Guidance for Clinical Trial design for evaluation of GenAI

Spiros Denaxas

University College London & British Heart Foundation Data Science Center

ESC
European Society of Cardiology

What do we mean by generative AI?

- Generative AI is a **subset of AI**
  - a class of models that learn the joint probability distribution of inputs and outputs - may or may not use deep learning
  - main task: generate new data similar to the training data e.g. text, images, sound etc.
  - autoregressive: a model that generates one piece at a time, using what it just made to decide the next piece.
- **Common clinical tasks:**
  - Document summarization
  - Information extraction or retrieval
  - Discharge letter generation
  - Chatbots
  - Outcome forecasting

Gen AI models compared to traditional clinical prediction models using ML

- **Broadly, Gen AI models are:**
  - less data-complex than traditional clinical predictive model because they require less data fusing, imputation, and feature engineering.
  - less deployment-complex than traditional clinical predictive model, because they enable real-time inference as physicians write notes and require fewer labelled examples.
  - more computational intensive as larger models require significant hardware
  - more storage complex because they require the storage of large amounts of information in a rapidly accessible manner

RESEARCH METHODS AND REPORTING

OPEN ACCESS

Check for updates

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

RESEARCH METHODS AND REPORTING

OPEN ACCESS

Check for updates

TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods

RESEARCH METHODS AND REPORTING

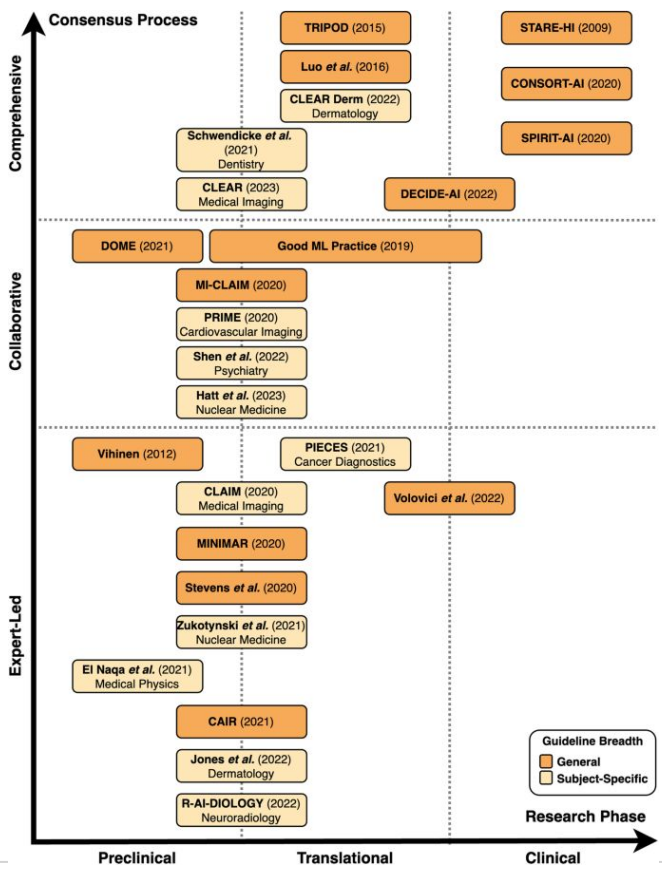
OPEN ACCESS

Check for updates

PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods

(most) AI reporting guidelines are not GenAI-ready!

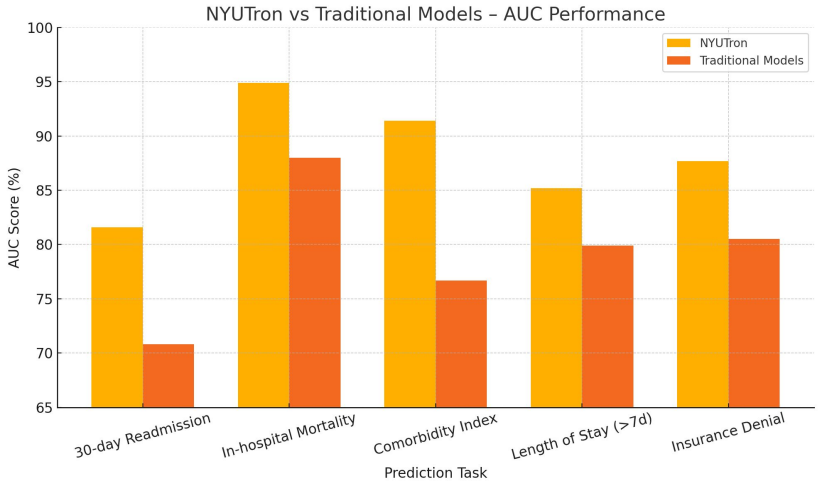
- **Medical AI challenges:**
  - Bias and Fairness
  - Explainability & Interpretability
  - Validation
- **Additional Gen AI specific challenges:**
  - Hallucinations
  - Hyperparameters & prompting
  - Reasoning & Grounding
  - Reproducibility
  - Human interaction / workflow



Article | Open access | Published: 07 June 2023

Health system-scale language models are all-purpose prediction engines

- **NYU Health**
  - Covers 2011–2020
  - Transformer based architecture trained on 387M clinical notes
- **Evaluated using prospective clinical trial**
  - Single arm, prospective
  - Real time inference at discharge of readmission scores
  - Integrated with EPIC
  - Median inference: 0.28sec per patient



How do we evaluate Gen AI ?

- **Clinical validation in real world settings is essential**
  - Clear reporting alone is not enough
  - Clinical outcomes equally important as human metrics

- **Evaluation depends on the use case**
  - Operational applications vs. clinical tasks
  - Do we need separate evaluation designs for each task ?
  - Do all tasks require a trial evaluation?

Are current trials sufficient? Probably not

- **Are trial designs sufficient?**
  - prospective, single-arm trials
  - pragmatic, embedded in standard EHR?
  - New endpoints on human-centered metrics
    - hard to define and measure

- **Challenges**
  - Feasibility equally important as clinical outcomes
  - How do we collect data on human interaction?
  - Many metrics are subjective and hard to quantify

A long way ahead

- **Other open challenges**
  - How do we grade evidence that has been generated by gen AI ?
  - Legal and ethical challenges associated with outputs
  - Subjective: many generative outputs (like writing, summarizing, or recommending actions) require human judgment to assess quality, coherence, and helpfulness.