ease of replicating the model's results. High-scoring models provided detailed methodologies and open-source code, facilitating reproducibility by other researchers.

- Fairness

Fairness evaluates the model's impartiality across different groups. Models that scored high in this domain demonstrated consistent performance without discriminatory outcomes.

- Accountability

Accountability assesses the mechanisms in place to ensure responsible model development and deployment. High-scoring models had clear accountability structures and processes for addressing ethical concerns

- Stakeholder Inclusiveness

Stakeholder inclusiveness evaluates the extent to which stakeholders are involved in the model's development. Models that scored well in this domain actively engaged stakeholders, ensuring that diverse perspectives were considered.

- Financial Impact

Financial impact assesses the potential financial repercussions of model errors. Models with low financial impact scores had minimal financial consequences for wrong predictions, reducing the risk of significant losses.

Discussion

Strengths

- Comprehensive approach
- Adaptability
- Stakeholder engagement

Limitations

- The tool's effectiveness depends on the quality of input data and the commitment of stakeholders to ethical principles.
- Need for continuous updates and improvements to keep pace with advancements in AI technology

Comparison with Existing Frameworks

- While **UNESCO's guidelines** provide broad ethical principles, the UBEP tool operationalizes these principles into specific, measurable criteria
- The general frame of the **Ethical Assessment Tool (ETHAS)** that was used in another study was based on two integrated checklists for self-assessment, the **Ethical Assessment Sheet (EES)** and the **Ethical Risk Assessment (ERA)**. Ethical assessment especially when performed in the guise of self-assessment, allows anticipating the critical aspects that can compromise the ethical acceptability of a procedure and intervening before their eventual occurrence could damage the reputation of the whole conversation project and alienate societal support
- The same applies with the **ALTAI tool** that was used in another study, as an interactive self-assessment tool and was implemented for the assessing AI-Bases technologies from SHAPES Pilots
- Another reliable tool is the **Responsible DAI solutions Assessment Tool** that was among the first tools to offer a comprehensive, valid and reliable means to measure the degree of responsibility of DAI health solutions

Future Prospects

- Refining the tool based on **user feedback** and **expanding its applicability to emerging AI technologies**.
- There is also a need for **longitudinal studies** to assess the long-term impacts of AI systems evaluated using the tool.
- Additionally, **integrating the tool with AI development platforms** could streamline the ethical assessment process, making it an integral part of AI development workflows.

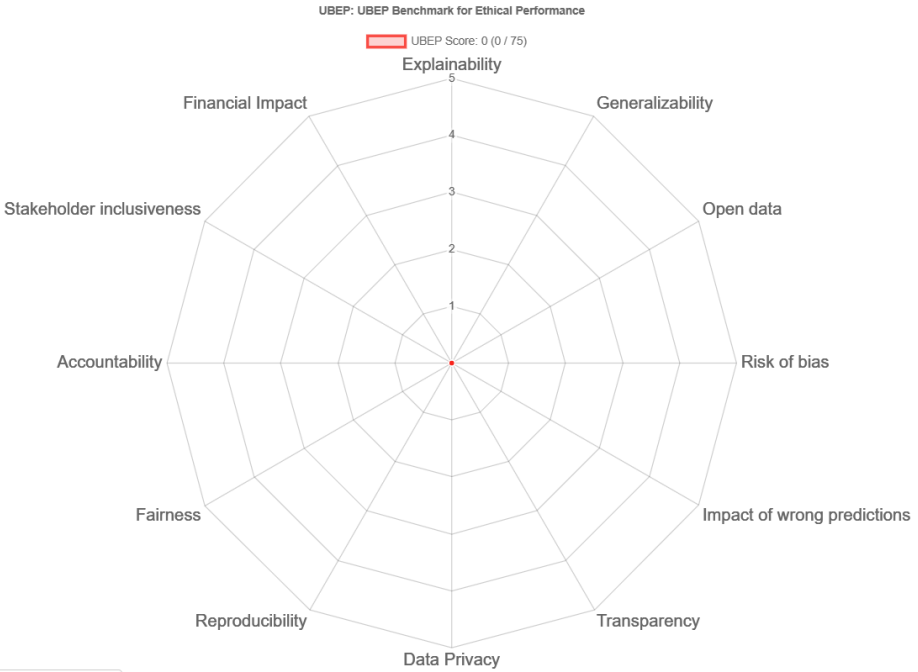
UBEP Ethical Assessment Tool

Explainability

- ○ The functioning of the model is completely hidden, similar to a 'black box' (A deep learning model without any form of interpretability tool.)
- ○ The model provides limited insights, leaving users mostly in the dark (A neural network with rudimentary metrics on the importance of features.)
- ○ The model is semi-transparent and offers some clarity on its decisions (A decision tree too complex to be fully understood.)
- ○ The model is clear on most of its decisions, with only small opaque areas (A random forest model with well-documented feature importance.)
- ○ The model is completely transparent and describes in detail every aspect of its decision-making process (A linear regression model with clear coefficients.)
- ◉ Not evaluated (-)

Generalizability

- ○ The model has extremely poor performance on new data (A cardiac diagnostic tool trained on elderly patients completely fails to identify common heart diseases in pediatric patients.)
- ○ The model struggles with new data, often making errors (A nutrition model trained on data from urban centers underperforms when predicting malnutrition trends in rural areas.)
- ○ The model manages decently with new data, but has occasional glitches (A thoracic model, although trained on a diverse dataset, occasionally misclassifies rare conditions when applied to data from a new epidemiological study.)
- ○ The model fits new data well, with errors being the exception (A vascular model trained on multicentric data successfully classifies most vascular conditions but occasionally struggles with edge cases from certain regions.)
- ○ The model adapts well to all new data it encounters (A pediatric model trained across various global multicentric studies provides accurate predictions for pediatric conditions regardless of the region of data origin.)
- ◉ Not evaluated (-)



Download Radar Chart

Published on 12.07.2023 in Vol 9 (2023)

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/44467, first published November 20, 2022.

Pediatric Injury Surveillance From Uncoded Emergency Department Admission Records in Italy: Machine Learning–Based Text-Mining Approach

Danila Azzolina[1]; Silvia Bressan[2]; Giulia Lorenzoni[3]; Giulia Andrea Baldan[3]; Patrizia Bartolotta[3]; Federico Scognamiglio[3]; Andrea Francavilla[3]; Corrado Lanera[3]; Liviana Da Dalt[2]; Dario Gregori[3]

Abstract

Background:
Unintentional injury is the leading cause of death in young children. Emergency department (ED) diagnoses are a useful source of information for injury epidemiological surveillance purposes. However, ED data collection systems often use free-text fields to report patient diagnoses. Machine learning techniques (MLTs) are powerful tools for automatic text classification. The MLT system is useful to improve injury surveillance by speeding up the manual free-text coding tasks of ED diagnoses.

Objective:
This research aims to develop a tool for automatic free-text classification of ED diagnoses to automatically identify injury cases. The automatic classification system also serves for epidemiological purposes to identify the burden of pediatric injuries in Padua, a large province in the Veneto region in the Northeast Italy.

Citation

Please cite as:

Azzolina D, Bressan S, Lorenzoni G, Baldan GA, Bartolotta P, Scognamiglio F, Francavilla A, Lanera C, Da Dalt L, Gregori D
Pediatric Injury Surveillance From Uncoded Emergency Department Admission Records in Italy: Machine Learning–Based Text-Mining Approach
JMIR Public Health Surveill
2023;9:e44467
doi: 10.2196/44467
PMID: 37436799
PMCID: 10372563

Copy Citation to Clipboard

Export Metadata

END for: Endnote

BibTeX for: BibDesk, LaTeX

RIS for: RefMan, Procite, Endnote, RefWorks

Add this article to your Mendeley library

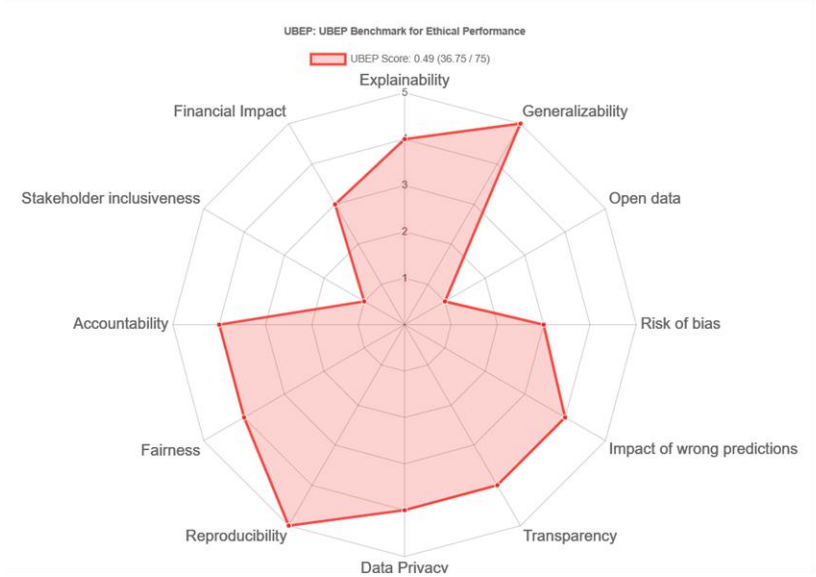
This paper is in the following e-collection/theme issue:

Infoveillance, Infodemiology, Digital Disease Surveillance, Infodemic Management (1087)

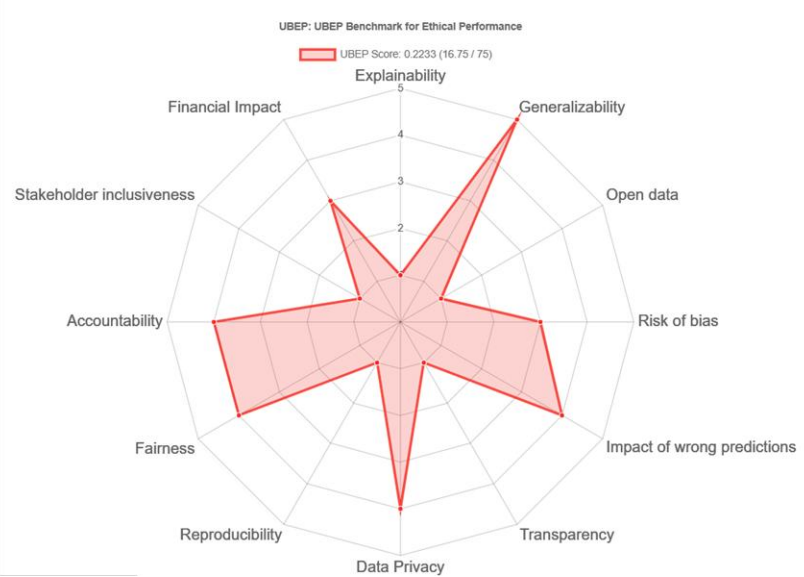
Machine Learning (1730)

UBEP-AE for injury coding

Classical ML AE score 0.49



GPT AE score 0.22



Thanks

Unit of Biostatistics, Epidemiology and Public health
University of Padova
segreteria@ubep.unipd.it

https://www.unipd-ubep.it