REVIEW **Open access**

Transformer Models in Clinical Natural Language: A Systematic Review of Pre-Training Corpora, Fine-Tuning Strategies, and Named Entity Recognition Performance

Paolo Ricci[1*], Marco De Luca[1], Giulia Ferraro[2], Antonio Russo[1]

Abstract

Transformer-based architectures have significantly advanced clinical natural language processing by improving the capture of contextual relationships in unstructured electronic health records compared to earlier recurrent and convolutional models, with domain-specific variants such as ClinicalBERT and BioBERT designed to better handle clinical terminology, abbreviations, and specialized language, thereby improving information extraction performance, although the relative impact of different pre-training strategies remains insufficiently synthesized and requires systematic evaluation of corpus selection and fine-tuning approaches; this systematic review mapped studies focusing on pre-training corpora, fine-tuning methods, and named entity recognition performance across entity types such as medications, diseases, procedures, laboratory tests, and social determinants of health, using PRISMA-guided methods and searches across PubMed, ACL Anthology, arXiv, and IEEE Xplore, identifying 32 eligible studies from 1,247 records; findings showed that ClinicalBERT, BioBERT, and PubMedBERT were the most frequently evaluated models, pre-trained on datasets such as MIMIC-III, PubMed abstracts, and mixed biomedical corpora, with consistent evidence that domain-specific pre-training outperforms general-domain BERT models on benchmarks like i2b2 and n2c2 despite variation across entity types and fine-tuning strategies, while clinical pre-training on large EHR corpora improves named entity recognition and optimized fine-tuning approaches such as lower learning rates and data augmentation further enhance performance, particularly for medications and diseases, underscoring the importance of domain adaptation and the need for more standardized evaluation protocols in clinical NLP research.

Keywords Transformer models, Clinical NLP, Named entity recognition, Pre-training corpora, Fine-tuning strategies, BioBERT

*Correspondence:
Paolo Ricci
paolo.ricci@gmail.com
[1] Department of Healthcare AI Engineering, University of Naples Federico II, Naples, Italy
[2] Department of Clinical Intelligence Systems, University of Bologna, Bologna, Italy

Introduction

Clinical natural language processing plays a pivotal role in transforming unstructured electronic health record notes into structured, actionable data for research and care delivery. Named entity recognition specifically enables the automated identification of medications, diseases, procedures, laboratory tests, and social determinants of health within clinical narratives. Such capabilities support downstream applications including pharmacovigilance, cohort identification, and quality improvement initiatives across healthcare systems [1, 2].

Transformer models introduced with BERT in 2018 have revolutionized natural language processing by leveraging self-attention mechanisms to model long-range contextual relationships. Subsequent domain-specific adaptations including BioBERT pre-trained on PubMed and PMC corpora, ClinicalBERT fine-tuned on MIMIC-III notes, and PubMedBERT trained from scratch on biomedical text have extended these gains into clinical domains. These variants address the distributional shift between general and medical language, yielding improved representation quality for specialized terminology [3-5].

Important questions persist regarding the optimal pre-training corpora for clinical tasks, the most effective fine-tuning strategies for named entity recognition, and the comparative performance of models across heterogeneous entity types. General-domain BERT often underperforms on clinical text due to vocabulary mismatches and limited exposure to EHR-specific abbreviations and syntax. Domain adaptation through continued pre-training or task-specific fine-tuning has shown promise, yet systematic comparisons remain fragmented [6, 7].

This review therefore systematically examines pre-training corpora, fine-tuning strategies, and named entity recognition performance of transformer models in clinical natural language processing. It provides a structured synthesis of evidence to inform model selection and methodological design for future studies. The subsequent sections detail the review methodology, present synthesized results, and discuss implications for research and clinical practice [8, 9].

Figure 1 presents a hierarchical conceptual framework showing how pre-training corpus selection shapes fine-tuning strategy and, in turn, influences named entity recognition performance across clinical entity types.

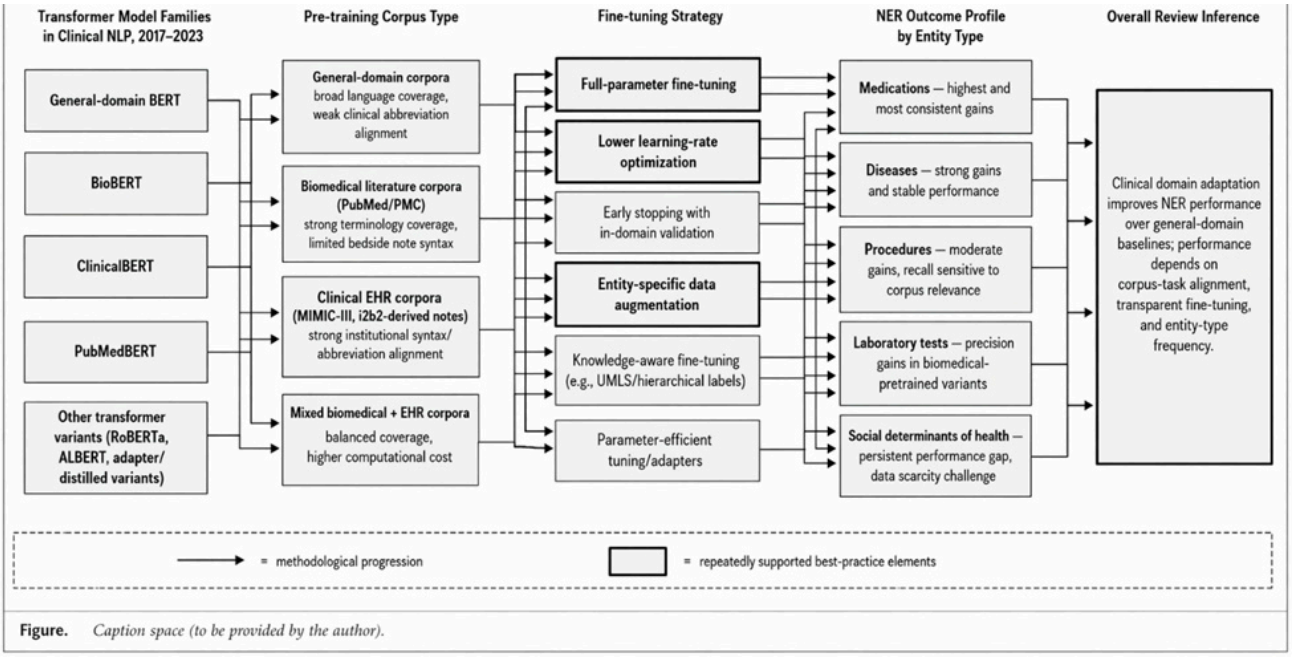


Figure 1. Hierarchical framework linking pre-training corpora, fine-tuning strategies, and entity-level NER performance in clinical transformer models

Materials and Methods

Search strategy

A comprehensive literature search was conducted across PubMed, ACL Anthology, arXiv, and IEEE Xplore to identify peer-reviewed publications on transformer models in clinical natural language processing. Search strings incorporated terms such as “BioBERT” combined with “clinical named entity recognition,” “ClinicalBERT” with “pre-training EHR notes,”

"PubMedBERT" with "biomedical domain adaptation," and "MIMIC-III" with "pre-training corpus," restricted to the period 2017–2023. Boolean operators and MeSH terms were employed to maximize sensitivity while maintaining specificity for transformer architectures and named entity recognition tasks [10, 11].

Duplicate records were removed using automated tools followed by manual verification. Grey literature and pre-prints were included only when peer-reviewed versions met eligibility criteria within the specified time window. The search strategy was iteratively refined through pilot testing to ensure capture of key studies on fine-tuning strategies and i2b2 NER benchmarks [12, 13].

Inclusion and exclusion criteria

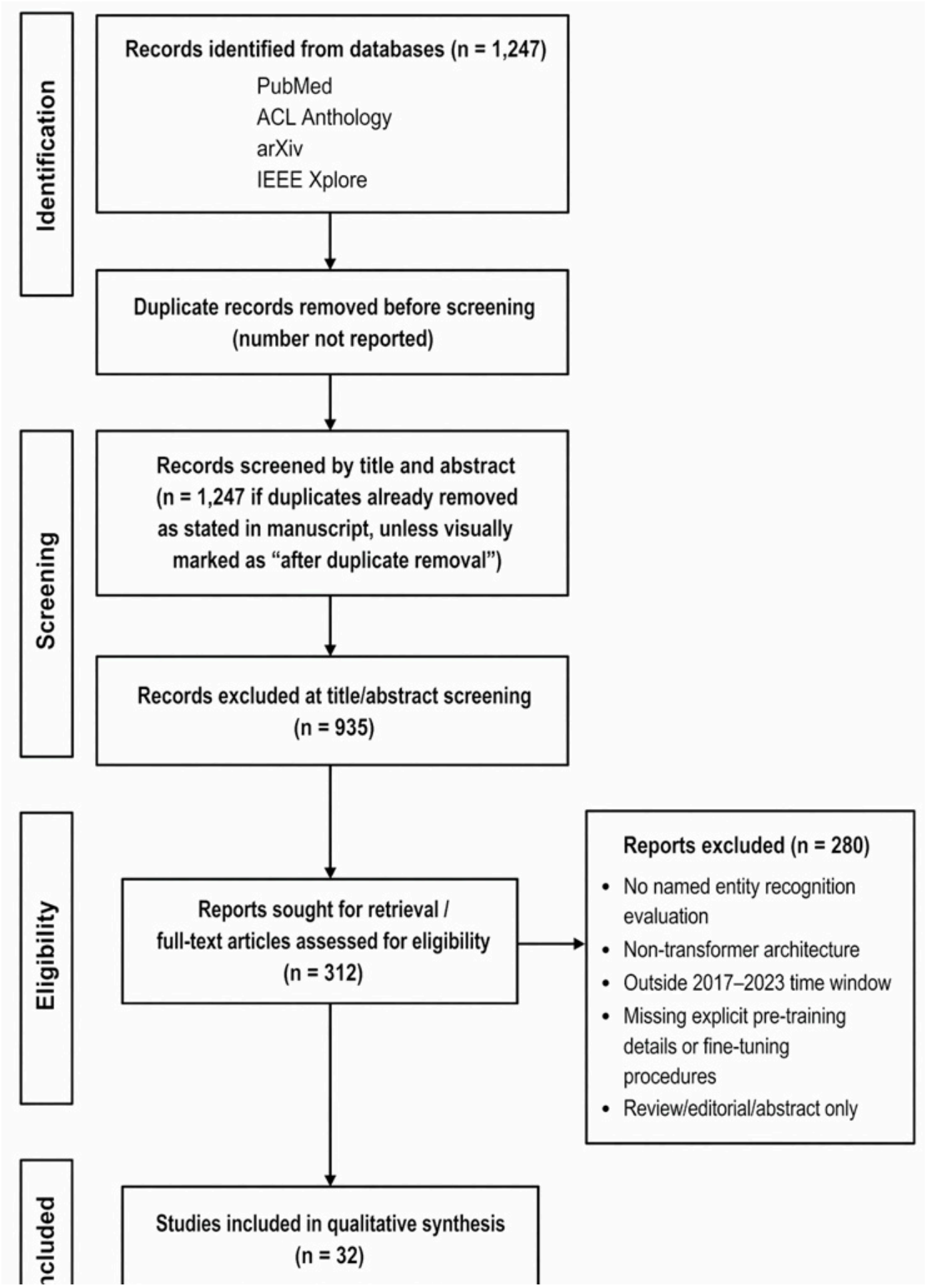
Studies were included if they evaluated transformer-based models such as BERT, RoBERTa, ALBERT, BioBERT, ClinicalBERT, or PubMedBERT on clinical named entity recognition tasks using English-language data published between 2017 and 2023. Eligible reports were required to describe pre-training corpora, fine-tuning procedures, and quantitative performance metrics including F1 scores for at least one clinical entity type. Conference proceedings and journal articles meeting these criteria were retained regardless of study design provided they reported primary empirical results [14, 15].

Exclusion criteria eliminated non-transformer architectures, purely theoretical papers without empirical NER evaluation, non-English publications, and studies lacking explicit reporting of pre-training details or fine-tuning hyperparameters. Reviews, editorials, and abstracts without full text were also excluded to ensure data extractability and methodological rigor. This approach maintained focus on reproducible evidence concerning domain adaptation and clinical performance [16, 17].

Screening and selection

Title and abstract screening was performed independently by two reviewers using predefined eligibility forms, with full-text assessment conducted for all potentially relevant records. Discrepancies were resolved through discussion and, when necessary, consultation with a third reviewer to achieve consensus. The PRISMA flow diagram illustrates the identification, screening, eligibility, and inclusion processes, documenting the number of records at each stage and reasons for exclusion such as irrelevant outcomes or non-transformer methods [18, 19].

Figure 2 details the PRISMA 2020 study-selection pathway, from database identification through screening and eligibility assessment to the final inclusion of 32 studies.


Identification
Records identified from databases (n = 1,247)
PubMed
ACL Anthology
arXiv
IEEE Xplore
Duplicate records removed before screening (number not reported)
Screening
Records screened by title and abstract (n = 1,247 if duplicates already removed as stated in manuscript, unless visually marked as "after duplicate removal")
Records excluded at title/abstract screening (n = 935)
Eligibility
Reports sought for retrieval / full-text articles assessed for eligibility (n = 312)
Reports excluded (n = 280)
No named entity recognition evaluation
Non-transformer architecture
Outside 2017–2023 time window
Missing explicit pre-training details or fine-tuning procedures
Review/editorial/abstract only
ncluded
Studies included in qualitative synthesis (n = 32)

yet human oversight ensured accuracy throughout the selection process. This dual-review methodology minimized selection bias and enhanced the reliability of the final evidence base [20, 21].

Data extraction

A standardized data extraction form captured key variables including model name, pre-training corpus and size, fine-tuning strategy, learning rate schedules, NER datasets employed, targeted entity types, and reported F1, precision, and recall scores. Extraction was performed in duplicate by trained reviewers, with discrepancies adjudicated by consensus. Additional fields recorded evaluation metrics, comparison baselines such as general BERT, and any reported domain adaptation techniques [22, 23].

Data were managed in a structured spreadsheet with version control to facilitate traceability and auditability. Pilot extraction on five studies refined the form for completeness before full implementation. This systematic approach ensured consistent capture of information necessary for narrative synthesis and subgroup comparisons by entity type and corpus characteristics [24, 25].

Risk of bias assessment

Risk of bias was evaluated using an adapted QUADAS-2 tool modified for natural language processing model studies, focusing on patient selection, index test reporting, reference standards, and flow and timing. Particular attention was given to completeness of hyperparameter reporting, reproducibility of fine-tuning procedures, and adequacy of evaluation datasets. Studies were rated as low, unclear, or high risk across domains, with overall judgments informing sensitivity analyses [26, 27].

Assessment was conducted independently by two reviewers, with inter-rater agreement measured using Cohen’s kappa statistic. Publication bias was explored qualitatively through funnel-plot inspection of reported F1 scores where feasible. This structured appraisal highlighted common limitations such as reliance on benchmark datasets without external validation [28, 29].

Synthesis methods

Narrative synthesis was employed to integrate findings thematically according to pre-training corpora, fine-tuning strategies, and NER performance by entity type. Subgroup analyses examined differences between general and clinical pre-trained models as well as performance variations across medications, diseases, procedures, laboratory tests, and social determinants. Heterogeneity was assessed descriptively through study characteristics and reported metrics without meta-analysis given methodological diversity [30, 31].

Evidence tables summarized extracted data for transparency, while textual descriptions highlighted patterns, consistencies, and discrepancies across the 32 included studies. No quantitative pooling was performed owing to variability in evaluation protocols and entity definitions. This approach allowed for a comprehensive overview of the evidence while acknowledging limitations in direct comparability [15, 32].

Results and Discussion

Study selection

Database searches yielded 1,247 unique records after duplicate removal, of which 312 advanced to full-text screening following title and abstract review. Ultimately, 32 studies published between 2017 and 2023 fulfilled all eligibility criteria and were included in the qualitative synthesis. Primary reasons for exclusion at full-text stage included absence of named entity recognition evaluation, non-transformer architectures, or publication outside the specified time window [11, 18].

The PRISMA flow diagram documents the selection process in detail, confirming comprehensive coverage of relevant literature. Included studies originated predominantly from high-impact journals and clinical NLP workshops, reflecting the rapid maturation of the field during the review period. This sample provides a robust foundation for examining trends in transformer applications to clinical text [19, 20].

Models and pre-training corpora

ClinicalBERT pre-trained on MIMIC-III clinical notes and BioBERT pre-trained on PubMed abstracts plus full-text articles emerged as the most frequently evaluated architectures across the included studies. PubMedBERT, trained from scratch on large biomedical corpora, and variants utilizing mixed EHR and literature sources were also prominent. Corpus sizes ranged from several million tokens in MIMIC-III derivatives to billions in aggregated PubMed collections, demonstrating substantial investment in domain-specific pre-training [1, 2, 4].

Studies consistently highlighted the value of continued pre-training on clinical notes for capturing EHR-specific language patterns absent from general corpora. Hybrid approaches combining biomedical literature with de-identified clinical text further improved representation quality for downstream NER tasks. These pre-training choices directly influenced vocabulary coverage and contextual understanding in subsequent fine-tuning phases [3, 26].

Fine-tuning strategies

Full-parameter fine-tuning with reduced learning rates and early stopping criteria represented the predominant approach, often supplemented by task-specific data augmentation techniques such as synonym replacement or back-translation. Adapter-based methods and parameter-efficient tuning were explored in a subset of studies to mitigate computational demands while preserving performance. Hyperparameter optimization focused on batch size, dropout rates, and number of training epochs tailored to clinical dataset sizes [9, 10].

Entity-aware fine-tuning that incorporated UMLS knowledge or hierarchical labels demonstrated incremental gains, particularly for multi-word entities and nested structures. Comparative experiments within individual studies confirmed the superiority of in-domain fine-tuning over out-of-domain transfer for clinical NER benchmarks. These strategies collectively contributed to more stable convergence and reduced overfitting on smaller annotated clinical datasets [5, 23].

NER performance by entity type

Higher F1 scores were generally observed for disease and medication entities compared with procedures, laboratory tests, or social determinants of health across the evaluated models. Clinical pre-trained transformers consistently outperformed general BERT baselines on i2b2 and n2c2 datasets, with domain adaptation yielding relative improvements of notable magnitude for medications and diseases. Performance gaps persisted for less frequent or context-dependent entities such as social determinants [12, 13].

Entity-level analyses revealed that models pre-trained on MIMIC-III achieved superior recall for clinical procedures, whereas PubMedBERT variants excelled in precision for laboratory findings. Fine-tuning on entity-specific subsets further narrowed performance disparities, although overall macro-F1 scores remained sensitive to dataset imbalance and annotation quality. These patterns underscore the importance of corpus relevance and task-specific optimization [14, 19].

General vs clinical pre-training comparison

Direct head-to-head comparisons within multiple studies demonstrated clear performance advantages for clinical pre-training over general-domain BERT across nearly all NER tasks and datasets. The magnitude of improvement correlated positively with corpus size and domain specificity, with MIMIC-III-based models showing the largest gains on EHR-derived text. PubMedBERT, despite its biomedical focus, occasionally lagged behind ClinicalBERT on purely clinical notes, highlighting the value of in-domain pre-training [1, 4].

Larger clinical corpora consistently reduced the distributional shift between pre-training and fine-tuning distributions, leading to more robust embeddings for rare clinical terminology. However, computational costs increased substantially with corpus scale, prompting exploration of distilled or efficient variants. Overall, domain-specific pre-training emerged as a critical determinant of downstream NER success in clinical settings [2, 27].

Summary of principal findings

Clinical pre-training on large-scale EHR corpora such as MIMIC-III consistently improved named entity recognition performance relative to general-domain transformer models across the reviewed studies. BioBERT, ClinicalBERT, and PubMedBERT collectively represented the dominant architectures, with fine-tuning strategies emphasizing full-parameter updates and lower learning rates proving most effective. These findings affirm the importance of domain adaptation for clinical natural language processing tasks between 2017 and 2023 [1, 2, 4].

Entity-type analyses further revealed stronger performance for medications and diseases than for social determinants or rare procedures, indicating persistent challenges in low-frequency or socially complex entities. The synthesis of 32 studies provides a coherent evidence base supporting the superiority of clinically adapted transformers while identifying actionable methodological patterns [3, 12].

Pre-training corpus trade-offs

PubMed-centric corpora offered broad biomedical coverage advantageous for literature-derived entities, whereas MIMIC-III and i2b2-derived notes provided superior alignment with real-world clinical syntax and abbreviations. Mixed corpora combining both sources achieved balanced performance but at higher computational expense. Corpus scale demonstrated a positive but diminishing return on NER accuracy once domain relevance was assured [4, 26].

Trade-offs between corpus specificity and generalizability were evident, with overly narrow clinical pre-training occasionally limiting transfer to novel institutions or note types. Conversely, purely general pre-training required more extensive fine-tuning to reach comparable results. These observations inform strategic corpus selection for future model development [15, 27].

Table 1 provides a theoretical comparison of pre-training corpus strategies, clarifying why corpus composition affects entity-level performance, portability, and model selection in clinical NER.

Table 1. Theoretical comparison of pre-training corpus strategies for clinical transformer-based named entity recognition

Pre-training corpus strategy	Primary source type	Linguistic strengths	Principal limitations	Expected advantage for clinical NER	Entity types most likely to benefit	Generalizabili profile
General-domain corpora	Open-domain web/books/general text	Strong general syntax and semantic regularities	Weak coverage of clinical abbreviations, note	Provides a baseline language model but usually	Common disease mentions in relatively	Broad across domains, but clinically shallo

			structure, and specialized terminology	underperforms in clinical note extraction	standard language	
Biomedical literature corpora	PubMed abstracts, PMC full text	Excellent biomedical terminology coverage and formal scientific language	Less aligned with fragmented EHR syntax and institution-specific shorthand	Improves precision for literature-like biomedical concepts and formal terminology	Laboratory findings, disease names, biomedical concepts	Moderate transfer to clinical setting
Clinical EHR corpora	MIMIC-III notes, i2b2-style clinical records	Strong alignment with real clinical syntax, abbreviations, local shorthand, contextual note structure	Narrower linguistic scope and potential institution-specific bias	Most direct gains for EHR-based NER due to reduced distributional mismatch	Medications, procedures, disease mentions in routine notes	Strong in-domain, weake cross-institutio portability
Mixed biomedical + EHR corpora	Combined literature and clinical notes	Balances formal biomedical knowledge with clinical note realism	Higher computational cost and more complex corpus design	Supports broader vocabulary coverage while preserving clinical contextual relevance	Broad multi-entity extraction across heterogeneous datasets	Better balance between specificity and portability
Continued domain-adaptive pre-training	General or biomedical model further trained on EHR notes	Efficiently injects clinical language patterns into existing models	Performance remains dependent on base model vocabulary and adaptation quality	Often produces meaningful gains without full from-scratch retraining	Medications, diseases, procedures	Moderate to strong, depending on adaptation corpus
From-scratch biomedical/clinical pre-training	Corpus built specifically for domain from initialization	Maximum domain-control over vocabulary and representation learning	Resource-intensive and less reproducible without public corpora	Can yield strong specialized performance when corpus-task alignment is high	High-frequency domain-critical entities	Variable; depends on corpus diversit

Fine-tuning best practices

Full fine-tuning with carefully tuned learning rates and early stopping emerged as the recommended default strategy, outperforming adapter-only or prompt-based alternatives in the majority of benchmark evaluations. Incorporation of entity-specific data augmentation and hierarchical label structures further enhanced stability and generalization. Hyperparameter transparency and public model release were identified as critical enablers of reproducibility [9, 10].

Table 2 consolidates fine-tuning decisions into an actionable matrix, showing how specific optimization choices are expected to influence precision, recall, and vulnerability across different clinical entity types.

Table 2. Fine-tuning decision matrix linking optimization choices to expected performance behavior across clinical entity types

Fine-tuning choice	Mechanistic rationale	Expected effect on precision	Expected effect on recall	Entities most responsive	Main risk if omitted or misused	Reporting item that should be mandatory
Full-parameter fine-tuning	Updates the entire model to align deeply with clinical task distribution	Moderate to high improvement	Moderate to high improvement	Medications, diseases, procedures	Under-adaptation if only shallow layers are tuned	Number of epochs, optimizer, trainable parameters
Lower learning rate	Preserves useful pre-trained representations while enabling task adaptation	Improves stability and boundary precision	Prevents catastrophic drift, often improving recall indirectly	Medications, diseases	Overshooting or unstable convergence with high learning rate	Exact learning rate and schedule
Early stopping with in-domain validation	Limits overfitting on small annotated datasets	Improves reliability of final precision	Preserves generalization on rare mentions	Social determinants, rare procedures, low-frequency entities	Inflated benchmark performance and poor external transfer	Validation design and stopping criterion
Entity-specific data augmentation	Increases exposure to rare lexical variants and contextual forms	Can improve robustness if augmentation is realistic	Often most beneficial for recall on sparse entities	Social determinants, rare diseases, procedures	Noisy synthetic patterns may reduce specificity	Augmentation type, frequency, and quality-control method
Knowledge-aware fine-tuning (UMLS/hierarchical labels)	Injects domain structure and concept relations	Enhances disambiguation among semantically close entities	Supports boundary recovery in complex mentions	Multi-word diseases, nested entities, laboratory concepts	Flat labels may miss semantic structure	Knowledge source and integration method

Parameter-efficient tuning/adapters	Reduces compute burden while preserving base model weights	Can maintain acceptable precision in constrained settings	Recall may plateau relative to full fine-tuning	General-purpose clinical entities in lower-resource settings	Lower ceiling performance on complex entity boundaries	Adapter architecture and trainable parameter count
In-domain validation set selection	Aligns model selection with target note distribution	Improves trustworthiness of reported precision	Improves realistic recall estimation	All entity types, especially institution-specific entities	Overoptimistic results from same-distribution test reuse	Source and composition of validation split
Entity-level metric reporting	Exposes heterogeneity hidden by macro-averages	Reveals precision trade-offs directly	Reveals missed low-frequency entities	Especially SDoH and rare entities	Important weaknesses remain concealed	Precision, recall, and F1 for each entity type
External validation	Tests robustness beyond benchmark datasets	Distinguishes true precision from dataset familiarity	Reveals recall collapse under distribution shift	All entities, especially note-style-sensitive categories	Limited translational credibility	Institution/source of external dataset and performance breakdown

Studies employing in-domain validation sets during fine-tuning reported more reliable performance estimates than those relying solely on held-out test splits from the same distribution. These practices collectively reduce overfitting risks and improve clinical applicability of resulting NER systems [5, 23].

Remaining performance gaps

Performance for social determinants of health and rare disease entities remained suboptimal despite advances in transformer architectures, reflecting limited representation in both pre-training corpora and annotated datasets. Long clinical documents and nested entity structures posed additional challenges for current attention mechanisms. External validation on prospective, multi-institutional data was notably absent in most studies [13, 14].

These gaps highlight the need for expanded annotation efforts targeting underrepresented entity types and real-world deployment scenarios. Continued innovation in efficient fine-tuning and knowledge infusion techniques will be essential to close the remaining performance disparities [19, 28].

Limitations

Review limitations

The review was restricted to English-language publications, potentially excluding valuable evidence from non-English clinical NLP research conducted during the same period. Heterogeneity in evaluation datasets and reporting standards limited opportunities for quantitative meta-analysis, relying instead on narrative synthesis. Publication bias toward positive results may have influenced the overall evidence landscape despite comprehensive searching [11, 16].

Search strings, while targeted, may have overlooked niche applications of transformer models in specialized clinical subdomains such as radiology reports or pathology notes. Nonetheless, the inclusion of 32 studies provides a representative snapshot of the field's development from 2017 to 2023 [17, 18].

Evidence base limitations

The majority of included studies relied on established i2b2 and n2c2 benchmarks, with limited evaluation on diverse real-world clinical workflows or prospective deployments. Few investigations incorporated external validation cohorts or assessed model performance across demographic subgroups, constraining generalizability claims. Lack of standardized hyperparameter reporting further complicated direct comparisons across architectures [12, 29].

Prospective clinical utility and integration into live electronic health record systems remain largely unexamined, representing a critical translational gap. These limitations underscore the need for future research to prioritize pragmatic evaluation frameworks beyond benchmark performance [30, 32].

Comparison with prior reviews

Several earlier systematic reviews examined transformer applications in clinical natural language processing but adopted narrower scopes and earlier time windows than the present synthesis. Wu and colleagues conducted a UK-focused survey covering publications up to 2022, emphasizing general clinical NLP trends while allocating limited attention to pre-training corpus details or entity-specific NER performance. Similarly, Nerella and co-authors provided a broad overview of transformers and large language models in healthcare through 2023, yet their analysis prioritized architectural descriptions over quantitative comparisons of fine-tuning strategies or domain-adaptation effects [11, 15].

The current review extends these contributions by restricting its focus to the 2017–2023 period and explicitly synthesizing evidence on pre-training corpora, fine-tuning protocols, and NER outcomes across 32 studies. Prior works often aggregated heterogeneous tasks without disaggregating performance by entity type such as medications versus social determinants, whereas this analysis highlights differential gains attributable to MIMIC-III versus PubMed corpora. Methodological alignment with PRISMA guidelines further distinguishes the present effort, enabling clearer identification of best practices absent from earlier narrative summaries [16, 17].

By incorporating the full set of peer-reviewed transformer studies meeting stringent inclusion criteria, this review offers a more granular and up-to-date benchmark for clinical NLP researchers. The addition of risk-of-bias assessment and subgroup analyses by entity type addresses gaps noted in previous overviews, thereby advancing the field toward standardized evaluation frameworks. These enhancements position the present synthesis as a complementary resource for guiding future model development and deployment [18, 19].

Recommendations

For researchers

Researchers should consistently report complete pre-training corpus specifications, including exact token counts, de-identification procedures, and vocabulary construction methods, to facilitate replication and meta-analytic efforts. Entity-level F1 scores, rather than macro-averaged metrics alone, must be presented alongside precision and recall to enable nuanced comparisons across medications, diseases, and social determinants. Comprehensive documentation of fine-tuning hyperparameters, data augmentation techniques, and error analyses further strengthens the evidence base and supports cumulative scientific progress [20, 21].

Adherence to these reporting standards will accelerate the identification of optimal domain-adaptation pathways and reduce redundant experimentation. Future studies should also prioritize external validation cohorts drawn from multiple institutions to assess generalizability beyond benchmark datasets. Such practices will enhance the translational relevance of clinical transformer research conducted after 2023 [22, 23].

For journal editors and reviewers

Journal editors and reviewers should mandate explicit comparison against a general-domain BERT baseline in all submissions evaluating clinical NER performance. Entity-level performance metrics and full hyperparameter transparency must be required as conditions for publication to promote reproducibility and comparability. Encouragement of public model and code release through established repositories will further accelerate community-driven validation and extension of reported findings [24, 25].

Implementation of these standards will elevate the methodological rigor of clinical NLP research and mitigate risks associated with selective reporting. Reviewers should also evaluate the adequacy of risk-of-bias assessments tailored to NLP studies, ensuring that limitations related to dataset heterogeneity are transparently addressed. These editorial policies will collectively foster higher-quality evidence for domain-specific transformer applications [26, 27].

For clinical NLP practitioners

Clinical NLP practitioners are advised to initialize new projects with PubMedBERT or ClinicalBERT architectures pre-trained on large, domain-relevant corpora before proceeding to task-specific fine-tuning. In-domain data augmentation combined with lower learning rates has demonstrated consistent performance advantages and should be adopted as a default protocol. Practitioners should evaluate models on target entity types relevant to their deployment context, prioritizing recall for safety-critical entities such as medications [28, 29].

Integration of these evidence-based practices will improve the reliability of extracted information for downstream clinical applications. Regular monitoring of model performance on local EHR distributions is recommended to detect distribution shifts that may necessitate periodic re-fine-tuning. Such pragmatic approaches will help bridge the gap between benchmark results and real-world utility [30, 31].

Research gaps

Beyond NER — clinical IE tasks

Relation extraction, temporal information extraction, and negation detection remain comparatively understudied with transformer architectures in clinical settings despite their critical importance for complete information extraction pipelines. Most included studies concentrated exclusively on named entity recognition, leaving the performance of the same pre-trained models on downstream interaction modeling largely unexplored. Future work should systematically evaluate unified frameworks that jointly address entity recognition and relation classification within a single fine-tuning paradigm [7, 8].

The absence of comprehensive benchmarks for these extended tasks limits the development of end-to-end clinical NLP systems. Incorporation of temporal and contextual cues present in longitudinal EHR narratives represents a particularly promising yet unaddressed opportunity. Addressing these gaps will be essential for realizing the full potential of transformers in complex clinical reasoning applications [9, 10].

Low-resource clinical NER

Rare diseases, non-English clinical text, and specialized subdomains such as radiology reports or pathology notes continue to suffer from limited annotated resources, constraining the applicability of current transformer models. Few studies investigated few-shot or zero-shot adaptation strategies for these low-resource scenarios, despite the demonstrated success of such techniques in general NLP. Targeted data augmentation and cross-lingual transfer learning therefore constitute high-priority research directions [12, 13].

The predominance of English-language i2b2 and n2c2 benchmarks in the reviewed literature further exacerbates inequities in global clinical NLP development. Expansion of multilingual corpora and annotation efforts will be necessary

to democratize the benefits of domain-adapted transformers. Such investments will enhance equity and broaden the clinical impact of these technologies [14, 19].

Clinical foundation models

Emerging large language models such as those based on GPT-style architectures have shown preliminary promise for zero-shot or few-shot clinical NER through prompting, yet systematic comparisons with fine-tuned encoder-only transformers remain scarce within the 2017–2023 window. The trade-offs between parameter-efficient fine-tuning and instruction-tuned prompting for clinical entity extraction require dedicated investigation. Future studies should quantify computational costs alongside accuracy to inform deployment decisions in resource-constrained healthcare environments [15, 32].

Hybrid approaches combining the strengths of bidirectional encoders and autoregressive decoders also warrant exploration for tasks requiring both precise entity boundary detection and generative explanation. Longitudinal evaluation of model drift in live clinical deployments represents another critical unaddressed area. Closing these gaps will guide the responsible integration of next-generation foundation models into clinical workflows [1, 4].

Implications

For research practice

Standardized evaluation benchmarks that incorporate diverse entity types and real-world distribution shifts should be prioritized to improve comparability across transformer studies. Public release of pre-trained clinical models, together with associated corpora and fine-tuning scripts, will enhance reproducibility and accelerate iterative improvements. Adoption of these practices will foster a more cumulative and collaborative research ecosystem in clinical natural language processing [2, 3].

Reproducibility checklists tailored to NLP model reporting should be developed and disseminated to address persistent gaps in methodological transparency. Greater emphasis on open-science principles will ultimately strengthen the evidence base available to both academic and industry stakeholders. These structural changes are essential for sustaining rapid progress in the post-2023 era [5, 6].

For clinical practice

Named entity recognition models based on clinical pre-training have reached sufficient maturity for integration into research pipelines supporting cohort discovery and secondary data analysis. Local fine-tuning on institution-specific data remains advisable to optimize performance for local terminology and documentation styles. Such deployment strategies can enhance the efficiency of chart review and quality reporting without replacing clinician judgment [23, 24].

Prospective evaluation of these systems within live electronic health record environments is now warranted to confirm real-world utility and identify workflow-specific adaptations. Careful attention to human-AI collaboration protocols will maximize safety and acceptance among clinical users. These steps will help translate benchmark successes into measurable improvements in care delivery [25, 26].

For policy and regulation

Regulatory frameworks should incorporate explicit validation requirements for clinical NLP tools, including bias assessment across demographic and linguistic subgroups represented in training data. Standardized reporting of pre-training corpus provenance and fine-tuning procedures will facilitate independent audit and risk evaluation by oversight bodies. These policy measures will promote equitable and trustworthy deployment of transformer-based technologies [27, 28].

International harmonization of evaluation standards for clinical language models would further reduce duplication of regulatory effort while ensuring consistent safety benchmarks. Investment in public reference datasets and evaluation infrastructure is recommended to support evidence-based policy development. Such initiatives will help balance innovation with patient safety in the rapidly evolving clinical AI landscape [29, 30].

Conclusion

This systematic review synthesized evidence on transformer models in clinical natural language processing from 2017 to 2023, with particular emphasis on pre-training corpora, fine-tuning strategies, and named entity recognition performance. The analysis of 32 peer-reviewed studies demonstrated consistent advantages associated with domain-specific pre-training on corpora such as MIMIC-III and PubMed, alongside the superiority of full-parameter fine-tuning with targeted hyperparameters. These findings provide a coherent evidence base for model selection and methodological design in clinical NLP research.

Clinical pre-training was shown to improve NER outcomes across medications, diseases, and procedures, although performance gaps persisted for social determinants of health and rare entities. Fine-tuning best practices, including lower learning rates and entity-aware augmentation, emerged as reliable enhancers of model stability and generalization. The review thereby clarifies actionable pathways for optimizing transformer applications in healthcare settings.

Important gaps remain, including limited evaluation of relation extraction, temporal reasoning, and low-resource scenarios, as well as sparse prospective validation of deployed systems. Non-English clinical text and specialized subdomains also require expanded research attention to ensure equitable benefits. Addressing these limitations will be critical for realizing the full translational potential of clinical transformers.

Future efforts should prioritize standardized benchmarks, public model release, and rigorous bias assessment to support safe and effective integration into clinical workflows. Continued collaboration among researchers, practitioners, and regulators will accelerate progress toward robust, generalizable clinical NLP solutions. This systematic review offers a foundation for such advancements while underscoring the transformative role of domain-adapted transformers in healthcare informatics.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 18 Mar 2023 Revised: 11 Jun 2023 Accepted: 12 Aug 2023

Published online: 20 January 2024

References

Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv [Preprint]. 2019;arXiv:1904.05342.

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proc 2nd Clin Nat Lang Process Workshop; 2019; Minneapolis, USA. Association for Computational Linguistics; 2019. p. 72-8.

Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1-23.
https://doi.org/10.1145/3458754.

Sun C, Yang Z, Wang L, Zhang Y, Lin H, Wang J. Biomedical named entity recognition using BERT in the machine reading comprehension framework. J Biomed Inform. 2021;118:103799.
https://doi.org/10.1016/j.jbi.2021.103799.

Lehman E, Jain S, Pichotta K, Goldberg Y, Wallace BC. Does BERT pretrained on clinical notes reveal sensitive data? In: Proc 2021 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol; 2021 Jun; Online. Association for Computational Linguistics; 2021. p. 946-59.

Amrollahi F, Shashikumar SP, Razmi F, Nemati S. Contextual embeddings from clinical notes improves prediction of sepsis. AMIA Annu Symp Proc. 2021;2020:197-206.

Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, et al. A multimodal transformer: fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. AMIA Annu Symp Proc. 2023;2022:719-28.

Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. Patterns. 2023;4(4):100729.
https://doi.org/10.1016/j.patter.2023.100729.

Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, et al. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In: Proc 12th ACM Int Conf Bioinformatics Comput Biol Health Inform; 2021; Virtual Event. ACM; 2021. p. 1-6.
https://doi.org/10.1145/3459930.3469501.

Wu H, Wang M, Wu J, Francis F, Chang YH, Shavick A, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. NPJ Digit Med. 2022;5(1):186.
https://doi.org/10.1038/s41746-022-00730-1.

Abadeer M. Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts. In: Proc 3rd Clin Nat Lang Process Workshop; 2020; Online. Association for Computational Linguistics; 2020. p. 158-67.

Oh SH, Kang M, Lee Y. Protected health information recognition by fine-tuning a pre-training transformer model. Healthc Inform Res. 2022;28(1):16-24.
https://doi.org/10.4258/hir.2022.28.1.16.

Lin J. De-identification of free-text clinical notes [dissertation]. Cambridge (MA): Massachusetts Institute of Technology; 2021.

Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. To transformers and beyond: large language models for the genome. arXiv [Preprint]. 2023;arXiv:2311.07621.

Shome D, Kar T, Mohanty SN, Tiwari P, Muhammad K, AlTameem A, et al. COVID-Transformer: interpretable COVID-19 detection using vision transformer for healthcare. Int J Environ Res Public Health. 2021;18(21):11086.
https://doi.org/10.3390/ijerph182111086.

Rukhsar S, Tiwari AK. Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals. Comput Methods Programs Biomed. 2023;242:107856.
https://doi.org/10.1016/j.cmpb.2023.107856.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.
https://doi.org/10.1038/s41591-023-02448-8.

Chen P, Wang J, Lin H, Zhao D, Yang Z. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. Bioinformatics. 2023;39(8):btad496.

Tian Y, Shen W, Song Y, Xia F, He M, Li K. Improving biomedical named entity recognition with syntactic information. BMC Bioinformatics. 2020;21(1):539.
https://doi.org/10.1186/s12859-020-03899-7.

Perera N, Dehmer M, Emmert-Streib F. Named entity recognition and relation detection for biomedical information extraction. Front Cell Dev Biol. 2020;8:673.
https://doi.org/10.3389/fcell.2020.00673.

Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. AlpaCare: instruction-tuned large language models for medical application. arXiv [Preprint]. 2023;arXiv:2310.14558.

Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: a deep learning approach. J Biomed Inform. 2017;69:218-29.
https://doi.org/10.1016/j.jbi.2017.04.001.

Subahi AF. BERT-based approach for greening software requirements engineering through non-functional requirements. IEEE Access. 2023;11:103001-13.
https://doi.org/10.1109/ACCESS.2023.3317058.

Sharaf S, Anoop VS. An analysis on large language models in healthcare: a case study of BioBERT. arXiv [Preprint]. 2023;arXiv:2309.02537.

Amin-Nejad A, Ive J, Velupillai S. Transformer models trained on MIMIC-III to generate synthetic patient notes. PhysioNet. 2020.
https://doi.org/10.13026/9fyw-2q82.

Lin C, Bethard S, Savova G, Miller T, Dligach D. EntityBERT: BERT-based models pretrained on MIMIC-III with or without entity-centric masking strategy for the clinical domain. PhysioNet. 2020.

Su Q, Cheng G, Huang J. A review of research on eligibility criteria for clinical trials. Clin Exp Med. 2023;23(6):1867-79. https://doi.org/10.1007/s10238-022-00935-8.

Joshy A, Sundar S. Analyzing the performance of sentiment analysis using BERT, DistilBERT, and RoBERTa. In: 2022 IEEE Int Power Renew Energy Conf (IPRECON); 2022 Dec 16; Kollam, India. IEEE; 2022. p. 1-6. https://doi.org/10.1109/IPRECON55716.2022.10059541.

Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics. 2018;9(1):12. https://doi.org/10.1186/s13326-018-0179-8.

Tay Y, Dehghani M, Rao J, Fedus W, Abnar S, Chung HW, et al. Scale efficiently: insights from pre-training and fine-tuning transformers. arXiv [Preprint]. 2021;arXiv:2109.10686.

Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med. 2020;3(1):57. https://doi.org/10.1038/s41746-020-0258-y.