Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice

Loïc Verlingue,[a,b,] Clara Boyer,[a] Louise Olgiati,[a] Clément Brutti Mairesse,[a] Daphné Morel,[b,c] and Jean-Yves Blay[a]*

[a]Centre Léon Bérard, Centre de Recherche en Cancérologie de Lyon, France
[b]INSERM U1030, Molecular Radiotherapy, Villejuif, France
[c]Department of Radiation Oncology, Gustave Roussy, Villejuif, France

The Lancet Regional Health - Europe
2024;46: 101064

Published Online xxx
https://doi.org/10.1016/j.lanepe.2024.101064

Summary

In this Personal View, we address the latest advancements in automatic text analysis with artificial intelligence (AI) in medicine, with a focus on its implications in aiding treatment decisions in medical oncology. Acknowledging that a majority of hospital medical content is embedded in narrative format, natural language processing has become one of the most dynamic research fields for developing clinical decision support tools. In addition, large language models have recently reached unprecedented performance, notably when answering medical questions. Emerging applications include prognosis estimation, treatment recommendations, multidisciplinary tumor board recommendations and matching patients to recruiting clinical trials. Altogether, we advocate for a forward-looking approach in which the community efficiently initiates global prospective clinical evaluations of promising AI-based decision support systems. Such assessments will be essential to validate and evaluate potential biases, ensuring these innovations can be effectively and safely translated into practical tools for oncological practice. We are at a pivotal moment, where continued advancements in patient care must be pursued with scientific rigor.

Keywords: Artificial intelligence; Natural language processing; Language models; Oncology; Healthcare; Clinical trials

Introduction

Numerous artificial intelligence (AI) applications designed to bolster medical decision-making have recently exhibited remarkable performance. In the field of oncology, there is a particular need of streamlining. Yet, therapeutic decision support is a relatively new application of AI in the field of oncology, which is slow in coming perhaps due to the complexity of the chain of decision-making and the continuous multiplication of therapeutic options. As an illustration, out of 71 devices granted by the FDA in oncology-related fields in 2022, only three could impact treatment decision,[1] all of which being radiotherapy treatment planning systems, i.e. AI-based algorithms used to contour the tumor and organs at risk on imaging data.

Beyond validation, in order to ensure the broad acceptance of a tool among the medical community, its use must be easy and intuitive, and limit prior time-consuming processing such as data structuration. In hospital systems, when browsing through the medical record of a patient, >80% of the number of files are in narrative format. Most of these clinical text files are inherently rich in key information, as they compile multiple patient-centered dimensions such as medical history, ongoing medications, medical examination, analyses of biological and imaging exams and clinical interpretation. Texts are daily used by doctors and caregivers to review a patient's history. It is a matter of course that natural language processing (NLP) underwent a great development in recent years, including for tasks related to therapeutic decision support in oncology.

NLP is dedicated to developing programs capable of integrating, interpreting and generating human language. Beginning in the 1950s with rule-based systems, NLP evolved to incorporate statistical methods like Hidden Markov Models in the 90s, and witnessed a breakthrough with compact word embedding spaces introduced in the early 2000s, popularized by the Word2vec algorithm in 2013, for example. The advent of attention mechanisms in 2017, embodied in Transformer models like BERT and GPT, revolutionized the field. Language Models such as BERT and its derivatives,[2] utilizing stacked encoder layers and attention mechanisms, have shown remarkable performance in tasks like Named Entity Recognition and Text Classification, particularly in biomedical domains.[3–5] Conversely, generative (Large) Language Models (LLMs), designed with a decoder-only architecture, excel in tasks like question-answering (detailed below) but pose challenges due to high computational demands, ethical considerations for the use of protected health information, potential bias, reliability, and interpretability in medical applications (Panel 1). This has prompted the development of strategies like In-Context Learning, Augmented

*Corresponding author. Centre Léon Bérard, Centre de Recherche en Cancérologie de Lyon, France.
E-mail address: loic.verlingue@lyon.unicancer.fr (L. Verlingue).

Panel 1: The dark side of LLMs and potential harm for healthcare applications

1. Environmental impact

Very large models consume substantial amounts of energy and water, contribute to the depletion of materials used in computer equipment manufacturing, and can lead to soil pollution.[54,55] Multi-purpose, generative architectures are orders of magnitude more expensive than task-specific systems for various tasks.[56] As an illustration, today's most energy-intensive generative models consume as much energy per inference as half a smartphone charge. This issue has been known for some time,[57] yet instead of addressing the ecological impact, some authors either avoid disclosing it or downplay its significance.[58]

2. Alienating work

The parceled work of click workers has become indispensable to the development of Large Language Models and the filtering of results. This work is often alienating, conducted under unacceptable and sometimes toxic social and health conditions, and frequently remains invisible to minimize its perceived impact.[59]

3. Disproportionate fashion

In 2022, 40 new LLMs were trained and published. They were more than 200 in 2023.[60] One might question the reasons behind this apparent redundancy in repeatedly training these models from scratch. The political and economic situations of countries can also influence the data types used for the development of large and costly AI systems, with underrepresentation of minority languages, for example.

4. Gender discrimination

Among many similar reports over recent years, a UNESCO report from March 2024 highlighted that gender biases influence the predictions of several Large Language Models, often resulting in sexist outputs. For instance, women are frequently associated with family, children, and home-related roles, while men are more commonly linked to work, pay, and business. Such biases can have significant consequences, including affecting automatic suggestions for postgraduate study options.[61]

5. Demographic origin and other stigmatisations

Bias in natural language processing applications, rooted in factors such as origin, religion, disabilities, can arise from various sources. These includes stereotypes present in the training data, the annotated data used to refine the models, the input representations, the models themselves, and especially the way the research task are defined.[62] As an illustration, Omiye et al. showed that four different commercialized LLMs were propagating race-based medical misconceptions and thus spreading harmful inaccurate content.[63] In another example, an algorithm designed to identify patients who would benefit from specialized referrals and that did not use race as an input to avoid bias still resulted in fewer Black patients being referred compared to White patients with similar disease burdens. This outcome was due to the model using healthcare costs as a proxy, which reflected systemic disparities in access to care.[64] This is not a new phenomenon in language models; even with improved controls over model outputs, recent models still frequently generate discriminative texts.

6. Models' ditractions

Deep learning models can rely on unexpected and sometimes undesired informations in the input data to make their predictions. Many classical examples exists in image analysis. For example a deep-learning method used across three hospitals for pneumonia screening mistakenly focused on text markers on X-rays rather than the pathology itself, leading to misdiagnosis.[65,66]

7. Bias transmission to offspring models

Generalist Large Language Models can be applied to a wide array of tasks beyond mere conversation, either with minimal modifications, through fine-tuning, or by generating synthetic data to train new models. The latter approach is particularly common among healthcare startups, as sharing patient data involves stringent controls and intense procedures. However, this method can inadvertently perpetuate and even amplify existing inequalities and biases learned by the foundation models.

8. Humans under influence

If an AI device finds a role in a repetitive application, it is likely intended to reduce the number of humans required for the task. This raises important questions about how decisions made by the AI can be scrutinized. An intriguing and potentially alarming phenomenon is that humans may internalize the biases present in the model's outputs, even after the AI is no longer in use, as seen in applications like image-based diagnosis.[67]

9. Challenge for reproducibility in AI

Deep learning methods inherently use non-deterministic tools to enhance their efficiency and learning speed. As a result, even with identical inputs (data and parameters), an algorithm can produce models with significantly varying accuracy and convergence speeds.[68] Moreover, the underlying libraries, such as TensorFlow (an open-source machine learning tool developed by Google), can introduce additional variance. This variance arises from parallelization, which can lead to a variable order of operations and can lead to results that is not reproducible even by its author(s).

10. Impossible evaluation

Evaluations are most often carried out retrospectively on small datasets that may not accurately represent the language, demographic and epidemiological characteristics of patients encountered in clinical practice. The confidentiality of health data and major constrains on data sharing further limit the possibility to assess the generalization performance of the language models. Moreover most public benchmark narrative datasets have recently become saturated, either because Language Models were trained on these datasets, or models are overfitting it by the number of times these datasets were used for benchmark.[42]

11. Hustling the physican-patient relationship

Decisions that lean heavily on AI recommendations can significantly impact the shared medical decision-making. Biases can emerge not merely because of biased datasets or algorithms, but because of factors involved in real-world implementation: Clinician-, patient-, and social-level factors can interact to create biases in the adoption of AI for clinical decision support.[69,70]

Language Models or multi agentic workflows to enhance their performance and mitigate biases.[3,6] Recent efforts explore avenues for improved efficiency and broader data analysis capabilities, including integration with other data formats like images. Despite advancements, the energy-intensive nature of LLMs also remains a concern, even for inference in cloud computing environments.[7]

Facing the growing demand, the trajectory of language processing-based medical decision support systems is currently fasting forward, but its progress towards regulatory approval is conditioned by their comprehensive understanding and their thorough evaluation by the caregiving community, with the objective of gaining the trust of all healthcare actors.[8] A carefull evaluation strategy is necessary to mitigate bias and potential harm throughout the AI algorithm's lifecycle (Panel 2). However, initiating this process can be challenging at this stage. In this Personal View, we present a classical evaluation approach used in medicine, which includes retrospective studies, non-interventional and interventional prospective studies, and multicenter evaluations, all of which could be applied to the most promising tools. We begin by exploring applications of language models in oncology, followed by an examination of the evaluation methods and their purposes. Finally, we outline strategies for prospective clinical evaluations, including the design of clinical trials.

Medical question answering: [medical student in training may ask] "what is the treatment of pulmonary embolism?"

Medical question answering serves as a prime example of the remarkable progress seen in NLP models over the past 15 years. It started with the public release of medical question answering databases such as the United States Medical Licensing Examination.[9] As NLP models evolved in architecture, size, and sophistication, their performance on medical tasks improved significantly. The best performing models to date achieved unprecedented test accuracy of 90.2% by the end of 2023[10,11] represented by generative language models, particularly LLMs, equipped with advanced prompting methods. Prompt engineering involves designing task-specific instructions to guide model outputs without altering the model parameters.[12] Recent promising methods in prompt engineering include few-shot, chain-of-thought, self-consistency, and ensemble refinement, achieving accuracy rates of [80–86%] on MedQA (USMLE) database.[13,14] As an illustration, the Chain-of-Thought prompting technique allows a large language model (LLM) to base its output on its own intermediate reasoning steps by simply adding 'Let's think step-by-step' at the end of the prompt. Additionally, a useful practice is to specify the context for the model by including a description such as 'You are a medical doctor answering real-world medical entrance exam questions.' Prompts are generally relatively long (with the paragraph headings of this manuscript not being typical prompts). Moreover, prompt engineering is a rapidly evolving field and is highly dependent on the specific LLM being used.

Even with the best prompting technics, LLMs' could have the tendency to produce incorrect responses resembling authentic ones, termed hallucinations, which may undermine clinical applications. For instance, to the question "what are the common side effects of metformin", a LLM may respond, "metformin side effects include nausea as well as trouble breathing", which is partly incorrect. The intricacies of clinical decision-making, which includes patient presentation and preferences, stepwise diagnostic processes, treatment planning based on evolving guidelines and reporting, creates significant challenges for answering medical questions in actual clinical environments. This is in contrast to the more straightforward scenarios presented in hypothetical patient vignettes or clinical case challenges. When applied to real patient data (derived from the MIMICS dataset), five LLMs consistently fell short of matching the diagnostic accuracy of clinicians across four urgent abdominal pathologies. Their accuracy further declined when required to independently gather diagnostic information, and they failed to consistently adhere to treatment guidelines.[15] It clould pose potential serious risks to the health of patients. Altogether, bridging the gap between medical question answering in medical exams and real-world clinical practice remains a significant challenge for LLMs applications. It is clear that dedicated real-world evaluations and adptations are crucial before these tools can be responsibly implemented in clinical settings.

Natural language processing for prognosis estimation: [clinical trial investigator may ask] "what is the predicted life expectancy of this patient given [...]?"

In oncology, therapeutic options often hinge on life expectancy considerations and prognosis estimation is often the first step towards treatment recommendation. Prognosis prediction in cancer care is useful to spare patients from intensive procedures with limited

Panel 2: Checklist to increase the bright side of language models in healthcare

These recommendations can help ensure that a project using language models for healthcare is well-structured, ethical, and effective in achieving its goals.

1. Data collection and cleaning: addresses Panel 1 *Gender discrimination, Demographic origin stigmatization, Models' distraction and Impossible evaluation*
Data quality is a cornerstone of any data science project. This is a broad and complex topic that applies to multiple types of projects. If the dataset is private, early discussions with the Informatic Team is highly recommended. For public or synthetic data especially if generated by a foundation Large Language Model, it is important to verify and address potential biases in the data. Debiasing is an intense filed of research and several approaches exist.[62] Finally, when data is in a structured format and some values are missing, one can consider imputation techniques or use models that can handle missing values effectively.

2. Model training: addresses Panel 1 *Environmental impact, Disproportionate fashion* and *Bias transmission to offspring models.*
Many educational materials exist on how to train a machine learning model and select the appropriate model architecture, among other considerations. Intense research continues with the objective to improve speed, efficiency, reduce energy consumption, among other factors. But in the era of Large Language Models, the primary question should be: « Do I really need to train a new model?». One solution could be to use a model pre-trained on the data of interest, and prompt it appropriately. Another option is to fine-tune a pre-trained model, but you should evaluate whether you have sufficient computational resources and consider the quality (and biases) of the foundation model you choose.

3. Selection of the outcome: addresses Panel 1 *Alienating work* and *Impossible evaluation*
You may want your model to predict something useful for the clinic. Obtaining labels can be challenging, either due to the considerable effort required or regulatory constraints. A statistician can provide you with guidance on the type of label data you have: survival data, continuous or categorical data, text (as for generation)—as well as the methodological challenges, such as competitive biases. Additionally, consider the impact of the predictions, how humans will use it, and whether you have sufficient data considering the complexity of the task and the size of the model you intent to use.

4. External validation: addresses Panel 1 *Challenge for reproducibility in AI*
A longstanding rule (but worth repeating) in predictive tasks is to split your data into training, validation and testing sets. Validation in machine learning serves as a proxy for generalization performance. The true evaluation of the generalization performance—i.e. testing how your model behave in real-world scenarios—requires new, ideally external data, and continuous monitoring.

5. Data sharing: addresses Panel 1 *Challenge for reproducibility in AI* and *Impossible evaluation*
To evaluate the generalization performance, it is crucial to use data that the models have not been trained on, ideally sourced from a different hospital if relevant to the task. Additionally, sharing data with others can help reproduce our work, promoting transparency and reproducibility. However, data sharing comes with ethical challenges, particularly the protection of patient privacy. Ensuring the privacy of patient data is essential for maintaining a trust-based relationship between caregivers and patients.

6. Federated learning: addresses Panel 1 *Challenge for reproducibility in AI* and *Impossible evaluation*
Data sharing may be impossible for various reasons, such as challenges in fully anonymizing data, privacy concerns, and technical requirements, among others. Federated learning infrastructure can help to train or evaluate models across different decentralized data sources while preserving privacy. Federated learning can nevertheless be complex due to technical requirements (such as installing and managing the infrastructure), interoperability of the data, computational resources, and monitoring of the process.

7. Generalization performance and evaluation: addresses Panel 1 *Models' ditractions, Challenge for reproducibility in AI* and *Impossible evaluation*
Evaluating generalization performance involves assessing how well the model performs on unseen data to ensure it extends beyond the training set. This helps verify that the model can make accurate predictions on new, real-world examples and is not merely overfitting to the training data. Effective generalization is critical for ensuring that the model maintains its utility and reliability in practical applications. Generalization evaluation is also the good moment to evaluate potential biases in the model such as language formatting, regional or cultural influences, among others.

8. Deployment of the tools: addresses Panel 1 *Humans under influence and Hustling the physican-patient relationship*
This refers to elaborating a plan for the practical deployment of a model, including integration into existing systems and user training. It is the good moment to plan the monitoring and how it integrates into practice, how it can affect the physician-patient relationship, among other considerations.

9. Calibration drifts: addresses Panel 1 *Models' ditractions*
Calibration drifts require regular monitoring and adjustment of model predictions to maintain accuracy over time, particularly after deployment.[71]

10. Monitoring for dataset shifts: addresses Panel 1 *Models' ditractions* and *Impossible evaluation*
Monitoring for dataset shifts involves regularly checking for changes in data distributions that could impact model performance. It is crucial to identify these shifts promptly and update the models accordingly to maintain their accuracy and relevance.[72]

11. Master regulatory requirements: addresses Panel 1 *Challenge for reproducibility in AI, Impossible evaluation* and *Hustling the physican-patient relationship*
see Panel 3 and look for the help of experts.

12. Ecological impact of your project: adresses Panel 1 *Environmental impact*
You can learn about the environmental impact of the tools you intent to use, make a Life Cycle Assessment, and consider frugal AI, for example.

13. Evaluate impact on patient-physician relationship: addresses Panel 1 *Hustling the physican-patient relationship*
It aims at reducing biases in how clinicians and patients use AI-based algorithms. To mitigate potential biases, physicians must emphasize patient-centered communication, ensuring AI's role remains that of a support tool rather than a decision maker. Human supervision of AI is a good practice, for example in image analysis application, but for NLP, human-aware supervision is preferable.[70,73]

probability of clinical benefit when the prognosis is poor. It is also a valuable indicator of whether or not a patient could be a good candidate to enter a clinical trial.[16]

Caregivers are known to overestimate their patient's life expectancy, which could paradoxically negatively impact the quality and the quantity of the end of life. As a baseline reference for 'human' performance, expert oncologists are expected to accurately predict the survival at two years of patients affected by cancer with an Area Under the ROC Curve (AUC) ranging between 72% and 81% (with AUC values of 100% indicating perfect prediction and 50% indicating random prediction), which suggests that their prognosis is wrong in about one case out of four.[17,18] In comparison, the NLP model NYUTron, trained and tested on over 7.2 million clinical notes from >380,000 patients, demonstrated reliable performance in predicting in-hospital mortality (AUC of 94.9%) and 30-day readmission risk (AUC of 79.9%), potentially surpassing human estimation.[19]

Importantly, while NYUTron excelled in most medical specialties, its efficiency in estimating cancer prognosis was lower (overall AUC of 63.8% for oncology patients versus 90.1% for neurology and 67.9% for internal medicine excluding oncology), highlighting the need for specialized oncology-specific NLP models. For example, our team developed K-memBERT, a French NLP model trained solely on text data from 2053 K deceased cancer patients, achieving superior 3-month survival predictions (AUC of 85% on an unseen test cohort and 87% on an external dataset) compared to NYUTron for oncology patients.[20] This trend suggests

Panel 3: Regulations on IA

In Europe, AI tools intended for clinical use are classified as *medical devices* and fall under the regulatory oversight of the EU AI Act and the CE Marking. After intense negotiations, the world's first legislation on artificial intelligence (AI) was definitively adopted by the European Union (EU) on 21 May 2024. The objective of this Regulation is 'to improve the functioning of the internal market and to promote the uptake of human-centred and trustworthy artificial intelli gence (AI).' (Article 1 of the AI Regulation). Tthe EU AI Act follows a 'risk-based' approach, categorizing AI systems into four groups: systems presenting an unacceptable risk, which must be prohibited; systems presenting a high risk and requiring control measures adapted to these risks; 'AI models for general use' and, among them, those presenting a 'systemic risk'. These systems are subject to specific control measures, which may be added to the measures to be taken if the system is also high-risk; and other systems which are not subject to specific control procedures. The only measure in this regulation that applies to all authorized systems (including those classified as risk-free) is a simple invitation to '[take] measures to ensure, as far as possible, a sufficient level of AI proficiency for their staff and other persons involved in the operation and use of AI systems » (Article 4), as well as an encouragement to "draw up codes of conduct […] designed to promote the voluntary application' of the requirements applicable to high-risk systems. The CE mark is mandatory for the clinical use, marketing and/or commercialization of these systems. Two years ago, the EU implemented new regulations pertaining to medical devices with clinical application. These regulations, in addition to enforcing a stringent and controlled development approach, mandate clinical trials, including those involving multi-center cohorts. From a technical standpoint, these regulations necessitate, among other requirements, comprehensive planning and documentation of all code development phases (including details such as personnel involved, methods employed, data sources, etc.), clear differentiation between training and validation cohorts, and obtaining authorization for all clinical validation phases, mirroring the procedures of drug clinical trials. Additionally, guidelines are emerging on how to design protocols of prospective evaluations of AI applications in the clinic (SPIRIT-AI) and how to report their results (CONSORT-AI).[74,75]
Altogether, it is worth noting that such evaluations can be expensive and, in some cases, challenging to implement, particularly for systems like clinical trial matching tools. A typical scenario might involve randomizing patients into two groups: one that uses clinical trial matching tools and one that does not (relying on standard procedures to access clinical trials), with overall survival comparison as the primary endpoint. However, this approach may be unethical, as it involves withholding potentially beneficial information from the control group about therapeutic options. An alternative approach could involve using historical controls, comparing recruitment rates in clinical trials before and after the intervention. While this method seems straightforward, it carries potential selection and historical biases. We are currently at the stage of prospective non-interventional evaluation of trial matching tools, and innovative designs are required to go to interventional evaluation to seek for reglementary approval, conditioned by the demonstration of a positive impact on patients.

that model performance may improve with training specificity, akin to physicians' expertise in specialized areas. Nevertheless, the external comparison of studies should be carefully interpreted as each study used different sources of data, and in some cases, proprietary data that precludes reproducibility checks with external researchers. In this regard, the generalization performance of models across hospitals is often very difficult to evaluate, meanwhile it is mandatory for any distribution of the tools. Collaborative federated learning, which utilizes multiple independent datasets to train or evaluate a common algorithm, offers an approach to address this challenge in health data applications.[21] This advocates for propsective and multi-centric evaluation of AI devices.

Therapeutic decision support systems including: 1) treatment effect prediction: [medical oncologist may ask] "what is the predicted benefit of maintenance olaparib for this patient, given [...]?" and 2) treatment recommender systems: [medical student in training may ask] "what treatment(s) should I prescribe for a patient with cough, fever and gram-positive bacillus in sputum?"

Once the disease is precisely diagnosed and the prognosis is estimated, oncologists envisage treatment among a wide variety of modalities including surgery, radiation therapy, cancer-directed systemic therapies (for example, chemotherapy and targeted therapies) and co-medications (for example, antiemetic agents and painkillers). All of these fields could benefit from the development of decision support applications using AI with the aim of saving time and energy for tasks that are more 'humanly-relevant', possibly with significant benefit in terms of survival and quality of life for patients (Fig. 1).

Two main ways to address medical decision assistance can be distinguished: 1) treatment effect prediction: "what are the chances of success of this precise strategy in this context?" and 2) treatment recommendation: "what is the best treatment option in this situation?". Depending on the clinical situation, one approach can be favored over the other. For example, when radiotherapy is planned, the act itself (tumor and organs at risk contouring) is prescribed directly onto the patient's imaging data, which is a task that can be highly accurately performed by AI as a treatment recommendation proposal and directly plugged in the treatment planning systems that radiation oncologists already use in routine.[22] Conversely, systemic therapy decisions often require broader information of different natures (imaging, clinical, biological, tumor molecular data, patient's wish etc.) and can largely benefit from both treatment recommendation and individualized treatment effect predictions.

In the clinic, personalized treatment effect prediction consists of AI-based estimation of individual outcomes of patients on a specific treatment. The system would rely on clinical context, genomic information and complementary exams. Most existing treatment effect prediction approaches are composed of shallow machine learning tools requiring structured clinical variables or distribution based analysis that utilize gene signatures requiring molecular data.[23] Some deep learning models were published to predict specific treatment effect, such as FOLFOX regimen in patients with colorectal cancer, adjuvant chemotherapy for patients with breast cancer, or immune checkpoint inhibitors, often with limited number of patients but sometimes with extensive data types (histopathological slides, multi-omics data, and even multiple data integration).[24,25] Interestingly but rarely, some models were trained to predict the onset of treatment-related adverse events.[26,27]

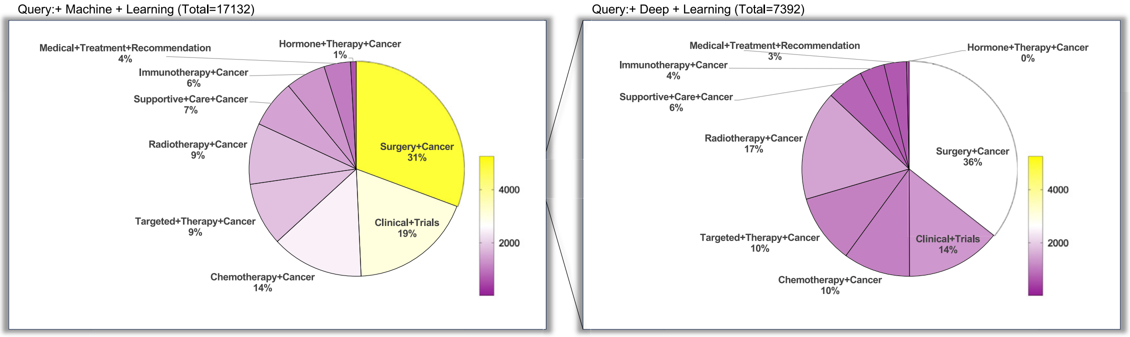


Fig. 1: Repartition of the number of machine learning methods applied to oncology-related tasks (left panel), as per PubMed query (assessed in February 2024). The right panel specifies the repartition of methods utilizing deep learning. MeSH terms that were used for each query are specified in bold with the related fraction. The color legend indicates the absolute count of results per query.

Conversely, the concept of treatment recommendation systems is more complex as they aim to output the best treatment among all possible options upon reviewing the input data, which could be of any kind (structured or free text, biological exams, imaging, etc). The main limitation of recommendation systems trained on retrospective medical data is that they may become quickly obsolete in rapidly-evolving fields such as oncology. Treatments should be chosen according to constantly-evolving guidelines that follow up-to-date evidence-based-medicine demonstrations.[15] These systems may lack the ability to guide treatment over novel authorized options. To address this issue, one approach is to develop tools that can learn from recent clinical data and/or incorporate high-impact literature, as seen with Retrieval-Augmented Language Models.

In oncology, there is a long history of systems mimicking therapeutic recommendations of multidisciplinary tumor boards. The main advantage of such systems is that they are often trained on both hospital and literature data. The most visible and well-evaluated examples are Watson for oncology and CSCO AI in China. Two meta-analyses showed performance that ranged between 74% and 81% of concordant recommendations compared to tumor boards.[28,29] The performances varied depending on the medical specialty, the type of treatment and the treatment line, requiring either refining of the predictive models or defining the clinical situations where it can be applied for prospective, comparative and interventional evaluation.

Clinical trial matching systems in oncology: [clinical trial investigator may ask] “does this patient’s profile […] matches with accessible ongoing clinical trials given that […]?”

In cancer care, participation to a clinical trial is a way for patients to access therapeutic innovation and increase their therapeutic options. Yet, there is a wide discrepancy in participation rates across cancer centers and less than 8% of all patients participate in a clinical trial worldwide.[16] The trial refereeing process depends mainly on relationships and knowledge shared between doctors, and thus often confined to one hospital.[30] NLP can be useful to automatically match patients’ profiles with clinical trials according to study’s selection criteria. If generalized, these tools could drastically improve patients’ access to clinical trials even beyond their local cancer center.

Automatic clinical trial matching approaches in oncology have shown remarkable retrospective performances with on average 90.5% sensitivity and 99.3% specificity according to a meta-analysis published in 2023 that scanned over 50,000 patients from 19 datasets.[31] Nearly all studies used a list containing expert annotated pairs of patient/trial on retrospective cases. Most approaches relied on structured clinical and trial data, and few of them used advanced Language Models. In information retrieval applications (i.e. evaluating the performance of search engines), common performance metrics include Mean Average Precision (MAP, using binary relevance) and Normalized Discounted Cumulative Gain (NDCG using graded relevance). Our team has prospectively evaluated the performance of four web-based trial matching tools on sequential patients from the Molecular Tumor Board of our institution. We found a drop in performance, with a mean MAP@5 (i.e. for the first 5 results) of 0.51 (SD 0.47) and a mean NDCG@5 of 0.44 (SD 0.43). These results underscore the need for prospective evaluation and improvement of these tools.

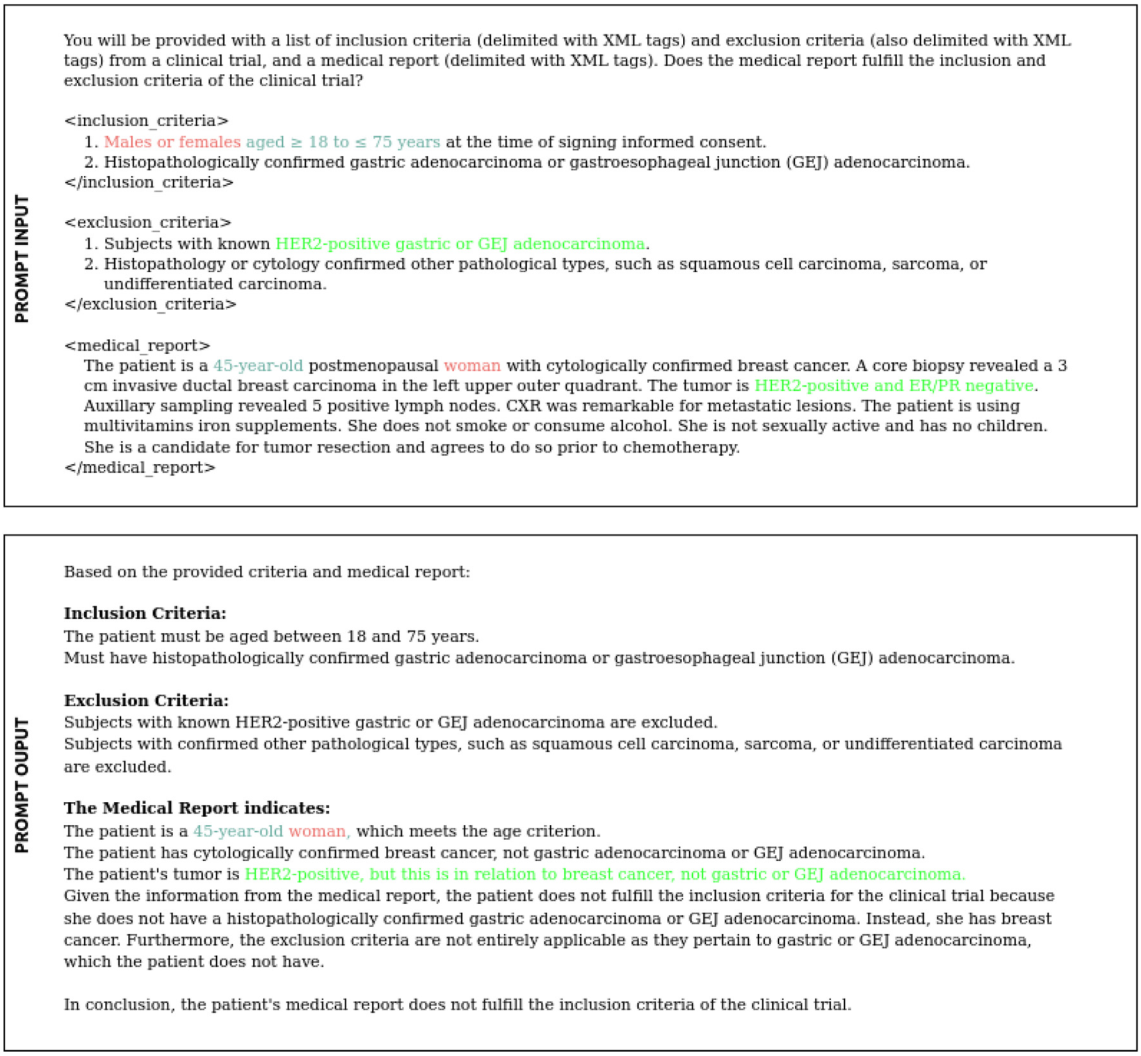
Without great surprise, Large Language Models have recently entered the arena. The Text REtrieval Conference (TREC) 2021 and 2022 released two public clinical trial matching datasets.[32,33] Initial NDCG@10 performances ranged around 0.6125 (i.e. the ranking is 60% right compared to the ideal ranking), then trialGPT improved it to an NDCG@10 of 0.82, and more recently the best performing models at SemEval F1 Scores of 0.8 (but unfortunately without NDCG scoring available).[33–35] All performed on retrospective, and synthetic data, lacking prospective real world evaluation for now. Compared to structured data, Large Language Models for trial matching can limit the data curation burden, can virtually review all selection criteria of clinical trial protocols, can allow human interpretation of the model reasoning, and motivates multiple methodological improvements (Fig. 2).

Beyond trial matching, AI can be valuable to estimate whether a patient will remain long enough in the trial. This can limit the risk of early-discontinuation and improve the quality of the data to allow robust clinical trial conclusions. Few examples exist but a Language Model successfully predicted screening and dose-limiting toxicity period completion with an AUC of 0.8822.[36]

When mature enough, these approaches will require prospective and multicentric evaluation to confirm their generalization performance, utility and impact for patients.

Evaluation of AI applications in healthcare

Numeric medical device development in healthcare faces several challenges, the major being the including non-reproducibility of results.[37] For language models in healthcare, evaluation on external datasets is highly recommended to evaluate the impact of multiple possible undesirable biases (Panel 1). Evaluation in real-world setting is important in technologies where machines operate with varying degrees of self-governance, such as self-driving cars, in dynamic environments with multiple stakeholders.[38] This issue has pushed the community toward data and model sharing to confirm observations and analyze further models’ performances.



You will be provided with a list of inclusion criteria (delimited with XML tags) and exclusion criteria (also delimited with XML tags) from a clinical trial, and a medical report (delimited with XML tags). Does the medical report fulfill the inclusion and exclusion criteria of the clinical trial?

```
<inclusion_criteria>
  1. Males or females aged ≥ 18 to ≤ 75 years at the time of signing informed consent.
  2. Histopathologically confirmed gastric adenocarcinoma or gastroesophageal junction (GEJ) adenocarcinoma.
</inclusion_criteria>

<exclusion_criteria>
  1. Subjects with known HER2-positive gastric or GEJ adenocarcinoma.
  2. Histopathology or cytology confirmed other pathological types, such as squamous cell carcinoma, sarcoma, or
     undifferentiated carcinoma.
</exclusion_criteria>

<medical_report>
  The patient is a 45-year-old postmenopausal woman with cytologically confirmed breast cancer. A core biopsy revealed a 3
  cm invasive ductal breast carcinoma in the left upper outer quadrant. The tumor is HER2-positive and ER/PR negative.
  Auxillary sampling revealed 5 positive lymph nodes. CXR was remarkable for metastatic lesions. The patient is using
  multivitamins iron supplements. She does not smoke or consume alcohol. She is not sexually active and has no children.
  She is a candidate for tumor resection and agrees to do so prior to chemotherapy.
</medical_report>
```

Based on the provided criteria and medical report:

Inclusion Criteria:
The patient must be aged between 18 and 75 years.
Must have histopathologically confirmed gastric adenocarcinoma or gastroesophageal junction (GEJ) adenocarcinoma.

Exclusion Criteria:
Subjects with known HER2-positive gastric or GEJ adenocarcinoma are excluded.
Subjects with confirmed other pathological types, such as squamous cell carcinoma, sarcoma, or undifferentiated carcinoma are excluded.

The Medical Report indicates:
The patient is a 45-year-old woman, which meets the age criterion.
The patient has cytologically confirmed breast cancer, not gastric adenocarcinoma or GEJ adenocarcinoma.
The patient's tumor is HER2-positive, but this is in relation to breast cancer, not gastric or GEJ adenocarcinoma.
Given the information from the medical report, the patient does not fulfill the inclusion criteria for the clinical trial because she does not have a histopathologically confirmed gastric adenocarcinoma or GEJ adenocarcinoma. Instead, she has breast cancer. Furthermore, the exclusion criteria are not entirely applicable as they pertain to gastric or GEJ adenocarcinoma, which the patient does not have.

In conclusion, the patient's medical report does not fulfill the inclusion criteria of the clinical trial.

Fig. 2: Trial Matching example by prompting the Large Language Model GPT4-turbo from https://chat.lmsys.org/. The medical report and synthetic patient profile was extracted from SemEval 2024, and selection criteria of the trial extracted from clinicaltrials.gov.

In their work to curate a catalog of existing datasets about clinical and biomedical NLP, Blagec et al. identified 450 unique datasets, of which less than 30% were publicly available, and only 10% were deemed directly relevant to clinical tasks. Furthermore, among seven main applications to clinical care, only three matched to relevant benchmarks in the catalog.[39] In this setting we can acknowledge GDPR-compliant data-sharing initiatives including MIMICS, TREC, and SemeVal, among others.[32,33,40]

The testing of reproducibility and the analysis of potential bias and model performances are essential for the evaluation of language models. LLMs produce text that are reflections of their training data and thus could perpetuate biases for example pertaining to race, sex, language, or culture and may negatively impact predictions in healthcare applications[8,41] (examples in Panel 1). In 2007, Aravind Joshi proposed the analogy that datasets are the telescopes of data science. Telescopes for languages models would analyze their representation of the data they have been trained on. Private LLMs are trained and validated on non-publicly available demographic data, and access to narrative clinical data in oncology remains a great challenge, and thus very limited. The community favored applications derived from publicly available educational healthcare datasets, which are inherently anonymous, such as MedMCQA.[14] However, most public benchmark datasets for language models have recently become saturated, with models outperforming human performance (either because Language Models were trained on these datasets, or models are overfitting it by the number of times these datasets were used for benchmark).[42] The difficulty in accessing clinical data, the scarcity of comprehensive oncology benchmarks, and the saturation of public educational medical datasets highlight the need for direct model evaluation for specific applications. In other words, initiating dedicated clinical evaluation of language models, ideally prospective.

Prospective clinical trials of AI applications

Few recent studies have proposed a non-interventional prospective comparison of the performance of NLP tools to that of physicians.[19,43,44] In these scenarios, the application does not impact clinical decisions. Non-interventional comparative analyses have been used for FDA approval and CE marketing, for instance for automatic breast cancer screening from mammography.[45] Two studies published in 2024 evaluated the performance of Large Language Models for diagnostic and treatment recommendations. One study focused on 11 cases of patients with gastrointestinal cancers and reported optimistic results.[46] The other study, which examined 2400 patients with abdominal emergencies (appendicitis, cholecystitis, diverticulitis or pancreatitis), highlighted significant concerns due to critical errors of the LLMs.[15]

Even though non-interventional studies are valuable elements to gain the trust among users, in medicine, prospective interventional and comparative studies are the gold standard for validating the performance and the clinical impact of a device or a treatment. Prospective evaluations in real-world clinical settings are important to assess and manage bias in AI models to support fair and equitable development[47,48] and to estimate their performance, usability, and impact on patient outcomes. This process enables to ensure the safety, reliability, and effectiveness of these systems before their deployment in clinical practice. For instance, a post-marketing prospective interventional evaluation has recently confirmed the usefulness of AI-aided breast cancer screening alongside standard procedure.[49]

A search in clinicaltrial.gov using 'artificial intelligence cancer' and filtering for 'recruiting' and 'interventional', in January 2024, returned 86 studies including 27 (31%) using a randomized design. The majority focused on imaging data (50% of studies

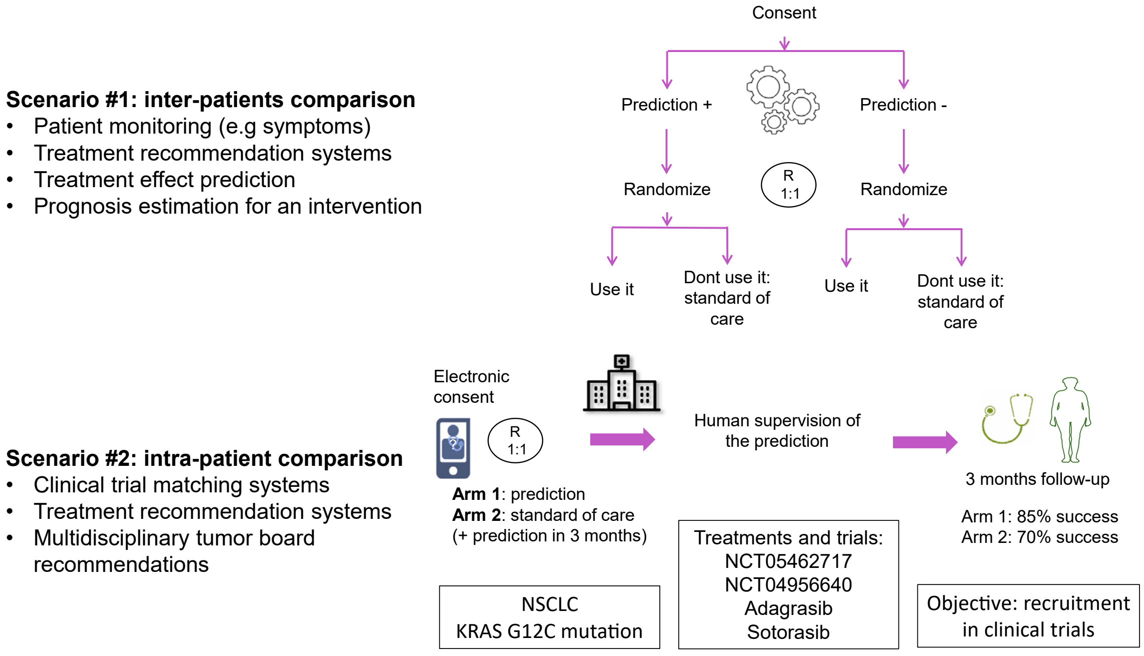

Scenario #1: inter-patients comparison
• Patient monitoring (e.g symptoms)
• Treatment recommendation systems
• Treatment effect prediction
• Prognosis estimation for an intervention
Consent
Prediction +
Prediction -
Randomize
R
1:1
Randomize
Use it
Dont use it: standard of care
Use it
Dont use it: standard of care
Scenario #2: intra-patient comparison
• Clinical trial matching systems
• Treatment recommendation systems
• Multidisciplinary tumor board recommendations
Electronic consent
R
1:1
Arm 1: prediction
Arm 2: standard of care
(+ prediction in 3 months)
Human supervision of the prediction
3 months follow-up
Arm 1: 85% success
Arm 2: 70% success
NSCLC
KRAS G12C mutation
Treatments and trials:
NCT05462717
NCT04956640
Adagrasib
Sotorasib
Objective: recruitment in clinical trials

Fig. 3: Examples of AI-stratified prospective clinical trial designs with randomisation. Scenario#1: randomization to assign patients to use or not the predictions of the tool, inspired from.[53] This applies to situations such as patients monitoring (e.g. using automatic altert systems versus standard procedures), orientation to a specific interventions or another (e.g. orientation to targeted therapies based on the presence of a biomarker), or evaluate a new prognosis tool to decide for an intervention or not (e.g. a surgery, a new treatment line, a clinical trial). This designs allows to derive clear conclusion of the superiority of the prediction compared to the standard of care. Scenario #2: randomization obtain the information now or in a pre-defined periode of time, i.e. intra-patient comparison, also known as N-of-1 trials. In practice, when AI models are used as search engines for medical treatments or clinical trials, it seems unethical to spare patients from the information. It can apply to several clinical decision support systems, to the evaluation of the efficacy of a treatement (using overall reponse rate as primary endpoint), to the impact on the work load of careguivers, the quality of life of patients, among other criterias. Patients that are randomized in the standard of care arm, can have acess to the results of the system after an observation period, here 3 months. The comparison of the endpoint measures between the two period of time allows concluding on the impact of the intervention.

Search strategy and selection criteria

The search and selection criteria used were hybride. As per 2023, 17K papers on Large Language Models were published in Arxiv; we thus could not be totally exhaustive in our search method. We have extracted the most significant messages, derived from our experience and literature watch in the field of Natural Language Processing for Oncology during the last 5+ years, to illustrate our Personal View article. The reviewing process spanned over 6 months and influenced significatively the literature search and the topics presented.

containing 'image' in the description), 24% were surgery-related, 10% radiotherapy-related, and only three studies included NLP methods. Few publications report on the randomized evaluation of antitumor treatment predictions.[50,51] There is a clear need to prospectively evaluate the performance of NLP therapeutic decision support systems in medicine and for us, in oncology.

Interventional prospective evaluations for medical devices demonstrate how AI interventions can enhance existing procedures, such as patient monitoring during radiotherapy,[50] selecting adjuvant treatment for localized RH + HER2-breast cancer,[51] or breast cancer creening with mammographies.[45] These studies require a multidisciplinary team, including medical investigators, engineers, biostatisticians, regulators, ethicists, and patient representatives. Their design and execution must consider specific patient recruitment, data collection methods, and outcome measures tailored to the tool being evaluated and its potential clinical impact. Randomized participant recruitment allows for comparison between intervention and control groups, with preplanned statistical analyses revealing the intervention's impact on outcomes like adverse event rates, treatment completion, and survival. Additionally, prospective interventions can assess the tool's adoption probability in the medical field, examining its utility, impact, logistical requirements, and effects on caregiver workload.[52] We can imagine various designs, inspired by biomarker-based clinical trial designs, for the evaluation of more recent devellopements such as prognosis estimation, treatment recommendations, multidisciplinary tumor board recommendations and matching patients to recruiting clinical trials[53] (Fig. 3).

Conclusion

The realm of language models is bursting with promising experimental models that could drastically facilitate a wide range of daily decisions made by caregivers in oncology but, unfortunately, it often lacks clinical validation for now. NLP can support physicians in identifying the most (predicted) effective treatment options or clinical trials for a patient. It can match or even surpass caregivers in estimating prognosis, predicting duration of participation in clinical trials, and assessing the risk of acute medical events. This makes NLP a valuable tool for guiding patient monitoring, among other applications. AI-driven image analysis has already streamlined numerous diagnostic tasks and is gradually garnering prospective validation. Nonetheless, translating AI research into practical applications for language models mandates a meticulous approach emphasizing trustworthiness, responsibility, and ethics. The design of prospective clinical evaluations for medical decision support systems based on NLP must be carefully tailored to the specific application, as traditional randomization schemes may not be suitable for all scenarios. Multidisciplinary teams with dedicated time are thus required for designing the evaluations that will shape our future practice. Finally, caregivers should be well informed about the practical pitfalls of artificial intelligence to genuinely improve patient care and avoid potential harm.

Contributors

Loic Verlingue: Writing -original draft and reviews, Figures, Conceptualization, Supervision, Project Administration. Clara Boyer: Writing–original draft. Louise Olgiati: Writing—reviews. Clément Brutti Mairesse: Writing -original draft and reviews. Daphné Morel: Writing -original draft and reviews, Supervision of the reviews. Jean-Yves Blay: Writing–original draft. All authors read and approved the final version of the manuscript.

Declaration of interests

LV reports personal fees from Adaptherapy, is CEO of RESOLVED, has received non-personal fees from Pierre-Fabre and Servier, and a grant from Bristol-Myers Squibb, all outside the submitted work. As part of the Phase 1 unit of Centre Léon Bérard, as medical doctor, LV report being: Principal/sub-Investigator of Clinical Trials for Abbvie, Adaptimmune, Aduro Biotech, Agios Pharmaceuticals, Amgen, Argen-X Bvba, Arno Therapeutics, Astex Pharmaceuticals, Astra Zeneca Ab, Aveo, Basilea Pharmaceutica International Ltd, Bayer Healthcare Ag, Bbb Technologies Bv, Beigene, Blueprint Medicines, Boehringer Ingelheim, Boston Pharmaceuticals, Bristol Myers Squibb, Ca, Celgene Corporation, Chugai Pharmaceutical Co, Clovis Oncology, Cullinan-Apollo, Daiichi Sankyo, Debiopharm, Eisai, Eisai Limited, Eli Lilly, Exelixis, Faron Pharmaceuticals Ltd, Forma Tharapeutics, Gamamabs, Genentech, Glaxosmithkline, H3 Biomedicine, Hoffmann La Roche Ag, Imcheck Therapeutics, Innate Pharma, Institut De Recherche Pierre Fabre, Iris Servier, Janssen Cilag, Janssen Research Foundation, Kura Oncology, Kyowa Kirin Pharm. Dev, Lilly France, Loxo Oncology, Lytix Biopharma As, Medimmune, Menarini Ricerche, Merck Sharp & Dohme Chibret, Merrimack Pharmaceuticals, Merus, Millennium Pharmaceuticals, Molecular Partners Ag, Nanobiotix, Nektar Therapeutics, Novartis Pharma, Octimet Oncology Nv, Oncoethix, Oncopeptides, Orion Pharma, Ose Pharma, Pfizer, Pharma Mar, Pierre Fabre, Medicament, Roche, Sanofi Aventis, Seattle Genetics, Sotio A.S, Syros Pharmaceuticals, Taiho Pharma, Tesaro, Xencor. Research Grants from Astrazeneca, BMS, Boehringer Ingelheim, Janssen Cilag, Merck, Novartis, Onxeo, Pfizer, Roche, Sanofi. Non-financial support (drug supplied) from Astrazeneca, Bayer, BMS, Boringher Ingelheim, Medimmune, Merck, NH TherAGuiX, Onxeo, Pfizer. ED reports grants and personal fees from Roche Genentech, grants from Boehringer, grants from Astrazeneca, grants and personal fees from Merck Serono, grants from BMS, and grants from MSD Roche. JYB declares Grant from Roche for research program Profiler 2, Grant from the French NCI for LYRICAN+, Grant from MSD Avenir for an AI project, grant from ANR for the PORTRAIT program. The other authors declare no potential conflicts of interest.

Acknowledgements

Loic Sebileau and Hugo Crochet for their support in the paragraph explaining CE marking and their views on multidisciplinary collaborations.

Funding source: Grant from the European Regional Development Fund FEDER, TRIAL MATCH 2, N° SYNERGIE: 20023.

References

1 Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. *Br J Cancer*. 2022;126(1):4–9.

2 Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*; 2019 [cité 5 oct 2020]; Disponible sur: http://arxiv.org/abs/1810.04805.

3 He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv*; 2023 [cité 18 mars 2024]. Disponible sur: http://arxiv.org/abs/2310.05694.

4 Keraghel I, Morbieu S, Nadif M. A survey on recent advances in named entity recognition. *arXiv*; 2024 [cité 22 mars 2024]. Disponible sur: http://arxiv.org/abs/2401.10825.

5 Chen S, Li Y, Lu S, et al. Evaluation of ChatGPT family of models for biomedical reasoning and classification. *arXiv*; 2023 [cité 18 mars 2024]. Disponible sur: http://arxiv.org/abs/2304.02496.

6 Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. *arXiv*; 2024 [cité 15 avr 2024]. Disponible sur: http://arxiv.org/abs/2312.10997.

7 Samsi S, Zhao D, McDonald J, et al. From words to watts: benchmarking the energy costs of large language model inference. In: *2023 IEEE high performance extreme computing conference (HPEC)*; 2023:1–9 [cité 26 mars 2024] Disponible sur: https://ieeexplore.ieee.org/abstract/document/10363447.

8 Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):e333–e335.

9 Yu H, Lee M, Kaufman D, et al. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *J Biomed Inform*. 2007;40(3):236–251.

10 Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv*; 2023 [cité 22 janv 2024]. Disponible sur: http://arxiv.org/abs/2311.16452.

11 Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: scaling medical pretraining for large language models. *arXiv*; 2023 [cité 25 mars 2024]. Disponible sur: http://arxiv.org/abs/2311.16079.

12 Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A systematic survey of prompt engineering in large language models: techniques and applications. *arXiv*; 2024 [cité 30 juill 2024]. Disponible sur: http://arxiv.org/abs/2402.07927.

13 Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv*; 2023 [cité 30 mai 2024]. Disponible sur: http://arxiv.org/abs/2305.09617.

14 Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: *Proceedings of the conference on health, inference, and learning*. PMLR; 2022:248–260 [cité 11 juin 2024]. Disponible sur: https://proceedings.mlr.press/v174/pal22a.html.

15 Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024:1–10.

16 Corbaux P, Bayle A, Besle S, et al. Patients' selection and trial matching in early-phase oncology clinical trials. *Crit Rev Oncol Hematol*. 2024;196:104307.

17 Krishnan M, Temel JS, Wright AA, Bernacki R, Selvaggi K, Balboni T. Predicting life expectancy in patients with advanced incurable cancer: a review. *J Support Oncol*. 2013;11(2):68–74.

18 Smith-Uffen MES, Johnson SB, Martin AJ, et al. Estimating survival in advanced cancer: a comparison of estimates made by oncologists and patients. *Support Care Cancer*. 2020;28(7):3399–3407.

19 Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357–362.

20 Piat C, Blampey Q, Joutard A, et al. *A validated and explainable deep learning model instantly predicts survival from consultation reports*. Rochester, NY; 2023 [cité 26 mai 2023]. Disponible sur: https://papers.ssrn.com/abstract=4410792.

21 Ogier du Terrail J, Leopold A, Joly C, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat Med*. 2023;29(1):135–146.

22 Landry G, Kurz C, Traverso A. The role of artificial intelligence in radiotherapy clinical practice. *BJR Open*. 2023;5(1):20230030.

23 Al-Tashi Q, Saad MB, Muneer A, et al. Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. *Int J Mol Sci*. 2023;24(9):7781.

24 Geaney A, O'Reilly P, Maxwell P, James JA, McArt D, Salto-Tellez M. Translation of tissue-based artificial intelligence into clinical practice: from discovery to adoption. *Oncogene*. 2023;42(48):3545–3555.

25 Zhang Z, Wei X. Artificial intelligence-assisted selection and efficacy prediction of antineoplastic strategies for precision cancer therapy. *Semin Cancer Biol*. 2023;90:57–72.

26 On J, Park HA, Yoo S. Development of a prediction models for chemotherapy-induced adverse drug reactions: a retrospective observational study using electronic health records. *Eur J Oncol Nurs*. 2022;56:102066.

27 Di Meglio A, Havas J, Soldato D, et al. Development and validation of a predictive model of severe fatigue after breast cancer diagnosis: toward a personalized framework in survivorship care. *J Clin Oncol*. 2022;40(10):1111–1123.

28 Jie Z, Zhiying Z, Li L. A meta-analysis of Watson for Oncology in clinical application. *Sci Rep*. 2021;11(1):5792.

29 Li J, Yuan Y, Bian L, et al. A comparison between clinical decision support system and clinicians in breast cancer. *Heliyon*. 2023;9(5):e16059.

30 Charton E, Baldini C, Fayet Y, et al. Inequality factors in access to early-phase clinical trials in oncology in France: results of the EGALICAN-2 study. *ESMO Open*. 2023;8(4) [cité 26 août 2023] Disponible sur: https://www.esmoopen.com/article/S2059-7029(23)00845-1/fulltext.

31 Chow R, Midroni J, Kaur J, et al. Use of artificial intelligence for cancer clinical trial enrollment: a systematic review and meta-analysis. *JNCI J Natl Cancer Inst*. 2023;115(4):365–374.

32 Truong TH, Otmakhova Y, Mahendra R, et al. ITTC @ TREC 2021 clinical trials track. *arXiv*; 2022 [cité 20 févr 2024]. Disponible sur: http://arxiv.org/abs/2202.07858.

33 Jullien M, Valentino M, Freitas A. SemEval-2024 task 2: Safe biomedical Natural Language inference for clinical trials. *arXiv*; 2024 [cité 28 mai 2024]. Disponible sur: http://arxiv.org/abs/2404.04963.

34 Jin Q, Wang Z, Floudas CS, Sun J, Lu Z. Matching patients to clinical trials with Large Language Models. *arXiv*; 2023 [cité 31 août 2023]. Disponible sur: http://arxiv.org/abs/2307.15051.

35 Jullien M, Valentino M, Frost H, O'Regan P, Landers D, Freitas A. NLI4CT: multi-evidence Natural Language inference for clinical trial reports. *arXiv*; 2023 [cité 13 sept 2023]. Disponible sur: http://arxiv.org/abs/2305.03598.

36 Delorme J, Charvet V, Wartelle M, et al. Natural language processing for patient selection in phase I/II oncology clinical trials. *medRxiv*. 2021, 2021.02.07.21249271.

37 Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

38 Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. *Npj Digit Med*. 2024;7(1):1–3.

39 Blagec K, Kraiger J, Frühwirt W, Samwald M. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *J Biomed Inform*. 2023;137:104274.

40 Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1.

41 Celi LA, Cellini J, Charpignon ML, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digit Health*. 2022;1(3):e0000022.

42 Kiela D, Bartolo M, Nie Y, et al. Dynabench: rethinking benchmarking in NLP. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al., eds. *Éditeurs. Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics:*

human language technologies. Online: Association for Computational Linguistics; 2021:4110–4124 [cité 11 juin 2024]. Disponible sur: https://aclanthology.org/2021.naacl-main.324.

43 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023:1–9.

44 Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024:1–9.

45 Potnis KC, Ross JS, Aneja S, Gross CP, Richman IB. Artificial intelligence in breast cancer screening: evaluation of FDA device regulation and future recommendations. *JAMA Intern Med*. 2022;182(12):1306–1312.

46 Ferber D, Nahhas OSME, Wölflein G, et al. Autonomous artificial intelligence agents for clinical decision making in oncology. *arXiv*; 2024 [cité 31 juill 2024]. Disponible sur: http://arxiv.org/abs/2404.04667.

47 Weng WH, Sellergen A, Kiraly AP, et al. An intentional approach to managing bias in general purpose embedding models. *Lancet Digit Health*. 2024;6(2):e126–e130.

48 Baumgartner R, Arora P, Bath C, et al. Fair and equitable AI in biomedical research and healthcare: social science perspectives. *Artif Intell Med*. 2023;144:102658.

49 Ng AY, Oberije CJG, Ambrózay É, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nat Med*. 2023;29(12):3044–3049.

50 Hong JC, Eclov NCW, Dalal NH, et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning–directed clinical evaluations during radiation and chemoradiation. *J Clin Oncol*. 2020;38:3652.

51 Zeng C, Zhang J. A narrative review of five multigenetic assays in breast cancer. *Transl Cancer Res*. 2022;11(4):897–907.

52 Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of Diabetic retinopathy. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery; 2020 [cité 15 avr 2024]. (CHI '20). Disponible sur: https://dl.acm.org/doi/10.1145/3313831.3376718.

53 Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nat Rev Clin Oncol*. 2014;11(2):81–90.

54 Wu CJ, Raghavendra R, Gupta U, et al. Sustainable AI: environmental implications, challenges and opportunities. *arXiv*; 2021 [cité 2 août 2024]; Disponible sur: https://www.semanticscholar.org/paper/Sustainable-AI%3A-Environmental-Implications%2C-and-Wu-Raghavendra/2c6df83795cd5baf3b8c6e2639b85e2df0cee1d0.

55 Luccioni AS, Viguier S, Ligozat AL. Estimating the carbon footprint of BLOOM, a 176B parameter Language Model. *J Mach Learn Res*. 2024, 253:11990–253:12004.

56 Luccioni S, Jernite Y, Strubell E. Power hungry processing: watts driving the cost of AI deployment?. In: *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery; 2024:85–99 [cité 2 août 2024].(FAccT '24). Disponible sur: https://dl.acm.org/doi/10.1145/3630106.3658542.

57 Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: Korhonen A, Traum D, Màrquez L, éditeurs, eds. *Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence, Italy: Association for Computational Linguistics; 2019 [cité 2 août 2024]. Disponible sur: https://aclanthology.org/P19-1355.

58 Ligozat AL, Lefevre J, Bugeau A, Combaz J. Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability*. 2022;14(9):5172.

59 Fort K, Adda G, Cohen KB, Words L. Amazon mechanical turk: gold mine or coal mine? *Comput Linguist*. 2011;37(2):413–420.

60 Movva R, Balachandar S, Peng K, Agostini G, Garg N, Pierson E. Topics, authors, and institutions in Large Language Model research: trends from 17K arXiv papers. In: Duh K, Gomez H, Bethard S, éditeurs, eds. *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*. Mexico City, Mexico: Association for Computational Linguistics; 2024:1223–1243 [cité 2 août 2024]. Disponible sur: https://aclanthology.org/2024.naacl-long.67.

61 Challenging systematic prejudices: an investigation into bias against women and girls in large language models - UNESCO Bibliothèque Numérique [cité 2 août 2024]. Disponible sur: https://unesdoc.unesco.org/ark:/48223/pf0000388971.

62 Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Linguist Compass*. 2021;15(8):e12432.

63 Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *Npj Digit Med*. 2023;6(1):1–4.

64 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453.

65 Banerjee I, Bhattacharjee K, Burns JL, et al. Shortcuts » causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *J Am Coll Radiol JACR*. 2023;20(9):842–851.

66 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683.

67 Vicente L, Matute H. Humans inherit artificial intelligence biases. *Sci Rep*. 2023;13(1):15737.

68 Pham HV, Qian S, Wang J, et al. Problems and opportunities in training deep learning software systems: an analysis of variance. In: *Proceedings of the 35th IEEE/ACM international conference on automated software engineering*. New York, NY, USA: Association for Computing Machinery; 2021:771–783. [cité 2 août 2024]. (ASE '20). https://doi.org/10.1145/3324884.3416545.

69 Verma H, Mlynar J, Schaer R, et al. Rethinking the role of AI with physicians in oncology: revealing perspectives from clinical and research workflows. In: *Proceedings of the 2023 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery; 2023:1–19 [cité 2 août 2024]. (CHI '23). Disponible sur: https://dl.acm.org/doi/10.1145/3544548.3581506.

70 DeCamp M, Lindvall C. Mitigating bias in AI at the point of care. *Science*. 2023;381(6654):150–152.

71 Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th international conference on machine learning - volume 70*. Sydney, NSW, Australia: JMLR.org; 2017:1321–1330 (ICML'17).

72 Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. *ACM Comput Surv*. 2014;46(4):44:1–44:37.

73 Heudel PE, Crochet H, Blay JY. Impact of artificial intelligence in transforming the doctor–cancer patient relationship. *ESMO Real World Data Digit Oncol*; 2024:3 [cité 2 août 2024];Disponible sur: https://www.esmorwd.org/article/S2949-8201(24)00004-3/fulltext.

74 Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–1363.

75 Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374.