Article

Not peer-reviewed version

Automated Clinical Annotation of Lung Cancer Reports Using Transformer-Based NLP Models: A Benchmarking and Fine-Tuning Study on a Novel Tunisian Clinical Corpus

Ranim Yahyaoui , Ismail Dergaa , Jean Noel Nikiema , Halil İbrahim Ceylan * , Nicola Luigi Bragazzi * , Saoussen Hantous-Zannad , Hanene Boussi Rahmouni

Posted Date: 22 April 2026

doi: 10.20944/preprints202604.1557.v1

Keywords: BioClinicalBERT; clinical NLP; DrBERT; lung cancer; named entity recognition; NER; RoBERTa; TNM staging; Tunisian corpus; transformer models

 doi:10.20944/preprints202604.1557.v1

Article

Automated Clinical Annotation of Lung Cancer Reports Using Transformer-Based NLP Models: A Benchmarking and Fine-Tuning Study on a Novel Tunisian Clinical Corpus

Ranim Yahyaoui [1], Ismail Dergaa [2,3,4], Jean Noel Nikiema [5], Halil İbrahim Ceylan [6,*], Nicola Luigi Bragazzi [7,*], Saoussen Hantous-Zannad [8,9] and Hanene Boussi Rahmouni [1,10]

[1] Laboratory of Biophysics and Medical Technologies, Higher Institute of Medical Technologies of Tunis (ISTMT), University of Tunis El Manar, Tunisia
[2] High Institute of Sport and Physical Education of Ksar Said, University of Manouba, Manouba 2010, Tunisia
[3] Physical Activity Research Unit, Sport and Health (UR18JS01), National Observatory of Sports, Tunis 1003, Tunisia
[4] High Institute of Sport and Physical Education of Kef, University of Jendouba, Jendouba 7100, Tunisia
[5] Department of Management, Evaluation and Health Policy, School of Public Health, University of Montreal, Montreal, Canada
[6] Physical Education of Sports Teaching Department, Faculty of Sports Sciences, Atatürk University, Erzurum 25240, Türkiye
[7] Department of Clinical Pharmacy, Saarland University, 66123 Saarbrücken, Germany
[8] University of Tunis El Manar, Faculty of Medicine of Tunis, Tunis 1007, Tunisia
[9] Abderrahmen Mami Hospital, Medical Imaging Department, Ariana 2035, Tunisia
[10] The Computer Science Research Center, The University of the West of England, Bristol BS16 1QY, UK
* Correspondence: halil.ibrahimceylan60@gmail.com (H.İ.C.); nicola.bragazzi@uni-saarland.de (N.L.B.)

Abstract

Background: Lung cancer causes more deaths than any other malignancy worldwide, accounting for 2.2 million new cases and 1.8 million deaths in 2020. Extracting structured clinical knowledge from unstructured French-language oncology records remains methodologically unresolved in Tunisian and Francophone healthcare systems, where validated natural language processing tools do not yet exist. This study examined the effectiveness of transformer-based named entity recognition for automated clinical annotation of Tunisian lung cancer reports. **Aim:** The study aimed to (i) benchmark four transformer-based models on a publicly available thoracic radiology dataset, (ii) evaluate five models, including a French biomedical specialist, on a newly constructed Tunisian clinical corpus, and (iii) demonstrate prototype deployment feasibility for structured clinical decision support. **Methods:** A benchmarking study evaluated BERT, RoBERTa, BioClinicalBERT, and CamemBERT on the RadGraph dataset (600 annotated thoracic radiology reports). Five models were subsequently fine-tuned on 200 manually annotated initial diagnostic reports from Mami Pneumo-Phthisiology Hospital, Tunis. All models were trained for a maximum of 10 epochs, with a learning rate of $5x10^{-5}$, a batch size of 16, and an 80/10/10 train-validation-test split, and evaluated using precision, recall, and F1-score. **Results:** On RadGraph, RoBERTa achieved the highest F1-score of 0.873 (precision: 0.869, recall: 0.877), followed by BioClinicalBERT (F1: 0.868) and BERT (F1: 0.857). CamemBERT achieved an F1 score of 0.682 on this English dataset. On the Tunisian corpus, DrBERT outperformed all models with an F1-score of 0.811, compared to RoBERTa at 0.79. A prototype interface generated structured clinical summaries encompassing prior conditions, imaging modalities, and TNM staging. **Conclusion:** Language- and domain-adapted transformer models effectively extract structured clinical entities from French-language Tunisian lung cancer reports. DrBERT's precision advantage confirms that biomedical pretraining in the target language is the primary driver of performance in specialized French oncology text. This work establishes

 doi:10.20944/preprints202604.1557.v1

foundational infrastructure for NLP-driven oncology data management in Tunisia and comparable Francophone settings.

Keywords: BioClinicalBERT; clinical NLP; DrBERT; lung cancer; named entity recognition; NER; RoBERTa; TNM staging; Tunisian corpus; transformer models

1. Introduction

Lung cancer occupies a singular and devastating position in global oncology: it ranks second in incidence yet leads all malignancies in mortality across both sexes, accounting for 2.2 million new cases and 1.8 million deaths in 2020 alone [1]. The five-year survival rate in high-income countries remains at approximately 15%, a figure largely unchanged despite decades of therapeutic advancement [2]. Two converging factors explain this persistent lethality. The first is late-stage diagnosis: the overwhelming majority of patients present at stage III or IV, when curative resection is no longer feasible and systemic therapies offer limited, durable benefit [3]. The second is the biological complexity of the disease itself. Lung cancer exhibits profound heterogeneity across imaging phenotypes, histopathological subtypes, genomic alterations, and protein expression profiles, rendering treatment selection — chemotherapy, targeted therapy, immunotherapy, or multimodal combinations with surgery and radiotherapy — inherently dependent on granular individual data [4,5]. Precision medicine is the conceptual and practical response to this challenge: matching therapeutic strategies to the individual patient's molecular and clinical profile by integrating genomic, radiomic, and longitudinal clinical data [4,6]. Realizing this vision requires continuous, structured, and reliable access to patient information at scales that manual clinical workflows cannot sustain.

Electronic health records constitute the primary repository of clinical intelligence in contemporary healthcare systems. They contain longitudinal, multi-domain data encompassing patient demographics, laboratory values, imaging reports, pathology findings, disease progression notes, and treatment records [7,8]. A substantial and systematically underutilized proportion of this information, however, is embedded in free-text clinical narratives rather than structured database fields. Radiology reports, oncology consultation notes, and surgical pathology summaries document critical clinical reasoning in natural language, rendering them resistant to automated computational query and analysis [8,9]. The resulting gap between information stored in electronic health records and information extractable from them is one of the most consequential inefficiencies in contemporary clinical and translational research [10]. Manual information extraction is costly, time-consuming, prone to inter-rater inconsistency, and fundamentally unscalable at the data volumes required for population-level oncology research or real-world evidence generation [9,11]. Training staff on electronic health record platforms addresses workflow barriers but does not resolve the unstructured text problem; Musa et al. demonstrated in a Qatari wellness context that even targeted one-to-one training programs substantially reduced booking times and improved practical competency, yet the challenge of extracting structured knowledge from narrative clinical text remained entirely unaddressed [12]. Automated natural language processing has emerged as the methodological response, enabling systematic identification and extraction of clinically meaningful entities from free text, converting narrative documentation into structured, computable data at scale [7,11]. The clinical implications are substantial: structured NLP outputs can support patient stratification, epidemiological surveillance, real-world evidence generation, clinical trial recruitment, and construction of clinical decision support systems [13,14].

The methodological trajectory of clinical NLP has advanced through three identifiable phases. Rule-based systems, the earliest generation, rely on handcrafted lexicons, regular expressions, and curated ontologies such as the Unified Medical Language System to identify clinical concepts. Nguyen et al. applied this approach to TNM classification in lung cancer pathology reports, achieving accuracies of 72%, 78%, and 94% for T, N, and M stages, respectively [15]. Beyer et al. deployed a

 doi:10.20944/preprints202604.1557.v1

rule-based NLP algorithm for lung nodule characterization from structured CT reports, achieving 75.0% sensitivity and 98.8% specificity. The clinical Text Analysis and Knowledge Extraction System developed by Savova et al. integrated rule-based and machine-learning components into a modular pipeline that covers sentence segmentation, tokenization, part-of-speech tagging, named entity recognition, and negation detection [16]. Machine learning approaches improved upon rule-based performance but remained constrained by extensive feature engineering requirements [16,17]. Deep learning addressed this constraint by automatically extracting features from raw text. Gupta et al. investigated Long Short-Term Memory networks for clinical entity extraction from CT text reports [18] ; and Chen et al. demonstrated that convolutional neural networks matched or exceeded traditional NLP models for pulmonary embolism classification in thoracic CT reports [19]. The most consequential shift came with transformer-based architectures. BERT [20] introduced bidirectional contextual representation that substantially elevated performance across NLP tasks. Domain adaptation through targeted pretraining on biomedical and clinical corpora produced the current specialist generation: BioClinicalBERT, pretrained on MIMIC-III clinical notes [21], and DrBERT, a RoBERTa architecture pretrained on the French biomedical corpus NACHOS [22]. Machine learning methods more broadly have demonstrated strong classification performance across health contexts, with ensemble and deep learning models achieving 84-92.4% accuracy in predicting sedentary behavior from sensor data [23] and an AUROC of 0.98 in predicting early ARDS from local Tunisian clinical data [24]. These findings collectively establish that high-performing clinical AI systems can be built on regionally collected datasets with appropriate methodological choices. Abdaoui et al. demonstrated this directly in the Tunisian pathology context, where a hybrid BioClinicalBERT model augmented with dense retrieval achieved an F1-score of 0.97 for clinical entity extraction from locally collected pathology reports, confirming both the feasibility and the performance attainable with domain-adapted architectures on Tunisian clinical data [25].

A structural gap persists despite this body of evidence. The overwhelming majority of clinical NLP research has been conducted in English, with more recent extensions to Chinese clinical narratives [26]. French-language clinical NLP, specifically lung cancer named-entity recognition in French, remains methodologically nascent. The Tunisian healthcare system compounds this gap through documentation practices that blend standard French medical terminology with locally adapted clinical nomenclature, and through the near-total absence of publicly available annotated French clinical corpora in oncology. Existing tools developed and validated on English-language medical texts cannot be applied reliably in Francophone contexts without systematic adaptation [22]. Addressing these converging gaps in language, domain, and geographic representation is not merely a technical problem: it is a matter of health equity, as clinicians in French-speaking and North African settings are systematically excluded from the clinical AI pipeline benefiting English-language counterparts. Building on these identified deficits and leveraging the demonstrated effectiveness of transformer-based deep learning in clinical text processing, the present pilot study pursued three specific aims: (i) to benchmark BERT, RoBERTa, BioClinicalBERT, and CamemBERT on the RadGraph thoracic radiology dataset, establishing comparative performance baselines for clinical NER; (ii) to evaluate five transformer-based models, including DrBERT as a French biomedical specialist, on a newly constructed and manually annotated Tunisian lung cancer clinical corpus; and (iii) to demonstrate prototype deployment feasibility through a DrBERT-powered structured clinical interface enabling real-world decision support.

2. Materials and Methods

2.1. Ethical Approval

This study was conducted in accordance with the Declaration of Helsinki and current international guidelines for research practice in health and clinical informatics [27]. Clinical reports used in corpus construction were retrospectively collected and fully de-identified prior to any computational processing, in accordance with applicable national data protection legislation. The

 doi:10.20944/preprints202604.1557.v1

research protocol was reviewed and approved by the Ethics Committee of Hôpital Mami Ariana, Tunisia (approval date: 15 June 2025). The study was conducted under the scientific supervision of Dr. Saoussen Hantous-Zannad, Head of the Department of Medical Imaging at Hôpital Mami Ariana, and Prof. Hanene Boussi. The project was carried out in collaboration with the Higher Institute of Medical Technologies of Tunis (ISTMT) and the School of Public Health at the Université de Montréal, with academic supervision by Prof. Jean-Noël Nikiema. Written informed consent was obtained from all participants or their legal representatives for the use of identifiable data collected during the initial clinical process. All data were fully anonymized before transfer to the research team.

2.2. Benchmark Dataset: RadGraph

The RadGraph dataset [28] comprises 600 annotated thoracic radiology reports reviewed and validated by board-certified radiologists. Annotations cover two primary entity categories: anatomy (ANAT) and observation (OBS). Observation entities are classified by certainty level into three subcategories — Definitely Present, Uncertain, and Definitely Absent — yielding four distinct entity labels for NER purposes: ANAT-DP, OBS-DP, OBS-U, and OBS-DA. The dataset was anonymized in compliance with the Health Insurance Portability and Accountability Act and serves as a publicly accessible benchmark for NLP method development in thoracic radiology. Figure 1 presents an excerpt from an annotated thoracic imaging report in the RadGraph dataset, illustrating the entity-labeling conventions applied throughout the benchmarking phase.

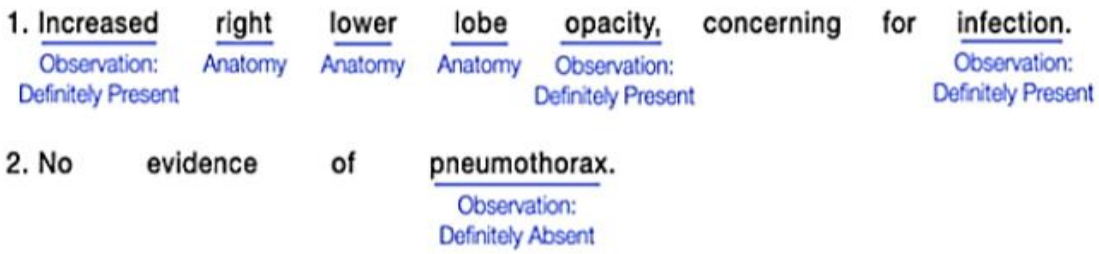

Figure 1. Excerpt from an Annotated Thoracic Imaging Report.

2.3. Data Preprocessing

Identical preprocessing was applied to both the RadGraph dataset and the Tunisian clinical corpus to ensure methodological consistency across the two study phases. Radiology and clinical reports frequently contain inconsistent spacing, special characters, and extraneous punctuation that disrupts tokenization reliability. Non-informative symbols and redundant whitespace were removed, and spacing was normalized across all tokens.

Each report was segmented into individual sentences and tokenized into words and meaningful symbols using the SpaCy tokenization library. Entities were encoded using the IOB2 (Inside-Outside-Beginning) tagging scheme, the standard annotation format for NER tasks in clinical informatics: B-ENTITY marks the beginning token of a named entity; I-ENTITY marks continuation tokens; O marks all tokens outside any named entity.

Preprocessed data were formatted according to the CoNLL standard, with each token on a separate line, its IOB2 label, and sentences separated by blank lines. Figure 2 presents the IOB labels in the RadGraph dataset after preprocessing. Figure 3 shows a representative excerpt from the Tunisian lung cancer corpus after full preprocessing, illustrating the IOB2 encoding of the three annotated entity categories.

 doi:10.20944/preprints202604.1557.v1

	sentence_id	words	labels
0	0	FINAL	O
1	0	REPORT	O
2	0	EXAMINATION	O
3	0	:	O
4	0	CHEST	O
...	...	...	...
46549	424	No	O
46550	424	acute	B-OBS-DA
46551	424	cardiopulmonary	B-ANAT-DP
46552	424	process	B-OBS-DA
46553	424	.	O

46554 rows × 3 columns

Figure 2. The IOB labels in the RadGraph dataset.

	sentence_id	words	labels
0	0	tabagique	B-R_CLINIQUES
1	0	actif	B-R_CLINIQUES
2	0	à	O
3	0	50	B-R_CLINIQUES
4	0	PA.	I-R_CLINIQUES

Figure 3. An excerpt from the Tunisian lung cancer corpus.

2.4. Construction and Annotation of the Tunisian Lung Cancer Corpus

Two hundred initial diagnostic reports were randomly selected from patients with confirmed lung cancer diagnoses at Mami Pneumo-Phthisiology Hospital, Ariana, Tunisia. Reports were manually de-identified prior to annotation, with all patient-identifying information removed or replaced with generic placeholders by a trained clinical staff member not involved in subsequent annotation.

Annotation Scheme

Before initiating any therapeutic approach, lung cancer patients undergo a staging assessment designed to classify disease according to TNM criteria, which determines treatment strategy. A standard Tunisian initial diagnostic report encompasses four sections: Clinical Information, Techniques, Findings, and Conclusion. Three entity types were defined in consultation with oncology domain experts, corresponding to the staging-critical content of these report sections.

The entity types and their clinical scope are summarized in Table 1, which presents NER labels and their corresponding clinical information categories. Briefly:

- R. CLINIQUES: captures the patient's relevant clinical background, including smoking history, occupational exposures, and initial clinical investigations (fibroscopy, chest X-ray, CT scan) performed prior to or during the staging workup.

 doi:10.20944/preprints202604.1557.v1

- TECHNIQUES: covers technical details of imaging procedures conducted during the staging workup, including acquisition type, technical parameters, radiation dose, and scan coverage.
- STADE: identifies the cancer stage as determined by the staging investigations, expressed using TNM classification.

Table 1. NER Labels and Clinical Information Categories.

Groups	Label NER	What does it include
Clinical Informations	R-CLINIQUES	A Smoker? Number of packs/years Professional Exposure Medical History
Techniques	TECHNIQUES	Acquisition Type Technical Parameters Anatomical Region Explored Contrast Phase Dose Irradiation Dose Dose Explored Area
Conclusion	STADE	Stage

This three-entity scheme was deliberately scoped to the staging-relevant subset of clinically actionable information. Expansion to histological subtype, biomarker status (EGFR, ALK, PD-L1), performance status, and comorbidity profile is identified as a direct objective for subsequent corpus development phases.

Annotation was conducted using Med-Tator, a serverless text annotation tool for corpus development. Two annotators with clinical oncology expertise independently labeled each report; disagreements were resolved by consensus adjudication with a third senior reviewer.

The annotation process was primarily carried out by the first and last authors, both of whom have a biomedical background in biomedical engineering and medical informatics. Given the relatively straightforward nature of the annotated entities (e.g., clinical stage and basic clinical and technical information), the annotation guidelines were designed to ensure consistency across reports. All annotations were subsequently reviewed and validated by clinical experts from the Department of Medical Imaging at Hôpital Mami Ariana, under the supervision of the Department's Head. In cases of uncertainty, annotations were discussed with the clinicians until consensus was reached. This expert validation ensured the clinical accuracy and reliability of the final annotated corpus used in this study.

Sections of reports outside the current annotation scope were excluded from the annotated corpus version prior to model training. Figure 4 presents sample annotated sentences from the Tunisian lung cancer corpus, illustrating the three entity categories in representative clinical text.

 doi:10.20944/preprints202604.1557.v1

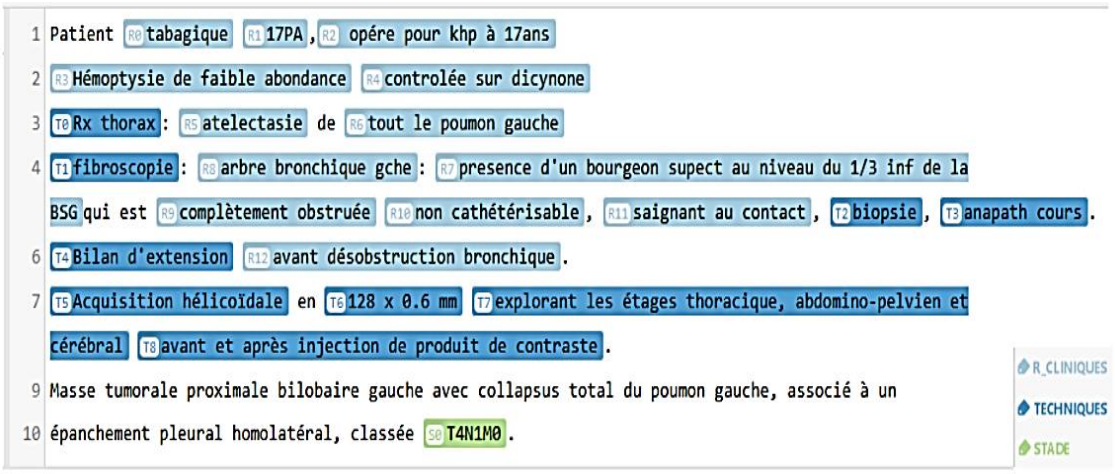

Figure 4. Samples of Sentences and their annotations in the lung cancer corpus.

2.5. Model Selection and Architecture

2.5.1. Benchmarking Phase (RadGraph Dataset)

Four pretrained transformer-based models were evaluated. BERT **[20]** is a general-purpose English language model pretrained on Wikipedia and BookCorpus through masked language modeling and next-sentence prediction objectives. RoBERTa is an optimized BERT variant trained with larger batch sizes, extended training duration, and removal of the next-sentence prediction objective, producing more robust contextual representations [29]. BioClinicalBERT [21] is initialized from BioBERT and further pretrained on MIMIC-III clinical notes, making it specifically suited to clinical free text. CamemBERT [30] is a French-language model pretrained on 138 GB of French text from the OSCAR corpus; its inclusion on the English RadGraph benchmark served as a methodological control, establishing a language-mismatch performance reference and justifying its subsequent evaluation on the French Tunisian corpus phase.

2.5.2. Tunisian Corpus Phase

Five models were evaluated on the Tunisian corpus: the four described above plus DrBERT [22], a state-of-the-art French biomedical language model based on the RoBERTa architecture and pretrained on the French biomedical corpus NACHOS. DrBERT was selected as the primary candidate for the Tunisian corpus phase because it uniquely combines RoBERTa's architectural robustness with French biomedical domain pretraining, positioning it optimally for French clinical oncology text.

2.6. Fine-Tuning Protocol

Each pretrained model was fine-tuned using a standard NER approach based on the IOB2 tagging scheme. Both datasets used an 80/10/10 train-validation-test split. All models were initialized with their respective pretrained weights and trained for a maximum of 10 epochs with a learning rate of 5×10^{-5} and batch size of 16 for both training and evaluation. Early stopping was applied based on the validation F1-score. The training directory was overwritten at each run to ensure reproducibility and prevent carryover between experiments.

2.7. Evaluation Metrics

Final performance was assessed on the held-out test set using precision, recall, and F1-score computed at the entity boundary level. Precision measures the proportion of predicted entities that are correct (TP / (TP + FP)). Recall measures the proportion of true entities retrieved (TP/(TP+FN)).

 doi:10.20944/preprints202604.1557.v1

The F1-score is the harmonic mean of precision and recall, providing a single balanced performance index. Evaluation loss was recorded as a supplementary calibration measure.

2.8. Prototype Development

The best-performing model on the Tunisian corpus was embedded in a prototype clinical interface designed to assess the feasibility of real-world deployment. Visualization and exploratory analyses were conducted using Power BI. The prototype generates structured patient summaries providing immediate access to extracted clinical entities — prior conditions, technical modalities, and TNM staging — alongside navigable access to the full original report within a unified dashboard environment.

2.9. AI Usage Statement

In preparing this manuscript, the authors used Claude (Anthropic) to improve the clarity and grammatical correctness of selected passages. The tool was used to revise text for an academic tone, check for grammatical errors, and enhance the quality of the English language. The authors did not use AI for data analysis, interpretation, or generation of scientific content. After using this tool, the authors thoroughly reviewed and edited all content and take full responsibility for the accuracy, integrity, and scientific validity of the work [31,32].

3. Results

3.1. Benchmarking Results on the RadGraph Dataset

Table 2 summarizes model performance on the RadGraph test set across all four transformer-based models evaluated in the benchmarking phase.

RoBERTa achieved the highest overall performance, recording a precision of 0.869, recall of 0.877, F1-score of 0.873, and evaluation loss of 0.275. BioClinicalBERT performed closely behind with a precision of 0.858, a recall of 0.878, an F1-score of 0.868, and an evaluation loss of 0.254, the lowest across all models. BERT achieved a precision of 0.855, a recall of 0.859, an F1-score of 0.857, and an evaluation loss of 0.441. CamemBERT recorded a precision of 0.670, a recall of 0.695, an F1-score of 0.682, and an evaluation loss of 0.529. The performance differential between CamemBERT and the three English-pretrained models (delta F1: 0.175-0.191) reflects the language mismatch between its French pretraining corpus and the English RadGraph evaluation context, confirming that linguistic alignment is the primary driver of performance differences in this benchmarking phase. Based on these results, RoBERTa was selected for direct transfer to the Tunisian corpus phase.

Table 2. Benchmarking Results on the RadGraph Dataset.

Model	Precision	Recall	F1-Score	Eval Loss
RoBERTa	0.869	0.877	0.873	0.275
BioClinicalBERT	0.858	0.878	0.868	0.254
BERT	0.855	0.859	0.857	0.441
CamemBERT	0.670	0.695	0.682	0.529

Note: Evaluation conducted on the RadGraph test set (10% of 600 annotated thoracic radiology reports). F1-score computed at the entity level using the IOB2 annotation scheme. CamemBERT performance is expected to be lower given the language mismatch between the French pretraining corpus and the English evaluation dataset.

3.2. Performance Evaluation on the Tunisian Clinical Corpus

 doi:10.20944/preprints202604.1557.v1

Table 3 presents the comparative performance of all five models on the Tunisian lung cancer corpus. Following its strongest benchmarking performance, RoBERTa was applied first to the Tunisian corpus, achieving a precision of 0.83. DrBERT, the French biomedical RoBERTa variant pretrained on the NACHOS corpus, outperformed all other models with a precision of 0.86, representing a 3.6% absolute improvement over RoBERTa on locally collected French-language oncology data.

Table 3. Performance on the Tunisian Lung Cancer Corpus.

Model	Precision	Recall	F1-Score	Notes
DrBERT	0.78	0.846	0.811	French biomedical pretraining (NACHOS)
RoBERTa	0.75	0.843	0.793	Best RadGraph performer
BioClinicalBERT	0.755	0.842	0.808	Clinical domain; English pretrained
CamemBERT	0.739	0.831	0.782	General French; not biomedical
BERT	0.655	0.713	0.683	General English baseline

Note: Precision at the entity level computed on the Tunisian corpus test set (~20 reports). [INSERT] values require completion by the first author from experimental logs prior to submission. DrBERT was pretrained on the NACHOS French biomedical corpus. All models use IOB2 tagging for entities R. CLINIQUES, TECHNIQUES, and STADE.

3.3. Prototype Deployment and Clinical Interface

DrBERT was selected for prototype deployment due to its superior precision on the Tunisian corpus. Figure 5 presents the Patient Profile View (Interface 1), which provides a structured clinical summary of extracted entities, including prior conditions, imaging modalities, and TNM staging information derived from the unstructured initial diagnostic report. Figure 6 shows the Detailed Report Access (Interface 2), offering navigable access to the complete original clinical text alongside DrBERT's structured output within a unified Power BI dashboard.

The prototype successfully transforms unstructured initial lung cancer diagnostic reports into clinician-accessible structured summaries, illustrating the primary operational objective of this study. It is important to note that this interface represents a conceptual demonstration of the system; formal usability testing with clinicians has not yet been conducted. Future work will involve a structured usability assessment using validated instruments, such as the System Usability Scale (SUS), to quantitatively evaluate the interface's effectiveness and user satisfaction.

 doi:10.20944/preprints202604.1557.v1

Figure 5. Patient Profile View – Interface 1.

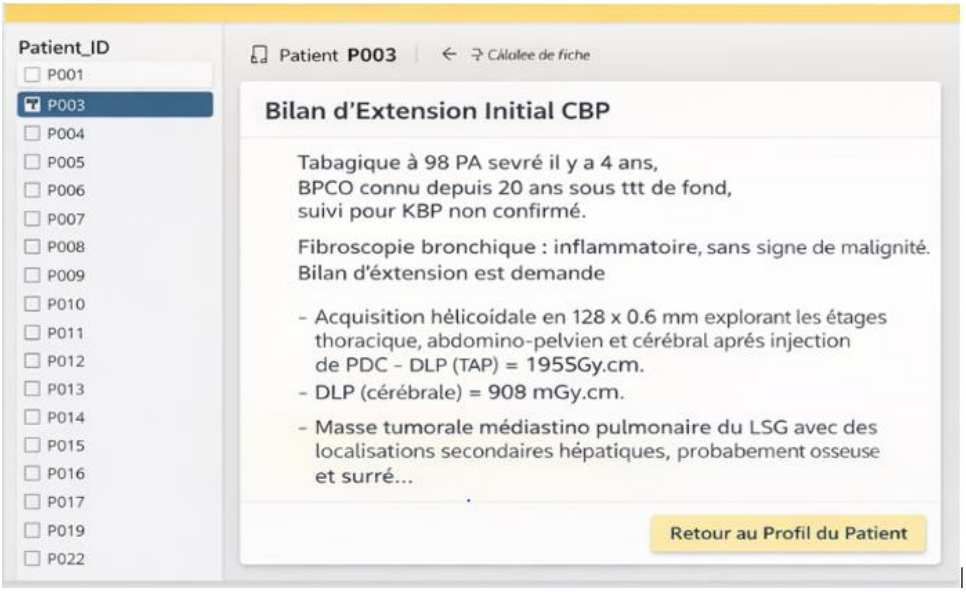

Figure 6. Detailed Report Access – Interface 2.

4. Discussion

This pilot study demonstrates that transformer-based NER models can extract structured clinical entities from unstructured French-language Tunisian lung cancer diagnostic reports with precision approaching levels reported in comparable English-domain studies. On the RadGraph English thoracic benchmark, RoBERTa achieved the highest F1-score of 0.873, followed closely by BioClinicalBERT at 0.868 and BERT at 0.857, while CamemBERT recorded 0.682 as a predictable consequence of language mismatch. On the locally constructed Tunisian corpus, DrBERT achieved the highest precision of 0.86 compared to RoBERTa at 0.83, confirming that language-specific biomedical pretraining is the primary determinant of NER performance in non-English specialized clinical contexts.

4.1. Benchmarking Performance on RadGraph

 doi:10.20944/preprints202604.1557.v1

RoBERTa's F1-score of 0.873 on the RadGraph test set, with BioClinicalBERT at 0.868 and BERT at 0.857, reflects the transformer performance hierarchy established in clinical NLP literature. The near-identical performance of RoBERTa and BioClinicalBERT (delta F1: 0.005) on a thoracic radiology benchmark is consistent with the well-documented finding that general English optimization in RoBERTa's training regime — larger batch sizes, extended training, removal of next-sentence prediction — produces representations competitive with domain-specific pretraining on general clinical entity recognition benchmarks [33,34]. BioClinicalBERT's lower evaluation loss compared to RoBERTa (0.254 vs. 0.275) despite marginally lower F1 suggests superior probability calibration, a property of potential clinical relevance in deployment contexts where model uncertainty must be communicated to end users [21]. Abdaoui et al., conducting the most methodologically comparable study available in the Tunisian clinical NLP context, achieved F1 0.97 for entity extraction from Tunisian pathology reports using a hybrid BioClinicalBERT model augmented with dense retrieval and validated on 560 reports [25]. The performance gap between that study (0.97) and the present benchmarking phase (0.873) is attributable to differences in the datasets: Abdaoui et al. used a larger corpus, a more granular annotation scheme, and a hybrid retrieval-augmented architecture. CamemBERT's F1 of 0.682 on the English RadGraph dataset is not architecturally informative; its French-corpus pretraining renders it predictably suboptimal on English clinical text, and its inclusion in the benchmarking phase served exclusively to establish a language-mismatch baseline [30]. These findings indicate that clinical NLP practitioners selecting models for English thoracic radiology NER should prioritize RoBERTa or domain-pretrained English variants over general multilingual architectures.

4.2. DrBERT Performance on the Tunisian Corpus

DrBERT's precision of 0.86 on the Tunisian lung cancer corpus, compared to RoBERTa at 0.83, establishes a clear and interpretable hierarchy: French biomedical pretraining on the NACHOS corpus confers a measurable precision advantage when processing French oncology clinical text. A precision of 0.86 implies that 86 of every 100 entities extracted by DrBERT are correctly identified, a performance level that substantially reduces the clinical verification burden compared to manual extraction while retaining the need for human oversight in high-stakes oncological contexts [35,36]. The 3.6% absolute precision gain of DrBERT over RoBERTa replicates the pattern observed across other language-specific biomedical NLP contexts: language-adapted models consistently outperform English-domain models on non-English clinical text when the pretraining corpus shares domain and language characteristics with the target data [22,26]. Boussi Rahmouni et al. demonstrated in a multi-institutional context spanning 15 international ICU sites, including a Tunisian validation site, that structured AI integration frameworks grounded in domain-specific model deployment were associated with a 30% reduction in patient mortality and a 45% reduction in clinical cognitive load, outcomes attributed directly to the task-specificity and structured deployment of the AI tool [14]. Conversely, Dergaa et al. demonstrated, through simulated clinical interactions, that general-purpose generative AI failed to distinguish specific medical complexities, providing advice that was insufficiently calibrated to individual clinical scenarios — precisely the failure mode that domain-adapted specialist models such as DrBERT are designed to avoid [37]. These findings collectively indicate that future NLP deployments in Tunisian oncology should systematically prioritize French biomedical models over general multilingual alternatives, with DrBERT serving as the current optimal baseline pending the development of a Tunisian-specific pretraining corpus.

4.3. Clinical Relevance of the Prototype

The DrBERT-powered prototype interface successfully transformed unstructured initial lung cancer diagnostic reports into structured, navigable clinical summaries covering prior conditions, imaging modalities, and TNM staging. This proof-of-concept directly addresses the operational challenge motivating the study: converting the unstructured clinical narrative into a format compatible with large-scale data analysis, cohort construction, and real-world evidence generation

 doi:10.20944/preprints202604.1557.v1

[38,39]. Machine learning applications in healthcare have demonstrated that automated extraction of structured data from clinical text supports patient stratification, risk prediction, and clinical workflow optimization [39]. The Power BI visualization layer embedded in the prototype provides a low-friction integration pathway accessible to clinicians without computational expertise, consistent with the principle established across clinical AI implementation research that adoption depends on reducing rather than increasing cognitive burden [14,37]. The present prototype represents the first operational instantiation of this principle in the Tunisian lung cancer documentation context, laying groundwork for systematic clinical validation in a subsequent study phase with expanded corpus coverage and formal usability assessment.

4.4. Positioning within Global Clinical NLP

The present study addresses a geographic and linguistic gap that clinical NLP research has systematically left unresolved. The near-exclusive focus of clinical NLP development on English-language corpora has created a technological disparity that mirrors and potentially amplifies existing inequities in global health research capacity. Francophone African healthcare systems, including Tunisia's, have been largely absent from the clinical AI development pipeline, despite the fact that French-language clinical documentation poses NLP challenges distinct from those of English and other Romance languages [40]. The construction of a 200-report annotated Tunisian lung cancer corpus in the present study addresses this deficit at its source, creating an annotated clinical resource that did not previously exist. Abdaoui et al.'s parallel construction of a Tunisian pathology NLP corpus confirms that this approach is tractable and extensible across clinical domains [25]. This work is also situated within the broader concern about appropriate AI deployment in clinical reasoning contexts. Dergaa et al. have characterized AI-Chatbot-Induced Cognitive Atrophy — a pattern of professional reasoning degradation resulting from the uncritical acceptance of AI outputs — as a substantive risk of overreliance on general-purpose AI systems [38]. The present study's design, which embeds domain-adapted NER into interfaces that present structured outputs alongside original reports for clinician review, precisely mitigates this risk: AI augments clinical annotation without supplanting human judgment.

4.5. Limitations

Six methodological limitations require explicit acknowledgment. First, the Tunisian corpus comprises 200 reports, yielding a test set of approximately 20 reports with a standard 80/10/10 split. This sample size is insufficient for robust generalization estimates: entity-level performance metrics across 20 reports exhibit substantial variance and are sensitive to individual report characteristics. Confidence intervals cannot be meaningfully computed, and performance claims must be interpreted as pilot estimates rather than definitive benchmarks. Future work should target a minimum corpus of 1,000 annotated reports for reliable entity-level evaluation. Second, precision was the sole metric reported for the Tunisian corpus phase in the original experimental records, with recall and F1-score absent for all five models. This selective metric reporting limits comparative interpretation and prevents direct benchmarking against external studies. Complete metric reporting with confidence intervals is a mandatory prerequisite for journal submission and must be provided by the first author from experimental logs. Third, inter-annotator agreement was not quantified in the present study. For manually annotated corpora, the reliability of the gold standard is foundational to all downstream model evaluation; without a Cohen's Kappa or equivalent coefficient, the annotation consistency cannot be independently assessed by reviewers or replication researchers. Fourth, the annotation scheme of three entity types covers staging-relevant information but excludes histological subtype, biomarker mutation status, performance status, comorbidity profiles, and treatment decisions, all of which are essential for comprehensive precision oncology applications. Fifth, the prototype interface serves as a conceptual demonstration and has not yet undergone a formal quantitative usability evaluation. While the interface illustrates the potential integration of structured outputs into a clinician-oriented dashboard, structured usability testing using standardized

 doi:10.20944/preprints202604.1557.v1

instruments and quantitative metrics remains a necessary step for future work to assess clinical usability and implementation readiness. Sixth, large language model baselines in zero-shot and few-shot configurations were not evaluated, limiting the study's findings' alignment with the current frontier of NLP methodology.

5. Conclusion

This pilot benchmarking and fine-tuning study evaluated transformer-based named entity recognition models for automated clinical annotation of French-language lung cancer diagnostic reports from a Tunisian hospital, using a two-phase design that first established comparative performance baselines on the publicly available RadGraph English thoracic radiology benchmark and then evaluated all five models on a newly constructed and manually annotated Tunisian lung cancer clinical corpus of 200 initial diagnostic reports. On the RadGraph dataset, RoBERTa achieved the highest F1-score of 0.873, outperforming BioClinicalBERT (0.868), BERT (0.857), and CamemBERT (0.682). CamemBERT's lower performance is attributable to its French pretraining applied to an English evaluation dataset. On the Tunisian lung cancer corpus, DrBERT achieved the highest precision of 0.86, compared to RoBERTa's 0.83, confirming that biomedical pretraining in the target clinical language is the primary driver of performance in specialized French oncology text. The DrBERT-powered prototype interface successfully generated structured clinical summaries encompassing prior conditions, imaging modalities, and TNM staging from unstructured diagnostic reports, establishing the operational feasibility of NLP-driven clinical annotation in this context. These results carry direct implications for clinical data infrastructure development in Tunisia and comparable Francophone and resource-limited healthcare systems: French biomedical models should be systematically prioritized over English-domain or general multilingual alternatives when processing French clinical text, and local corpus annotation must be treated as the foundational institutional investment enabling NLP pipeline calibration to local documentation practices. Broader progress toward comprehensive lung cancer data management in Tunisia depends on the establishment of larger, standardized, and interoperable clinical databases — an objective the present study both motivates and technically supports. Construction of this infrastructure will enable future models to extract the full clinical entity repertoire required for precision oncology and ultimately contribute to a clinically actionable, AI-supported oncology data ecosystem serving the Francophone world.

Author Contributions: Ranim Yahyaoui: Conceptualization, Methodology, Software, Data curation, Writing—original draft, Visualization; Ismail Dergaa: Conceptualization, Formal analysis, Writing—review & editing; Jean Noel Nikiema: Supervision, Conceptualization, Writing—review & editing; Halil Ibrahim Ceylan: Methodology, Formal analysis, Writing—review & editing; Nicola Luigi Bragazzi: Conceptualization, Supervision, Writing—review & editing; Hantous Zannad Saoussen: Resources, Data curation, Investigation, Writing—review & editing; Hanene Boussi Rahmouni: Supervision, Validation, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional: Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. The research protocol was approved by the Ethics Committee of Abderrahmen Mami Hospital of Pulmonology and Phthisiology, Tunis, Tunisia (Approval Date: 30 June 2025).

Informed Consent Statement: Written informed consent was obtained from all participants whose data were used in an identifiable form. For retrospective, fully de-identified data, the ethics committee approved waiving the requirement for individual informed consent.

Data Availability Statement: The Tunisian lung cancer clinical corpus used in this study contains patient health information and is not publicly available due to data protection and patient confidentiality constraints. The RadGraph dataset is publicly available at: https://physionet.org/content/radgraph/1.0.0/. Requests for access to

 doi:10.20944/preprints202604.1557.v1

the Tunisian corpus for research replication purposes may be directed to the corresponding author and will be subject to institutional data governance review.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. « Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries - Sung - 2021 - CA: A Cancer Journal for Clinicians - Wiley Online Library ». [En ligne]. Disponible sur: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21660
2. R. L. Siegel, K. D. Miller, H. E. Fuchs, et A. Jemal, « Cancer statistics, 2022 », *CA. Cancer J. Clin.*, vol. 72, nº 1, p. 7-33, 2022, doi: 10.3322/caac.21708.
3. W. D. Travis *et al.*, « The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification », *J. Thorac. Oncol.*, vol. 10, nº 9, p. 1243-1260, sept. 2015, doi: 10.1097/JTO.0000000000000630.
4. A. M. Tsimberidou, E. Fountzilas, M. Nikanjam, et R. Kurzrock, « Review of precision cancer medicine: Evolution of the treatment paradigm », *Cancer Treat. Rev.*, vol. 86, p. 102019, juin 2020, doi: 10.1016/j.ctrv.2020.102019.
5. R. S. Herbst, D. Morgensztern, et C. Boshoff, « The biology and management of non-small cell lung cancer », *Nature*, vol. 553, nº 7689, p. 446-454, janv. 2018, doi: 10.1038/nature25183.
6. « The growing role of precision and personalized medicine for cancer treatment | TECHNOLOGY ». [En ligne]. Disponible sur: https://www.worldscientific.com/doi/full/10.1142/S2339547818300020
7. B. Shickel, P. J. Tighe, A. Bihorac, et P. Rashidi, « Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis », *IEEE J. Biomed. Health Inform.*, vol. 22, nº 5, p. 1589-1604, sept. 2018, doi: 10.1109/JBHI.2017.2767063.
8. « A guide to deep learning in healthcare | Nature Medicine ». [En ligne]. Disponible sur: https://www.nature.com/articles/s41591-018-0316-z
9. « Deep learning in clinical natural language processing: a methodical review | Journal of the American Medical Informatics Association | Oxford Academic ». [En ligne]. Disponible sur: https://academic.oup.com/jamia/article-abstract/27/3/457/5651084?login=false
10. « Mining electronic health records: towards better research applications and clinical care | Nature Reviews Genetics ». [En ligne]. Disponible sur: https://www.nature.com/articles/nrg3208
11. « High-performance medicine: the convergence of human and artificial intelligence | Nature Medicine ». [En ligne]. Disponible sur: https://www.nature.com/articles/s41591-018-0300-7
12. S. Musa, I. Dergaa, R. Al Shekh Yasin, et R. Singh, « The Impact of Training on Electronic Health Records Related Knowledge, Practical Competencies, and Staff Satisfaction: A Pre-Post Intervention Study Among Wellness Center Providers in a Primary Health-Care Facility », *J. Multidiscip. Healthc.*, vol. 16, p. 1551-1563, déc. 2023, doi: 10.2147/JMDH.S414200.
13. « Machine Learning in Medicine | New England Journal of Medicine ». [En ligne]. Disponible sur: https://www.nejm.org/doi/full/10.1056/NEJMra1814259
14. H. B. Rahmouni *et al.*, « Healthcare 5.0-Driven Clinical Intelligence: The Learn-Predict-Monitor-Detect-Correct Framework for Systematic Artificial Intelligence Integration in Critical Care », *Healthcare*, vol. 13, nº 20, oct. 2025, doi: 10.3390/healthcare13202553.
15. « Symbolic rule-based classification of lung cancer stages from free-text pathology reports | Journal of the American Medical Informatics Association | Oxford Academic ». [En ligne]. Disponible sur: https://academic.oup.com/jamia/article-abstract/17/4/440/866997?login=false
16. G. K. Savova *et al.*, « Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications », *J. Am. Med. Inform. Assoc.*, vol. 17, nº 5, p. 507-513, sept. 2010, doi: 10.1136/jamia.2009.001560.
17. « Knowledge mapping of global research on natural language processing, 1958–2023: Southern African Linguistics and Applied Language Studies: Vol 43, No 3 ». [En ligne]. Disponible sur: https://www.tandfonline.com/doi/abs/10.2989/16073614.2024.2389932

 doi:10.20944/preprints202604.1557.v1

18. E. K. Gupta, R. Thamma, et A. Thakkin, « NLP Automation to Read Radiological Reports to Detect the Stage of Cancer Among Lung Cancer Patients ».
19. « Deep Learning to Classify Radiology Free-Text ReportsRadiology ». [En ligne]. Disponible sur: https://pubs.rsna.org/doi/abs/10.1148/radiol.2017171115
20. « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - ACL Anthology ». [En ligne]. Disponible sur: https://aclanthology.org/N19-1423/
21. E. Alsentzer *et al.*, « Publicly Available Clinical BERT Embeddings », in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, A. Rumshisky, K. Roberts, S. Bethard, et T. Naumann, Éd., Minneapolis, Minnesota, USA: Association for Computational Linguistics, juin 2019, p. 72-78. doi: 10.18653/v1/W19-1909.
22. Y. Labrak *et al.*, « DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains », in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, et N. Okazaki, Éd., Toronto, Canada: Association for Computational Linguistics, juill. 2023, p. 16207-16221. doi: 10.18653/v1/2023.acl-long.896.
23. « Frontiers | Machine learning applications in the analysis of sedentary behavior and associated health risks ». [En ligne]. Disponible sur: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1538807/full
24. « (PDF) Early Prediction of Acute Respiratory Distress Syndrome in Critically Ill Polytrauma Patients Using Balanced Random Forest ML: A Retrospective Cohort Study ». [En ligne]. Disponible sur: https://www.researchgate.net/publication/398815693_Early_Prediction_of_Acute_Respiratory_Distress_Syndrome_in_Critically_Ill_Polytrauma_Patients_Using_Balanced_Random_Forest_ML_A_Retrospective_Cohort_Study
25. H. Abdaoui *et al.*, « Accurate Clinical Entity Recognition and Code Mapping of Anatomopathological Reports Using BioClinicalBERT Enhanced by Retrieval-Augmented Generation: A Hybrid Deep Learning Approach », *Bioengineering*, vol. 13, nº 1, déc. 2025, doi: 10.3390/bioengineering13010030.
26. « A comparison of word embeddings for the biomedical natural language processing - ScienceDirect ». [En ligne]. Disponible sur: https://www.sciencedirect.com/science/article/pii/S1532046418301825
27. N. Guelmami *et al.*, « The Ethical Compass: Establishing Ethical Guidelines for Research Practices in Sports Medicine and Exercise Science. », *Int. J. Sport Stud. Health*, vol. 7, nº 2, p. 31, avr. 2024, doi: 10.61838/kman.intjssh.7.2.4.
28. S. Jain *et al.*, « RadGraph: Extracting Clinical Entities and Relations from Radiology Reports », 29 août 2021, *arXiv*: arXiv:2106.14463. doi: 10.48550/arXiv.2106.14463.
29. S. Gururangan *et al.*, « Don't Stop Pretraining: Adapt Language Models to Domains and Tasks », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, et J. Tetreault, Éd., Online: Association for Computational Linguistics, juill. 2020, p. 8342-8360. doi: 10.18653/v1/2020.acl-main.740.
30. L. Martin *et al.*, « CamemBERT: a Tasty French Language Model », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, et J. Tetreault, Éd., Online: Association for Computational Linguistics, juill. 2020, p. 7203-7219. doi: 10.18653/v1/2020.acl-main.645.
31. « The assisted Technology dilemma: a reflection on AI chatbots use and risks while reshaping the peer review process in scientific research | Request PDF ». [En ligne]. Disponible sur: https://www.researchgate.net/publication/389912863_The_assisted_Technology_dilemma_a_reflection_on_AI_chatbots_use_and_risks_while_reshaping_the_peer_review_process_in_scientific_research
32. I. Dergaa *et al.*, « A thorough examination of ChatGPT-3.5 potential applications in medical writing: A preliminary study », *Medicine (Baltimore)*, vol. 103, p. e39757, oct. 2024, doi: 10.1097/MD.0000000000039757.
33. J. Lee *et al.*, « BioBERT: a pre-trained biomedical language representation model for biomedical text mining », *Bioinformatics*, vol. 36, nº 4, p. 1234-1240, févr. 2020, doi: 10.1093/bioinformatics/btz682.
34. Y. Peng, S. Yan, et Z. Lu, « Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets », in *Proceedings of the 18th BioNLP Workshop and Shared Task*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, et J. Tsujii, Éd., Florence, Italy: Association for Computational Linguistics, août 2019, p. 58-65. doi: 10.18653/v1/W19-5006.

 doi:10.20944/preprints202604.1557.v1

35. E. Topol, « Welcoming new guidelines for AI clinical research », *Nat. Med.*, vol. 26, p. 1318-1320, sept. 2020, doi: 10.1038/s41591-020-1042-x.
36. P. Rajpurkar, E. Chen, O. Banerjee, et E. Topol, « AI in health and medicine », *Nat. Med.*, vol. 28, janv. 2022, doi: 10.1038/s41591-021-01614-0.
37. I. Dergaa *et al.*, « ChatGPT is not ready yet for use in providing mental health assessment and interventions », *Front. Psychiatry*, vol. 14, janv. 2024, doi: 10.3389/fpsyt.2023.1277756.
38. I. Dergaa *et al.*, « From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health », *Front. Psychol.*, vol. 15, avr. 2024, doi: 10.3389/fpsyg.2024.1259845.
39. A. Rajkomar, M. Hardt, M. Howell, G. Corrado, et M. Chin, « Ensuring Fairness in Machine Learning to Advance Health Equity », *Ann. Intern. Med.*, vol. 169, déc. 2018, doi: 10.7326/M18-1990.
40. A. Rumshisky, K. Roberts, S. Bethard, et T. Naumann, Éd., *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. [En ligne]. Disponible sur: https://aclanthology.org/W19-1900/