ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Starved by scarcity: A three scarcities framework and decision roadmap for medical named entity recognition in low-resource languages

Elira Hoxha and Polina Çeço*

Department of Statistics and Applied Informatics, Faculty of Economy, University of Tirana, Tirana, Albania

Abstract

Named entity recognition (NER) is a foundational task in medical natural language processing (NLP), enabling the extraction of clinically relevant information from unstructured text and supporting downstream applications such as information retrieval and clinical decision support. While pre-trained language models such as BioBERT, ClinicalBERT, and PubMedBERT have achieved state-of-the-art performance for English medical NER, progress in low-resource languages remains fragmented and uneven, even though healthcare systems worldwide generate vast amounts of textual data whose value depends on the availability of language technologies. Existing surveys provide valuable overviews but largely treat low-resource settings as a single category, overlooking important differences in the constraints practitioners face. To address this gap, we conducted a systematic review of 46 studies on medical NER in low-resource languages published since 2020. Based on the synthesized evidence, we introduce the three scarcities framework, which conceptualizes low-resource medical NLP as the interaction among data, model, and infrastructure scarcities. We also propose a practitioner-oriented decision roadmap that maps appropriate modeling and data augmentation strategies to different resource configurations. Our review develops a structured taxonomy of approaches, including cross-lingual transfer, in-domain pre-training, annotation projection, back-translation, and large language model-based synthetic data generation, and shows that the effectiveness of techniques depends less on the method itself than on the underlying scarcity profile. We also identify persistent weaknesses in evaluation practices, including inconsistent reporting and limited use of statistical significance testing.

Keywords: Medical natural language processing; Named entity recognition; Low-resource languages; Large language models; Transformer models

***Corresponding author:**
Polina Çeço
(polina.ceco@unitir.edu.al)

Citation: Hoxha E, Çeço P. Starved by scarcity: A three scarcities framework and decision roadmap for medical named entity recognition in low-resource languages. *Artif Intell Health.* 2026;3(4):026260068. doi: 10.36922/AIH026260068

Received: June 24, 2026

Revised: July 10, 2026

Accepted: July 13, 2026

Published online: July 23, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

More than 80% of daily healthcare data is produced in unstructured textual formats. Clinical documentation encapsulates patient health information in ways that outgrow the processing capabilities of traditional database systems.[1,2] Therefore, natural language processing (NLP) has emerged at the center of efforts to render clinical documentation computationally useful, whether for healthcare delivery and decision support or for broad demographic studies.[3,4]

Named entity recognition (NER) occupies a distinctive position within the broader NLP pipeline in the medical domain. It is pivotal, as it not only extracts essential information but also provides the foundation for subsequent downstream tasks, such as relation extraction, question answering, automatic summarization, topic modeling, and information retrieval.[5,6] For English, the infrastructure around medical NER is now substantial. Domain-specific pre-trained language models (PLMs), such as biomedical bidirectional encoder representations from transformers (BioBERT)[7], ClinicalBERT[8], SciBERT[9], PubMedBERT[10], and BlueBERT[11], have consistently outperformed general-purpose models owing to their pre-training on large-scale biomedical corpora, including annotated datasets such as MIMIC-III, the n2c2 shared task, and the i2b2 corpora, as well as extensive unannotated biomedical text from PubMed Central. In contrast, comparable resources remain scarce for most other languages, leaving clinical natural language processing in other languages considerably underdeveloped.

Despite the growing urgency, no systematic review to date has examined this literature through a unified methodological lens, nor has the compound nature of low-resource conditions been formally theorized. We aim to address this gap via the present systematic Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-compliant survey of 46 publications on medical NER for low-resource languages spanning from 2020 onward. The scope is deliberately narrow: we considered only NER applied to pure clinical documentation, such as clinical notes, discharge summaries, radiology reports, and electronic medical records, while excluding patient-generated text, social media, and general biomedical literature that do not embody the intrinsic linguistic and structural challenges of medical text. Moreover, we restricted the scope to solutions that employed large language models (LLMs) or transformer-based models, given their dominant position in current NLP research and practice.[12-14] This reflects a scope decision rather than a claim of universal superiority, as rule-based and dictionary-based methods remain competitive in specific settings. We revisit this distinction in Section 6, where it directly informs research question (RQ) 3 by examining how resource configurations influence methodological choices.

To structure the inquiry, this review is organized around four RQs:

(i) RQ1: What model-level and data-level techniques have been applied to medical NER in low-resource language settings, and under what resource configurations does each approach demonstrate effectiveness?
(ii) RQ2: To what extent are performance results in this literature comparable, and what structural factors undermine cumulative scientific progress in the field?
(iii) RQ3: What resource configurations characterize the languages studied, and how does the nature of resource scarcity shape methodological choices?
(iv) RQ4: What practical guidance can be derived from the reviewed evidence, and what unresolved challenges remain, to support researchers building medical NER systems in low-resource settings?

Addressing RQ1 and RQ2 yielded a structured taxonomy of model- and data-level techniques, encompassing fine-tuning, domain-adaptive pre-training, cross-lingual transfer (CLT), few-shot learning, back-translation, entity replacement, semantic similarity augmentation, annotation projection, and LLM-based synthetic generation. It also provided a critical analysis of the evaluation landscape, documenting the incomparability of F1 scores across papers and the near-inexistence of statistical significance testing. Addressing RQ3 yielded the original conceptual contribution of the three scarcities framework, demonstrating that low resources do not represent a single condition but rather a combination of three structurally distinct forms of scarcity—data, model, and infrastructure—each with its own causes, consequences, and potential solutions. Addressing RQ4 resulted in a practitioner's decision roadmap, guiding researchers towards the most viable methodological path given specific resource configurations, while surfacing the unresolved challenges that any such path must still contend with.

2. Background and related works

The biomedical NLP landscape has undergone a paradigm shift with the introduction of transformer-based PLMs, as documented in comprehensive surveys by Luo *et al.*[15] and Wang *et al.*[16] Both surveys converge on the finding that PLMs built on medical-specific corpora demonstrate a superior capacity to capture medical semantic knowledge compared to their general-domain counterparts, and that the pre-train-then-fine-tune paradigm has emerged as the default framework for biomedical NLP tasks. The most direct prior work to this literature review is the study by Shaitarova *et al.*[17], which surveyed clinical and biomedical NLP in languages other than English from 2020 to 2022. Their work reiterates the dominance of transformer-based models, the persistent gap between European and non-European languages, and, for the first time, accentuates the acceleration in Spanish-language NLP driven by targeted funding initiatives. However, their survey predates the emergence of generative LLMs as practical tools for synthetic corpus creation, addresses multiple NLP tasks

without the depth that a NER-focused review can achieve, and does not theorize the structural heterogeneity of low-resource conditions across languages.

The limitations outlined above define the scope of this review. Firstly, this study included publications up to 2025, covering an additional three years of advancements in generative LLMs that have introduced NLP techniques not present at the time of Shaitarova *et al.*[17]'s work. Secondly, it was based on a systematic and fully reproducible protocol, PRISMA, with explicit exclusion criteria applied consistently across 1,807 retrieved records. Thirdly, it deliberately confined the scope to NER and adjacent NLP tasks in the medical domain to achieve depth of technical and methodological findings that a broad multi-task survey cannot reach. Furthermore, the survey considered only pure clinical documentation, recognizing that it carries a distinctive set of linguistic challenges regardless of setting. Such prominent challenges include resolving abbreviations[18], correctly interpreting negation[19], and coping with flexible formatting (e.g., missing punctuation, parenthetical insertions, and variable documentation styles are commonplace).[20] Another difficulty concerns the presence of noisy and misspelled text[21], including spelling variations of transliterated words that have no immediate equivalent in the original language.[22] The list of challenges includes resolving semantic ambiguity among contextual synonyms[2,23], correctly identifying the boundaries of multi-line entities[5], and handling data sensitivity.[24] These challenges compound the resource scarcity problem in ways that no prior work has systematically mapped, motivating the three scarcities framework introduced in Section 5.1.

The authors of this review have a direct interest in the problem of medical NER in a low-resource setting. Our work on clinical NER for the Albanian language, a language for which no domain-specific language model and no annotated benchmark existed at the time of writing, has made us acutely aware of the practical choices facing researchers working at the genuinely resource-starved end of the spectrum. This review is a direct extension of that line of work, aiming to provide a structured, evidence-based account of the methodological landscape helpful to every researcher in the field in their early stages.

3. Materials and methods

This study was based on a systematic literature review conducted in accordance with the updated 2020 PRISMA guidelines. The adoption of PRISMA was motivated by the need to ensure transparency, robustness, and reproducibility of the review process. By following this framework, the study provides a clear and systematic rationale for conducting the review, as well as how the studies were identified, selected, and synthesized. Furthermore, the use of PRISMA ensures completeness of reporting while reducing the risk of bias and omissions, enabling other researchers to replicate the methodology or build upon the review by broadening the scope or expanding the temporal coverage.[25]

3.1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses Stage 1: Identification

The research query was designed to identify studies located at the intersection of four conceptual domains: NER, medical setting, LLMs, and low-resource languages. To achieve this, the query was structured as a four-clause exclusive Boolean expression, as presented in Figure 1. Each clause incorporates a range of contextual synonyms and commonly used abbreviations, thereby ensuring comprehensive search coverage and maximizing the retrieval of relevant studies. Aligned with the research scope, the first clause targeted the NLP task of interest, NER; the second clause restricted the domain to medicine, healthcare, and clinical settings; the third clause focused on transformer-based and LLM architectures; and the fourth clause constrained results to languages with scarce training materials, including, but not limited to, cross-lingual approaches that are typically employed to overcome data scarcity.

The query was submitted to seven acclaimed digital libraries to ensure broad and complementary coverage of the research in the fields of computer science, biomedical informatics, and computational linguistics: The Association for Computational Linguistics (ACL) Anthology, The Association for Computing Machinery Digital Library, arXiv, Elsevier ScienceDirect, IEEE Xplore, PubMed, and SpringerLink. The generic query shown in Figure 1 was adapted for two of the sources. Firstly, since Elsevier ScienceDirect's advanced search engine enforces a maximum of eight logical operators, the query was condensed to the core concepts, limiting the list of synonyms. Secondly, for arXiv, the low-resource language clause was omitted entirely because preliminary tests yielded near-empty result sets. However, the language criterion was meticulously applied during the downstream screening phase.

The query was executed on January 31, 2026, with a publication date filter spanning 2015 to 2025, and the results were ordered by relevance where available. The Association for Computational Linguistics search interface restricted access to the top 100 results of the search query. For all other sources, all returned results were retrieved and screened. In total, 1,800 results were identified across

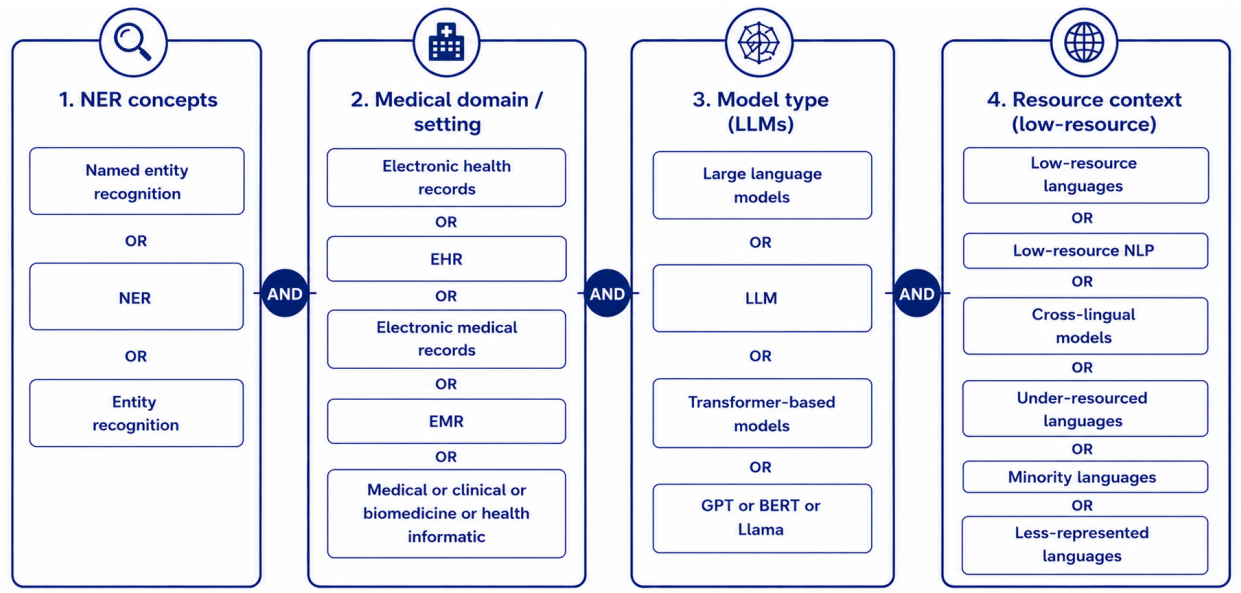

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses Stage 1: Identification using keywords
Abbreviations: EHR: Electronic health record; EMR: Electronic medical record; LLM: Large language model; NER: Named entity recognition; NLP: Natural language processing.

the seven digital sources. An additional seven publications were identified through backward citation chaining during the full-text review phase, yielding a combined result set of 1,807 records.

3.2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses Stage 2: Screening and Exclusion Criteria

Initial screening was conducted by reviewing titles and abstracts to assess their relevance to the research scope. However, more often than not, it was necessary to consult the full text rather than rely solely on the abstract, as determining the languages of the corpora, models, and techniques used in the studies was not always straightforward. Usually, this information was only provided in the methodology or dataset description sections. The exclusion criteria (EC) applied during screening and depicted in Figure 2 were:

(i) EC1: Non-medical domain. Publications proposing NLP solutions not specific to medicine or healthcare typically present models trained on general-purpose corpora or on corpora from unrelated domains such as finance, legislation, or the social sciences. This criterion also covers studies that, while set in a medical context, rely on text that lacks the terminology and structure characteristic of clinical documentation, e.g., electronic health records, radiology reports, or visit notes. Patient-generated content and social media text used for diagnostic purposes, e.g., detecting neurological disorders from online posts, fall under this exclusion.

(ii) EC2: Non-NER task. Publications that concerned non-NER tasks, e.g., machine translation, question answering, text generation, speech synthesis, were excluded from consideration. This criterion was extended to encompass only solutions based on LLMs or transformer-based models.

(iii) EC3: No low-resource language included. The scope of this review was publications aiming solutions for low-resource languages. Acknowledging that there is no universal agreement on measures for classifying languages as computationally high- or low-resource, this study excluded English and Chinese only. Indeed, the extent of research and the availability of unannotated and annotated corpora for these languages are unparalleled among the other languages considered here.

Additional criteria were employed to ensure that the literature review was based on original experimental results (EC4: Irrelevant publication type) and accessible to the authors through institutional or open-access channels (EC5: Inaccessible full text). Therefore, publications such as editorials, reviews, surveys, tutorial notes, extended abstracts, and conference proceeding notes were excluded

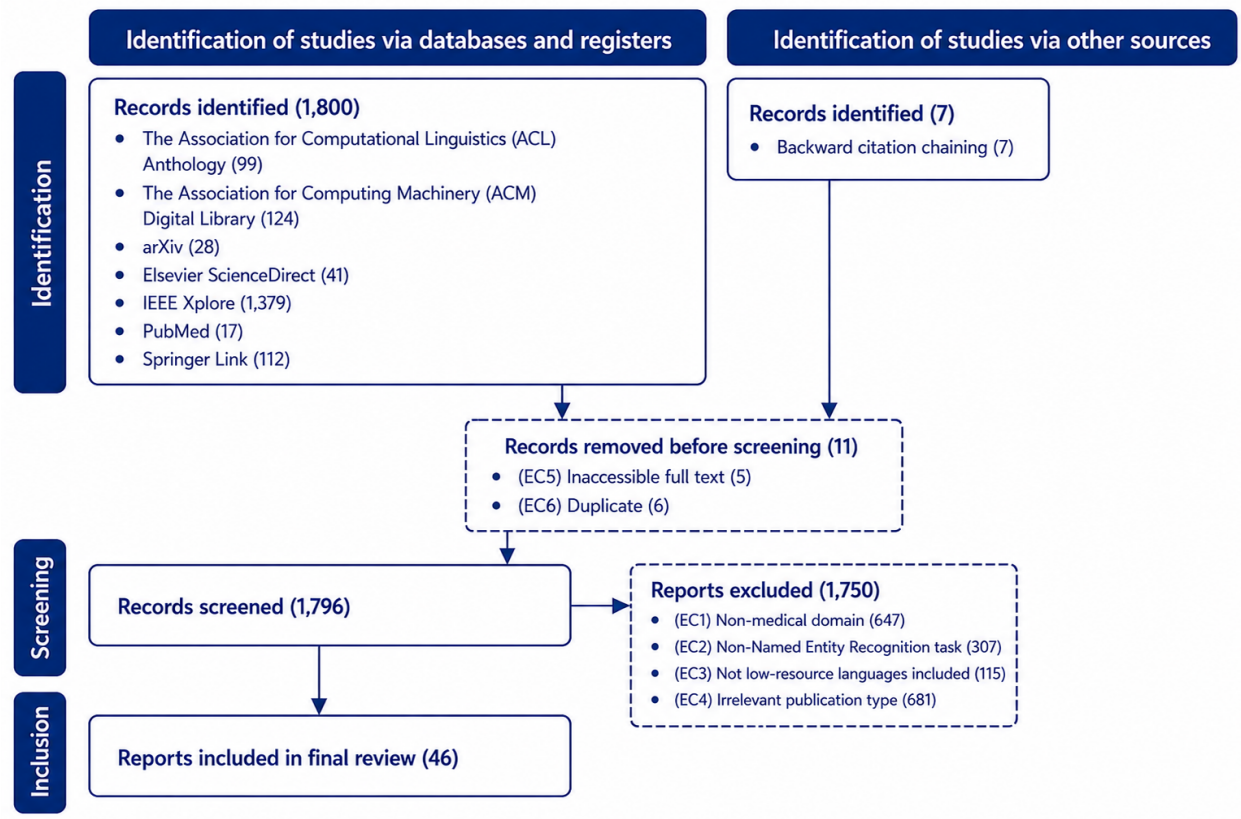

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses Stage 2: Screening and exclusion criteria

from consideration. Duplicates (EC6: Duplicate) were identified through manual comparison of the retrieved records in a reference spreadsheet, since the same publication was sometimes retrieved from more than one of the seven digital libraries with minor variations in formatting. Titles were normalized (case-insensitive, with punctuation and whitespace differences disregarded) before comparison, and records with matching normalized titles were cross-checked against the corresponding author list to confirm they referred to the same publication, after which they were removed as duplicates.

Based on the exclusion criteria, at the end of the PRISMA protocol screening phase 1,761 publications were excluded, and 46 were considered for review. The extended list of publications included is available in Table A1.

3.3. Preferred Reporting Items for Systematic Reviews and Meta-Analyses Stage 3: Inclusion for qualitative and quantitative analysis

The last stage of the PRISMA protocol required a close reading of the full texts of the selected publications. To structure this analysis and ensure consistency across studies, we established a framework with six thematic categories: publication identification, language and corpus, NER task specifics, model and technique, evaluation and results, and additional settings, as shown in Figure 3. The publications were systematically tagged by both authors independently. Ambiguous cases were resolved through discussions.

The extracted data were synthesized using a narrative synthesis approach, as the heterogeneity in entity schemas, evaluation metrics, and reporting practices documented in Section 4.2 precluded a quantitative meta-analysis. Themes were derived inductively through iterative coding of the 46 studies within the six extraction categories, then organized to address the four RQs introduced in Section 1: technique-level themes inform RQ1 and RQ2 (Sections 4.1–4.2), while patterns relating these techniques to language resource characteristics were synthesized into the three scarcities framework and roadmap addressing RQ3 and RQ4 (Sections 5.1–5.2).

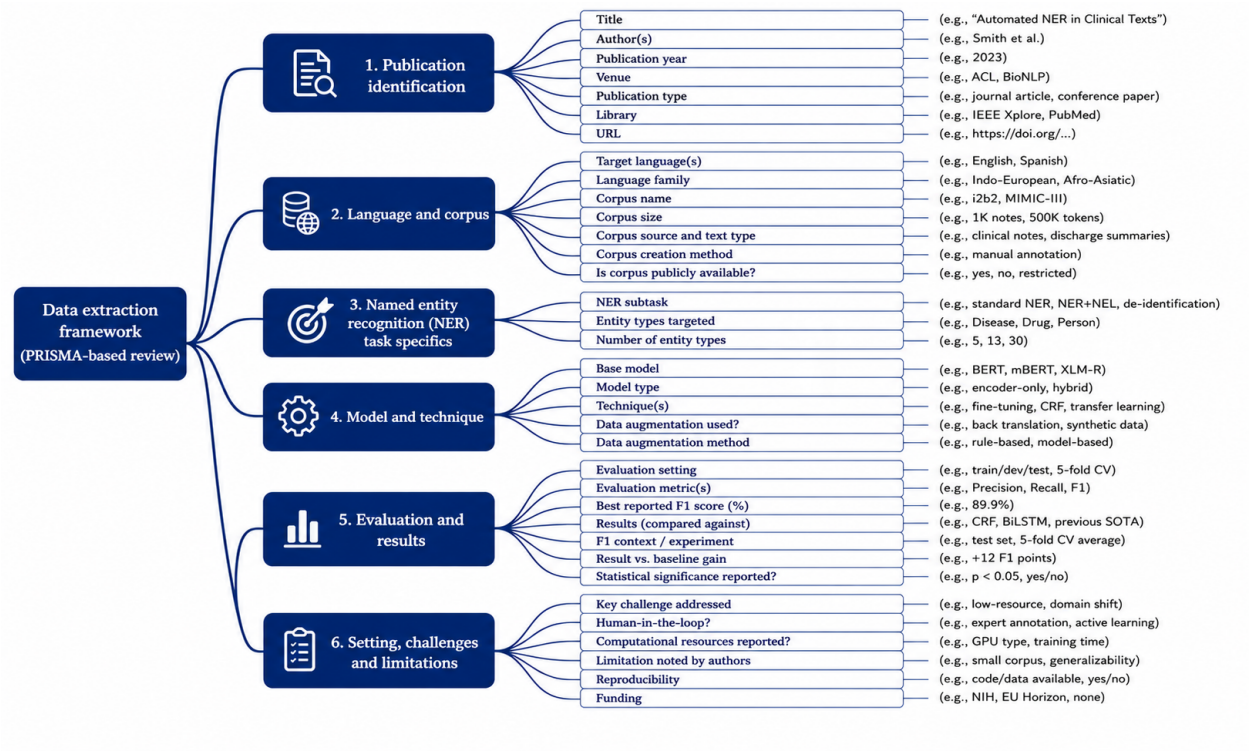

Figure 3. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Stage 3: Qualitative and quantitative analysis dimensions

3.4. Risk of bias assessment

To assess the methodological quality of the 46 included studies, we applied a five-domain checklist covering: (i) public availability of the corpus, (ii) reproducibility (code/data availability), (iii) presence of a reported baseline comparison, (iv) statistical significance testing, and (v) human-in-the-loop validation of annotations. These domains were selected as suitable indicators of methodological transparency and rigor for computational NLP method papers, a category not well served by existing clinical risk-of-bias tools such as Joanna Briggs Institute or Critical Appraisal Skills Programme, which are designed for trial and observational study designs. Each domain was scored 1 (fully met), 0.5 (partially met), or 0 (not met) using the data already extracted during PRISMA Stage 3, yielding a composite score out of 5 per study. Scoring followed the same independent dual-author process described in Section 3.3, with disagreements resolved through discussion. No study was excluded on the basis of its score; the assessment was applied descriptively, to characterize the evidence base rather than to filter it, since methodological weakness is itself part of the phenomenon this review documents.

4. Results

Publication output was roughly stable over the review period, ranging from 6 to 9 papers per year between 2020 and 2025, with no clear upward or downward trend. Target languages skew heavily toward European languages: Spanish[3,26–37], English[3,26,28,36–43], German[44-50], and French[36,43,45,51-53] together appear in at least one language in 29 of the 46 papers. The other 10 papers are European languages other than Spanish, English, French, and German, while genuinely non-European languages appear in only 7[54–59], reinforcing the resource concentration discussed in Section 5.1. Task focus was overwhelmingly on NER itself, with relation extraction and entity normalization each appearing in only a handful of studies. Of the 46 reviewed papers, 37 used a BERT-family encoder as their primary architecture, with generative LLMs appearing in only a handful of studies and typically not as NER engines but as data generators feeding a downstream BERT encoder. Consequently, the dominant practical framework remained structurally stable across the review period: selecting a PLM, optionally continuing pre-training on unlabeled in-domain text, fine-tuning on a labeled NER corpus, and evaluating on a held-out test set.

This section addresses two RQs against that backdrop. RQ1 asks which model-level and data-level techniques have been applied in this setting and, under which resource configurations, each approach demonstrates effectiveness; this is examined in Section 4.1. RQ2 asks to what extent performance results in this literature are comparable and what structural factors undermine cumulative scientific progress; this is addressed in Section 4.2.

4.1. Techniques for low-resource medical named entity recognition

4.1.1. Model-level techniques

Fine-tuning on in-domain labeled corpora is the foundational technique for NER, serving as the baseline against which all other models are measured. The technique starts with initializing a PLM encoder with its learned weights and continuing training on a labeled NER corpus. The model's performance is primarily determined by the specificity of the pre-trained data and the size and annotation quality of the labeled fine-tuning corpus. When both are well-developed, the results are promising: BER-Timbau, fine-tuned on a purpose-annotated oncology corpus for European Portuguese, achieved an F1 of 95.0% for drug entities[60], and mBERT, fine-tuned on Spanish clinical data, achieved 88.8% on PharmaCoNER.[26] When the labeled corpus is small, fine-tuning tends to overfit, and other NLP strategies become necessary.

A technique that accounts for the mismatch between the general-domain text on which most encoder models are pre-trained and the specialist vocabulary and syntactic structure of clinical language is continued pre-training, also referred to as domain-adaptive pre-training. This technique preserves the general linguistic representations learned during initial pre-training while adjusting them toward domain-specific patterns. BioBERTpt, the Brazilian Portuguese biomedical BERT, was initialized from mBERT and continued pre-training on de-identified clinical notes from Brazilian hospitals and Portuguese biomedical abstracts, achieving an F1 of 92.6%.[61] For Vietnamese, ViHealthBERT was created by continued pre-training of PhoBERT on health-domain texts, outperforming general-domain baselines across all evaluated health datasets.[54] The German benchmarking study[44] demonstrated that domain-adaptive pre-training on translated PubMed and MIMIC-III data yielded a consistent 6.9% point improvement on the GraSCCo clinical corpus compared to the general GBERTlarge baseline. However, these models exhibit large-scale dependence: as Šuvalov *et al.*[62] explicitly noted, the gains from continued pre-training are marginal when the in-domain unlabeled corpus is small.

Cross-lingual transfer leverages the multilingual representations learned by mBERT and XLM-RoBERTa to perform NER in a target language without labeled data for that language. The mechanism relies on the empirically demonstrated property that multilingual encoders trained on Wikipedia-scale corpora spanning more than 100 languages learn shared representations in which semantically equivalent tokens across languages occupy nearby positions in the embedding space. This property emerges from the scale and diversity of the pre-training corpus without requiring parallel data. In practice, the model is fine-tuned on a source-language NER dataset, typically English, and evaluated zero-shot on a target-language test set. On one hand, the technique is particularly efficient for linguistically proximate language pairs: mBERT fine-tuned on MEDDOCAN Spanish and evaluated zero-shot on Spanish–Catalan code-mixed clinical notes achieved an F1 of 73.7%.[27] XLM-RoBERTa base used as a CLT model for French and German clinical NER produces results comparable to translation-based approaches and substantially outperforms the smaller DistilBERT CLT baseline on the French test set.[45] However, CLT degrades in settings with linguistic distance and domain divergence. Mullick *et al.*[38], evaluating CLT for six Indic languages, found that zero-shot CLT to non-English targets degrades entity F1 by approximately 9–10 % points relative to the English baseline on the 1mg dataset. The results were attributed in part to typological distance and in part to the underrepresentation of Indic languages in the pre-training corpora.

Few-shot CLT occupies a practical middle ground between zero-shot CLT and fully supervised fine-tuning. A small number of labeled examples in the target language, typically 50 to 500, are used to adapt the CLT-initialized model to the target distribution. Amin *et al.*[27], addressing de-identification of Spanish–Catalan code-mixed stroke clinical notes, provided the most compelling illustration: fine-tuning the zero-shot mBERT baseline with a few hundred labeled target-domain examples raises the F1 from 73.7% to 91.2%. This result reveals a systematic principle: even small amounts of in-domain labeled data provide a disproportionate boost to CLT-based models. The improvement occurs because few-shot adaptation targets precisely the lexical and distributional shifts that zero-shot CLT fails to accommodate.

4.1.2. Data augmentation techniques

Almost one-third of the reviewed papers address scarcity via data augmentation. The methods were divided into five categories that differ in their technical assumptions, dependency on external resources, and applicability across the low-resource spectrum.

Back-translation produces training data by exploiting the variability of bidirectional machine translation. In back-translation, a clinical sentence in the target language is translated to a high-resource language and then translated back, generating a paraphrase that preserves semantic content while introducing lexical and syntactic differences, as demonstrated by Lancheros *et al.*[28], Vakili *et al.*[29], and Yada *et al.*[39] As an example, Yada *et al.*[39] addressed Japanese-English bilingual NER by applying back-translation via Chinese as a bridge language in combination with synonym replacement, segment shuffling, and label-wise token replacement, achieving an F1 of 89.26%. The effectiveness of back-translation is tightly coupled to translation quality. Mullick *et al.*[38] provided a direct illustration of this principle: in back-translation experiments on Indic languages, the information loss during back-translation was substantial for clinically specialized entities, with performance dropping by 9.66 % points for RoBERTa and 10.49 % points for BioClinicalBERT compared to the English baseline on the 1mg dataset.

Entity-level replacement produces synthetic data by substituting labeled entities in an annotated corpus with other entities of the same semantic type, retrieved from a domain terminology dictionary. Schneider *et al.*[61] deployed this technique by compiling entity lists from official ontologies and cross-checking them against the existing dataset vocabulary to ensure genuine novelty. Lancheros *et al.*[28] combined entity replacement with back-translation for Spanish biomedical NER, substituting 20% of entity spans with same-type dictionary entities, achieving an F1 of 86.39%. The principal advantage of entity-level replacement over back-translation is its independence from translation quality. In entity-level replacement, the substitution directly manipulates clinical entities, leaving the surrounding syntactic structure intact and ensuring that the resulting instances are grammatically well-formed. However, the technique requires a medical terminology dictionary, and coverage varies by language: Systematized Nomenclature of Medicine Clinical Terms, for instance, has official editions in English, Spanish, Danish, and Swedish, with French and Dutch translations still in progress, while Unified Medical Language System's non-English coverage is similarly concentrated in a small set of European languages. This coverage pattern closely overlaps with the institutionally well-resourced languages discussed in Section 5.1, so entity-level replacement remains dependent on a terminology infrastructure that, like annotated corpus availability, is largely absent in the genuinely low-resource languages where augmentation is most needed.

Semantic similarity-based augmentation is a more linguistically sensitive variant of entity-level replacement as it does not require a pre-existing terminology resource. Phan and Nguyen[40] propose replacing entity mentions and surrounding context tokens with semantically similar alternatives identified through cosine similarity in a pre-trained embedding space. On the VietBioNER corpus, this technique yielded a 1.3% point F1 improvement over the non-augmented baseline. The larger gains were observed in the low-resource regime of 50 to 500 training sentences, precisely the regime most relevant for genuinely low-resource scenarios. The language-agnostic nature of cosine similarity makes the technique applicable, provided a pre-trained embedding model for the target language exists.

Large language model-based synthetic corpus generation is the newest and most rapidly evolving augmentation technique in this review. The augmentation pipeline involves three stages: a generative LLM is prompted to produce clinical text containing entities of specified types; the same or a different model assigns BIO-format annotations; and the resulting labeled corpus is employed to fine-tune a BERT-family encoder. Šuvalov *et al.*[63] provided a methodologically detailed account for Estonian. Firstly, a GPT-2 model domain-adapted to Estonian electronic health record text via QLoRA quantization generated synthetic clinical documents that exceeded the gold corpus by a factor of 4. Secondly, GPT-3.5-Turbo and GPT-4 were used as annotation models across four prompt strategies, such as zero-shot, Estonian-language zero-shot, definitions-augmented, and few-shot, with few-shot prompting achieving the highest annotation quality and GPT-4 substantially outperforming GPT-3.5-Turbo across all entity types. Thirdly, the downstream NER model, trained entirely on synthetic data and annotations, achieved an F1 score of 69% for drug extraction. The result is remarkable considering that it was achieved without any human annotation effort beyond the expert sample used for quality evaluation. The authors nonetheless caution that hallucination in LLM-generated text can introduce noise into the resulting corpus, a risk inherent to any pipeline that trains a downstream model on machine-generated rather than human-written text. Frei and Kramer[46] reported a comparable LLM-based annotation pipeline for German and were among the few to explicitly quantify its computational cost, estimating 118 graphics processing unit-hours and approximately 15 kg of carbon dioxide emissions. They also highlighted the ethical concerns surrounding fully automated annotation as an important limitation of the approach. Byun *et al.*[55] applied a simpler variant for Korean, with ChatGPT-generated synthetic sentences producing a 20% improvement in medical NER over the general-domain baseline. The authors note, however, that ChatGPT-generated sentences may not fully reflect the authentic clinical writing style, making the quality of downstream annotation dependent on the

model's command of Korean medical language. Vakili *et al.*[29] provided another strong example of fully automated synthetic pipelines: a Swedish de-identification model trained on LLM-generated and machine-annotated data that falls only 2.9% F1 points short of the gold-data-trained benchmark. The most sophisticated architecture variant in this review is presented by Wang *et al.*[56] for Mongolian and classical Chinese. The technique employs advanced prompt engineering: a collaborative large-small model pipeline uses chain-of-thought prompting to decompose the generation task into multi-step reasoning constraints, reducing hallucinations and yielding F1 improvements of 3.52% to 9.42% over baseline data augmentation methods. The authors also note that LLMs' limited exposure to ultra-low-resource languages during pre-training constrains the quality of augmentation, and that the framework still depends on the availability of high-resource language descriptions for CLT. Taken together, these caveats—hallucination, dependence on access to proprietary models, and non-trivial compute and emissions costs—qualify the F1 gains reported above and echo the broader comparison in Section 4.2, where GPT-3.5 few-shot prompting without fine-tuning represents the current performance floor relative to fine-tuned encoder baselines.

Cross-lingual annotation projection creates target-language labeled corpora by transferring NER labels from annotated source-language text via machine translation and word alignment, a technique that sits at the boundary between corpus creation and augmentation.[64] Schäfer *et al.*[47] implemented this for German clinical NER using Fairseq WMT19 Neural Machine Translation and either the statistical FastAlign or the BERT-based AWESOME aligner. The BERT-based aligner substantially outperformed FastAlign for the German clinical domain, achieving F1 improvements of 5.4% points for medication, 8.7% for diagnosis, and 5.9% for treatment entities over the Conditional Random Field baseline, with an overall impressive F1 of 95.4%. Zaghir *et al.*[51] applied a closely related approach for French, producing the FRASIMED corpus of 2,051 synthetic French clinical cases. An important consideration is that projection quality degrades when the source and target languages exhibit substantial syntactic and morphological divergence.[41] Moreover, annotation projection is substantially more reliable for entities with high cross-lingual surface overlap, such as medications, than for diagnosis and treatment entities, whose annotation guidelines differ across languages and clinical traditions.[47] On the question of translate-train versus translate-test, Fontaine *et al.*[45] provided the most directly relevant comparative evidence: CLT with general-domain multilingual models yields results comparable to translating the training set, but training on translated data consistently outperforms translating at inference time. The translate-test approach should be reserved for settings where large multilingual models cannot be used.

4.1.3. Conditions on model selection: A comparative reading

The evidence from Sections 4.1.1 and 4.1.2 converges on several cross-cutting findings that directly condition model selection in low-resource settings. One finding that warrants particular emphasis, because it is counterintuitive and has direct practical implications, is that domain-specific monolingual pre-training is not universally superior to general-purpose multilingual pre-training. Fontaine *et al.*[45] demonstrated that XLM-RoBERTa Base, a general-domain multilingual model, achieved results comparable to translation-based approaches using monolingual clinical language models for French and German NER. According to the study, domain-specific monolingual models outperformed multilingual alternatives only when pre-trained on sufficiently large clinical corpora. The Swedish benchmarking study[65] reached a similar conclusion: larger generic models outperformed smaller domain-adapted clinical models on NER tasks, suggesting that model capacity is a stronger determinant of NER performance than domain specificity when training data is limited. This challenges the default assumption that a domain-specific BERT, however small its pre-training corpus, will outperform a larger, general-domain multilingual model for clinical NER.[66]

4.2. Performance reporting and the comparability problem

Addressing RQ2, the evidence reveals that performance results across the reviewed papers were not directly comparable, and that several structural features of the field actively undermined cumulative scientific progress. Thirty-seven of the 46 reviewed papers reported at least one F1 score; the remaining nine either presented corpus-only work without a trained model[57], addressed tasks for which F1 is not the primary metric[30,31,67] or did not report numerical results.[42,68,69] Among the 37 papers that reported F1 scores, scores ranged from 37.25% to 97.2%, with a mean of 85.3% and a median of 87.7%. The lower end was occupied by Villena *et al.*[32], who reported a GPT-3.5 few-shot NER result of 37.25% on Spanish clinical referral documents. This result is worth noting, as it establishes the floor for what current generative LLMs achieve in clinical NER without fine-tuning. The upper end, 97.2%, belongs to Amin *et al.*[27], achieved through few-shot CLT for Spanish–Catalan code-mixed de-identification. The Mongolian and classical Chinese paper[56] reported an F1 score of 65.61%, the lowest among encoder-based models,

reflecting the genuinely sparse resource environment for these languages.

A blunt comparison of F1 results across papers could be misleading for several reasons. Firstly, entity schemas differ radically: a paper targeting three entity types and a paper targeting 20 are measuring fundamentally different things with the same metric. Secondly, annotation guidelines are not standardized outside the few examples of shared tasks. Schäfer *et al.*[47] explicitly demonstrated the divergences in what constitutes a diagnostic entity between the English n2c2 guidelines and the German clinical annotation tradition: divergences that directly influence the cross-lingual projection for that entity class. Thirdly, train–test splits are neither uniform in size nor always drawn from the same corpus distribution. Indeed, several papers[45,46,48] evaluated on corpora that differ from the training distribution, making their F1 scores measures of cross-domain generalization rather than in-domain accuracy. The Swedish benchmarking paper[65] was one of the few to explicitly compare domain-adapted and generic models across multiple evaluation datasets. It found that domain-adapted models outperformed generic models on classification tasks, whereas larger generic models outperformed smaller clinical models on NER, a contrast that would have remained undetected if either result had been reported in isolation.

The statistical significance situation was inadequate. Of the 46 papers, only one[58] reported a significance test: a *t*-test yielding $T = -16.69$ and $p = 1.68\times10^{-7}$. The remaining 45 papers reported performance differences without assessing whether these exceeded what could be expected by chance, given the variability inherent in small evaluation sets, random model initialization, and the multiple comparisons involved in hyperparameter optimization.

4.3. Risk of bias

Composite risk-of-bias scores ranged from 0 to 3 out of 5, with a mean of 1.55. Forty-two of the 46 studies (91%) fell into a low tier (score < 2.5), four (9%) into a medium tier (2.5–3.5), and none reached a high tier (≥4). The weakest domains were statistical significance testing, reported in only 1 of 46 studies, and reproducibility, with 13 studies sharing neither code nor data. Rather than undermining the review's conclusions, this distribution is itself a substantive finding: it provides direct empirical evidence for the infrastructure scarcity dimension of the framework introduced in Section 5.1, where the absence of shared evaluation standards and methodological norms is argued to be a structural, not incidental, feature of the field.

5. Discussion

5.1. The three scarcities: A structural interpretation

Addressing RQ3, the evidence reveals that the resource configurations characterizing the reviewed languages were not reducible to a single dimension, thereby giving rise to the three scarcities framework developed below. The findings converge on the interpretation that the condition described by the term low-resource in the context of medical NLP is not a single condition but a compound of three structurally distinct scarcities, each with different causes, consequences, and remedies. Recognizing their independence is essential to accurately diagnosing the state of a given language and to prioritizing research and policy interventions appropriately.

The first is data scarcity, which is the absence of annotated corpora for training and evaluation. This is the most visible and discussed scarcity in individual papers. The methods documented in Section 4.1, such as CLT, continued pre-training, back-translation, entity replacement, annotation projection, and LLM-based synthetic generation, are all responses to data scarcity. The field has made real, documented progress on this dimension over the review period, and the emergence of LLM-based synthetic generation pipelines represents a genuine methodological advance that substantially expands what is achievable in zero- or near-zero-annotation settings.

The second is model scarcity, which is the absence of PLMs for the target language, particularly domain-adapted ones. This scarcity is partially technical and partially economic: training a monolingual BERT from scratch requires a large unlabeled corpus and substantial computational investment, both of which are concentrated in high-resource language communities. Multilingual models partially address model scarcity by extending transferability to additional languages during pre-training, but coverage remains uneven. XLM-RoBERTa was trained on 100 languages, yet the token budget per language scales roughly with the size of the available web corpus, leaving genuinely low-resource languages with shallow representations. Continued pre-training on existing in-domain text is the most accessible partial remedy, but it is only a remedy when some unlabeled text and some computational resources are available.

The third is infrastructure scarcity, which is the absence of shared task benchmarks, evaluation standards, institutional funding mechanisms, and the scholarly ecosystem that generates cumulative progress. This is the scarcity least visible in individual papers, never addressed

by any individual methodology, and most consequential for long-term research trajectory. A language with data scarcity and model scarcity but with a functioning shared task and funding mechanism can, over time, build both. Spanish is the evidence: the PharmaCoNER corpus was released in 2019, and within five years, the language had multiple domain-adapted models, a shared-task benchmark, and more papers in this review than any other language. A language with infrastructure scarcity has no such path. This review finds no evidence that any of the genuinely low-resource languages outside the European institutional context, Uyghur, Bangla, or Mongolian, is receiving the level of sustained institutional investment required to exit infrastructure scarcity within any foreseeable time horizon. Advancing medical NLP for these languages requires not only better methods but better institutions. Technical innovation addressing data and model scarcity, which is the focus of the vast majority of papers in this corpus, is valuable and should continue, but it operates within a ceiling set by infrastructure scarcity. The most sophisticated LLM-based synthetic generation pipeline cannot substitute for the shared task that would create a common evaluation standard, the government funding program that would sustain a research team, or the data use agreement framework that would make clinical records legally accessible to researchers.

5.2. A practitioner's roadmap: From resource configuration to a working named entity recognition system

Addressing RQ4, the following decision roadmap translates the evidence assembled in Chapter 4 and the structural diagnosis of Section 5.1 into actionable guidance for practitioners, as depicted in Figure 4.

The first question to resolve is whether a PLM that includes the target language with adequate coverage already exists. If a language-specific domain-adapted encoder exists, as is the case for Spanish, French, Brazilian Portuguese, Vietnamese, Arabic, and Korean at varying levels of biomedical specificity, the primary task is fine-tuning, provided that the encoder was pre-trained on a sufficiently large in-domain corpus. When the domain-adapted model's pre-training corpus is small, Section 4.1.3 indicates that a larger general-purpose multilingual encoder, such as XLM-RoBERTa, can outperform it; both should be evaluated empirically rather than defaulting to the domain-specific option by name alone. The model should be fine-tuned on the available labeled NER corpus, following the hyperparameter configurations established in Schneider *et al.*[61] and Copara *et al.*[52] If no language-specific domain-adapted model exists but a general-domain monolingual encoder is available, continued pre-training on any available unlabeled in-domain medical text should be attempted before fine-tuning.[44,54,61,62] If the in-domain unlabeled corpus is small (fewer than a few million words), the gains from continued pre-training may be marginal relative to the compute investment, and proceeding directly to fine-tuning with data augmentation may be more efficient.

When no monolingual model exists, the choice between multilingual CLT and corpus-creation strategies depends on the annotated-corpus situation. XLM-RoBERTa should be preferred over mBERT as the CLT base model, as multiple papers consistently favor its larger pre-training corpus and more robust multilingual representations, particularly for languages underrepresented in mBERT's training data.[29,38,45] Zero-shot CLT can serve as a lower bound on what is achievable without any target-language annotation and is a legitimate starting point when no labeled data exists. Few-shot CLT, providing as few as 50 to 500 labeled target-language examples, almost always produces meaningful improvements over the zero-shot baseline[27,55], and should be the default path when even minimal annotation capacity is available.

If an annotated corpus exists for the target language and clinical domain, such as a shared task corpus, it should be used as the primary fine-tuning resource, even if its domain does not fully overlap with the deployment context. Schäfer *et al.*[47] demonstrated that fine-tuning on radiological reports meaningfully transfers to magnetic resonance imaging and pediatric X-ray domains in low-data settings, and that a PLM fine-tuned on a foreign clinical domain outperformed a model trained from scratch on the target domain in resource-scarce settings.

If no annotated corpus exists in the target language, the practical path depends on the translation infrastructure. When a reasonably capable machine translation system exists, an assumption that holds for most European languages and is increasingly valid for major Asian languages, cross-lingual annotation projection from English, or other higher-resource language, should be considered. The pipeline documented in Schäfer *et al.*[47] and Zaghir *et al.*[51], neural machine translation of an English clinical NER corpus followed by BERT-based cross-lingual span alignment for label transfer, can produce silver-standard annotated corpora of sufficient quality to substantially outperform zero-shot CLT baselines. The choice of alignment model is critical: BERT-based contextualized alignment consistently outperforms statistical alignment in clinical domains where one-to-many token mappings are common. The resulting corpus should be treated as a starting point, and it is documented that any available capacity for human expert correction of a

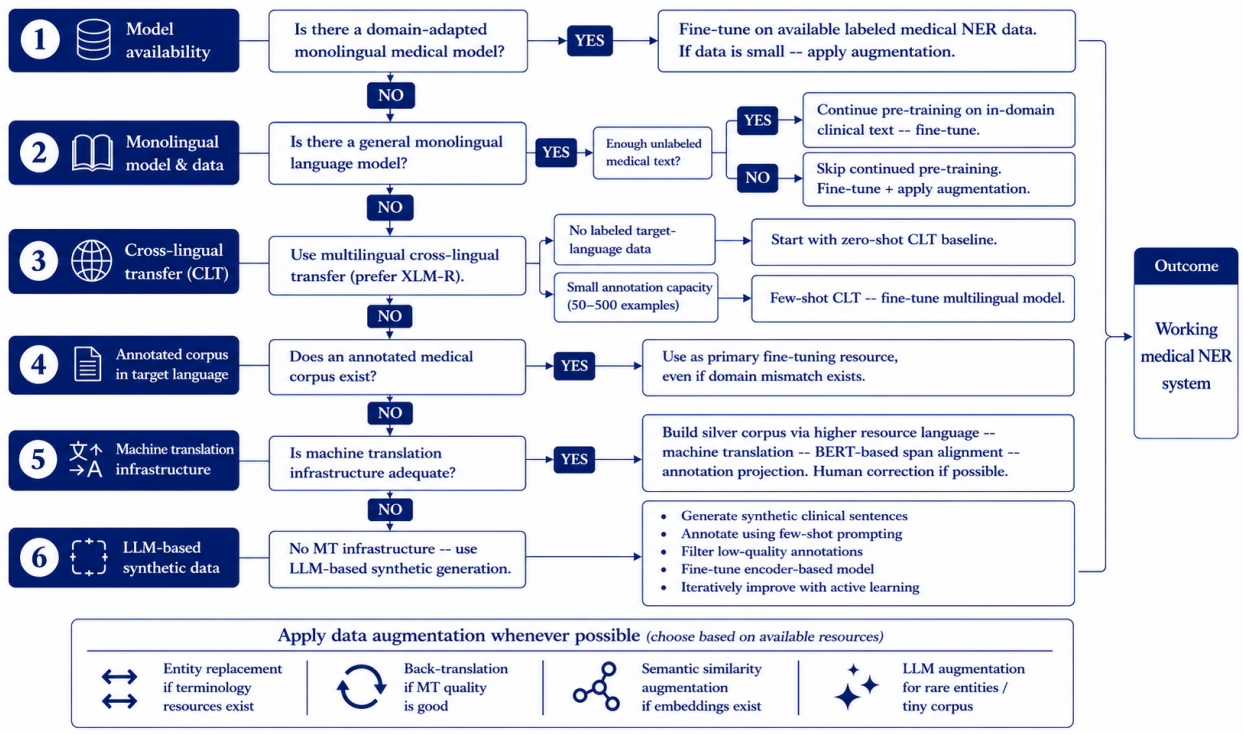

Figure 4. A practitioner's roadmap
Abbreviations: LLM: Large language model; MT: Machine translation; NER: Named entity recognition.

sample of projected annotations will improve downstream model quality.

When the translation infrastructure is underdeveloped, LLM-based synthetic corpus generation is the most viable alternative for bootstrapping from zero. The practical template established across papers[38,48,56,65] is to generate synthetic clinical sentences using a locally deployed or privacy-compliant LLM prompted with detailed clinical entity specifications; use few-shot prompting with example annotation instances rather than zero-shot to maximize annotation quality; apply the latest and greatest commercial rather than smaller models where budget permits; filter low-confidence annotations using model confidence scores or inter-model agreement; and fine-tune a BERT-family encoder on the resulting corpus. The expectation is not gold-standard performance but a meaningful baseline that can be iteratively improved through subsequent active-learning annotation cycles.

Augmentation should be applied wherever a working baseline exists, and resources permit. The hierarchy of practical applicability, as established by the evidence, is as follows. Entity-level replacement is the most reliable method when a domain terminology resource exists in the target language; it requires no machine translation system and preserves grammaticality by construction. Back-translation is highly effective when machine-translation quality is adequate but degrades significantly in morphologically complex languages and in entity types where translation produces systematic surface divergence. Semantic similarity-based augmentation is a language-agnostic alternative applicable wherever a pre-trained embedding model is available.[40] LLM-based augmentation can further expand training diversity, particularly for rare entity types and very small baseline corpora, provided quality control through few-shot prompting and human sample review is applied.[56,63] No augmentation technique is universally superior; each method's effectiveness is conditional on the specific resource configuration of the target language.

6. Conclusion

This review aimed to address four RQs. RQ1 is addressed by the taxonomy of model- and data-level techniques presented in Section 4.1, which establishes that effectiveness is conditional on resource configuration rather than on technique alone. RQ2 is addressed in Section 4.2, finding that F1 scores are not comparable across papers and

that the near-absence of statistical significance testing represents a structural weakness of the field. RQ3 is answered by the three scarcities framework in Section 5.1, demonstrating that low-resource is a compound of three distinct conditions requiring different remedies. RQ4 is addressed by the practitioner's decision roadmap in Section 5.2, translating the accumulated evidence into actionable guidance.

Medical NER for low-resource languages is a problem grounded in technical and institutional roots. On the technical side, the field has made genuine and measurable progress. BERT-family encoder models, particularly mBERT and XLM-RoBERTa, provide a viable starting point for virtually any language, and CLT from higher-resource languages can serve as a lower bound even when no labeled data exists in the target language. Few-shot adaptation with as few as 50 to 500 in-domain examples consistently delivers substantial improvements over the zero-shot baseline. LLM-based synthetic corpus generation, demonstrated most clearly for Estonian, Korean, and Mongolian, has emerged as a practical path to bootstrapping NER from zero, a development that simply did not exist at the start of the review period. The growing body of evidence for augmentation via back-translation, entity replacement, and semantic similarity substitution further extends the toolkit available to researchers working under data constraints. These advances are real. They reduce the annotation burden, lower the performance floor for genuinely low-resource languages, and make clinical NER more accessible to communities that previously had no viable path to it.

On the institutional side, however, it is less encouraging. The low-resource label, as this review has documented, spans a spectrum from Spanish, which has shared task benchmarks, multiple domain-adapted BERT models, and sustained government funding through Spain's Plan TL, to Uyghur, Mongolian, and Bangla, where a single paper may represent the entirety of published clinical NLP work. The three scarcities framework introduced in Section 5.1 formalizes this distinction. Data scarcity and model scarcity are addressable by the methods reviewed here. Infrastructure scarcity, as the absence of shared task benchmarks, evaluation standards, and research funding mechanisms, is not. No data augmentation technique can substitute for the institutional investment that produced PharmaCoNER, enabled IberLEF, and funded the WisPerMed doctoral program.

Several limitations of the reviewed literature warrant emphasis, as they also limit the conclusions that can be drawn here. Only one of the 46 reviewed papers reported statistical significance testing on its results. F1 scores were not comparable across papers due to heterogeneous entity schemas, annotation guidelines, and evaluation corpora. Reproducibility was partial at best: 13 papers shared neither code nor data, and 5 did not report the availability of their data. These were not isolated oversights; they reflect field norms that the community should consciously revise. Future reviews of this topic will be considerably stronger if significance testing, entity-level F1 reporting, and public release of both code and data become standard practice.

Looking ahead, three challenges follow directly from the evidence assembled here. Evaluation standardization is the most pressing: the incommensurable F1 scores documented in Section 4.2 mean that the field cannot build cumulatively until shared entity schemas and benchmark corpora, comparable to those for English, become available. Cross-lingual generalization beyond the largely Indo-European evidence base reviewed here remains open: the transfer degradation reported for Indic languages in Section 4.1.1 suggests typologically distant language pairs need dedicated study rather than extrapolation from these results. Privacy-preserving learning is an increasingly relevant direction as LLM-based synthetic generation matures: locally deployable, open-source models and on-premises fine-tuning warrant direct investigation as alternatives to proprietary cloud application programming interfaces for languages in which clinical data cannot leave institutional or national boundaries. Cross-lingual annotation projection methods, especially BERT-based span alignment, remain underexplored for non-Indo-European language pairs, and institutional investment in the development of Spanish-language clinical NLP at the national and European Union levels should be replicated to genuinely support underserved language communities. The methods required to enable clinical NER in low-resource languages are largely already available. The remaining challenge lies in ensuring their equitable development, adaptation, and deployment across diverse linguistic and healthcare settings.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: All authors
Data curation: Polina Çeço

Formal analysis: All authors
Investigation: All authors
Methodology: Elira Hoxha
Resources: All authors
Validation: Elira Hoxha
Visualization: Polina Çeço
Writing–original draft: All authors
Writing–review & editing: Elira Hoxha

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data are available from the corresponding author upon reasonable request. This includes the full list of papers retrieved from online sources as defined in PRISMA Stage 1 (Identification of keywords), the analysis of records against the exclusion criteria applied in PRISMA Stage 2, and the complete tagging of publications according to the PRISMA Stage 3 (Qualitative and quantitative analysis dimensions).

References

1. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform*. 2009;42(5):760-772.

   doi: 10.1016/j.jbi.2009.08.007

2. Holzinger A, Schantl J, Schroettner M, Seifert C, Verspoor K. Biomedical text mining: state-of-the-art, open problems and future challenges. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg; 2014:271-300.

   doi: 10.1007/978-3-662-43968-5_16

3. Rivera-Zavala R, Martinez P. The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: comparative study. *JMIR Med Inform*. 2020;8(12):e18953.

   doi: 10.2196/18953

4. Zhou B, Yang G, Shi Z, Ma S. Natural language processing for smart healthcare. *IEEE Rev Biomed Eng*. 2024;17:4-18.

   doi: 10.1109/RBME.2022.3210270

5. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: Challenges and opportunities. In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE; 2016:1937-1945.

   doi: 10.1109/bigdata.2016.7840814

6. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics; 2018:2145-2158. Accessed January 31, 2026. https://aclanthology.org/C18-1182/

7. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.

   doi: 10.1093/bioinformatics/btz682

8. Alsentzer E, Murphy JR, Boag W, *et al*. Publicly Available Clinical BERT Embeddings. *arXiv*. Preprint posted online 2019.

   doi: 10.48550/arXiv.1904.03323

9. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv*. Preprint posted online 2019.

   doi: 10.48550/arXiv.1903.10676

10. Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2022;3(1):1-23.

    doi: 10.1145/3458754

11. Peng Y, Yan S, Lu Z. 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*. Association for Computational Linguistics; 2019:58-65.

    doi: 10.18653/v1/W19-5006

12. Durango MC, Torres-Silva EA, Orozco-Duque A. Named entity recognition in electronic health records: a methodological review. *Healthc Inform Res*. 2023;29(4):286-300.

    doi: 10.4258/hir.2023.29.4.286

13. Fraile Navarro D, Ijaz K, Rezazadegan D, *et al*. Clinical named entity recognition and relation extraction using natural language processing of medical free text: a systematic review. *Int J Med Inform*. 2023;177:105122.

    doi: 10.1016/j.ijmedinf.2023.105122

14. Jehangir B, Radhakrishnan S, Agarwal R. A survey on named entity recognition—datasets, tools, and methodologies. *Nat Lang Process J*. 2023;3:100017.

    doi: 10.1016/j.nlp.2023.100017

15. Luo X, Deng Z, Yang B, Luo M. Y. Pre-trained language models in medicine: a survey. *Artif Intell Med*. 2024;154:102904.

    doi: 10.1016/j.artmed.2024.102904

16. Wang X, Sanders HM, Liu Y, *et al*. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac*. 2023;41:100905.

    doi: 10.1016/j.lanwpc.2023.100905

17. Shaitarova A, Zaghir J, Lavelli A, Krauthammer M, Rinaldi F. Exploring the latest highlights in medical natural language processing across multiple languages: a survey. *Yearb Med Inform*. 2023;32(01):230-243.

    doi: 10.1055/s-0043-1768726

18. Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. In: *Biocomputing 2003*. World Scientific; 2002:451-462.

    doi: 10.1142/9789812776303_0042

19. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34(5):301-310.

    doi: 10.1006/jbin.2001.1029

20. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform*. 2015;57:28-37.

    doi: 10.1016/j.jbi.2015.07.010

21. Lee EB, Heo GE, Choi CM, Song M. MLM-based typographical error correction of unstructured medical texts for named entity recognition. *BMC Bioinform*. 2022;23(1):486.

    doi: 10.1186/s12859-022-05035-9

22. Boudjellal N, Zhang H, Khan A, *et al*. ABioNER: a BERT-based model for Arabic biomedical named-entity recognition. *Complexity*. 2021;2021(1):6633213.

    doi: 10.1155/2021/6633213

23. Perera N, Dehmer M, Emmert-Streib F. Named entity recognition and relation detection for biomedical information extraction. *Front Cell Dev Biol*. 2020;8:673.

    doi: 10.3389/fcell.2020.00673

24. Berg H, Henriksson A, Dalianis H. The impact of de-identification on downstream named entity recognition in clinical text. In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics; 2020:1-11.

    doi: 10.18653/v1/2020.louhi-1.1

25. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021:n71.

    doi: 10.1136/bmj.n71

26. Rivera-Zavala RM, Martínez P. Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization. *BMC Bioinformatics*. 2021;22(S1):601.

    doi: 10.1186/s12859-021-04247-9

27. Amin S, Goldstein NP, Wixted MK, García-Rudolph A, Martínez-Costa C, Neumann G. Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts. *arXiv*. Preprint posted online 2022.

    doi: 10.48550/arXiv.2204.04775

28. Lancheros BS, Corpas Pastor G, Mitkov R. Data augmentation and transfer learning for cross-lingual named entity recognition in the biomedical domain. *Lang Resour Eval*. 2025;59(2):665-684.

    doi: 10.1007/s10579-024-09738-8

29. Vakili T, Henriksson A, Dalianis H. Data-Constrained Synthesis of Training Data for De-Identification. *arXiv*. Preprint posted online 2025.

    doi: 10.48550/arXiv.2502.14677

30. Gallego F, López-García G, Gasco-Sánchez L, Krallinger M, Veredas FJ. ClinLinker: Medical Entity Linking of Clinical Concept Mentions in Spanish. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2404.06367

31. López-García G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Transformers for clinical coding in Spanish. *IEEE Access*. 2021;9:72387-72397.

    doi: 10.1109/ACCESS.2021.3080085

32. Villena F, Bravo-Marquez F, Dunstan J. NLP modeling recommendations for restricted data availability in clinical settings. *BMC Med Inform Decis Mak*. 2025;25(1):116.

    doi: 10.1186/s12911-025-02948-2

33. Sun C, Yang Z, Wang L, Zhang Y, Lin H, Wang J. Deep learning with language models improves named entity recognition for PharmaCoNER. *BMC Bioinform*. 2021;22(S1).

    doi: 10.1186/s12859-021-04260-y

34. Gallego F, Veredas FJ. Recognition and normalization of multilingual symptom entities using in-domain-adapted BERT models and classification layers. *Database*. 2024;2024:baae087.

    doi: 10.1093/database/baae087

35. Akhtyamova L, Martinez P, Verspoor K, Cardiff J. Testing contextualized word embeddings to improve NER in Spanish clinical case narratives. *IEEE Access*. 2020;8:164717-164726.

    doi: 10.1109/ACCESS.2020.3018688

36. Sasikumar N, Mantri KSI. Transfer learning for low-resource clinical named entity recognition. In: *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2023:514-518.

    doi: 10.18653/v1/2023.clinicalnlp-1.53

37. Fabregat H, Martinez-Romo J, Araujo L. Understanding and improving disability identification in medical documents. *IEEE Access*. 2020;8:155399-155408.

    doi: 10.1109/ACCESS.2020.3019178

38. Mullick A, Mondal I, Ray S, Raghav R, Chaitanya G, Goyal P. Intent identification and entity extraction for healthcare queries in Indic languages. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics; 2023:1870-1881.

    doi: 10.18653/v1/2023.findings-eacl.140

39. Yada S, Nakamura Y, Wakamiya S, Aramaki E. Real-MedNLP: overview of REAL document-based MEDical natural language processing task. In: *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*. National Institute of Informatics; 2022:285-296.

40. Phan U, Nguyen N. Simple semantic-based data augmentation for named entity recognition in biomedical texts. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2022:123-129.

    doi: 10.18653/v1/2022.bionlp-1.12

41. Miftahutdinov Z, Alimova I, Tutubalina E. On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2020:281-288.

    doi: 10.1007/978-3-030-45442-5_35

42. Pengpun P, Tiankanon K, Chinkamol A, *et al*. On creating an English-Thai code-switched machine translation in medical domain. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics; 2024:6055-6073.

    doi: 10.18653/v1/2024.findings-emnlp.351

43. Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. *J Biomed Inform*. 2020;102:103356.

    doi: 10.1016/j.jbi.2019.103356

44. Idrissi-Yaghir A, Dada A, Schäfer H, *et al*. Comprehensive Study on German Language Models for Clinical and Biomedical Text Understanding. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2404.05694

45. Fontaine X, Gaschi F, Rastin P, Toussaint Y. Multilingual Clinical NER: Translation or Cross-lingual Transfer? *arXiv*. Preprint posted online 2023.

    doi: 10.48550/arXiv.2306.04384

46. Frei J, Kramer F. Annotated dataset creation through large language models for non-English medical NLP. *J Biomed Inform*. 2023;145:104478.

    doi: 10.1016/j.jbi.2023.104478

47. Schäfer H, Idrissi-Yaghir A, Horn P, Friedrich C. Cross-language transfer of high-quality annotations: combining neural machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In: *Proceedings of the 4th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2022:53-62.

    doi: 10.18653/v1/2022.clinicalnlp-1.6

48. Frei J, Kramer F. GERNERMED: an open German medical NER model. *Softw Impacts*. 2022;11:100212.

    doi: 10.1016/j.simpa.2021.100212

49. Jantscher M, Gunzer F, Kern R, Hassler E, Tschauner S, Reishofer G. Information extraction from German radiological reports for general clinical text and language understanding. *Sci Rep*. 2023;13(1).

    doi: 10.1038/s41598-023-29323-3

50. Abdulnazar A, Roller R, Schulz S, Kreuzthaler M. Large language models for clinical text cleansing enhance medical concept normalization. *IEEE Access*. 2024;12:147981-147990.

    doi: 10.1109/ACCESS.2024.3472500

51. Zaghir J, Bjelogrlic M, Goldman JP, Aananou S, Gaudet-Blavignac C, Lovis C. FRASIMED: a clinical French annotated resource produced through crosslingual BERT-based annotation projection. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL; 2024:7450-7460.

    doi: 10.63317/4z8mxey3dmxp

52. Copara J, Knafou J, Naderi N, Moro C, Ruch P, Teodoro D. Contextualized French language models for biomedical named entity recognition. In: *Actes de la 6e Conférence Conjointe JEP-TALN-RÉCITAL, Atelier DÉfi Fouille de Textes*. ATALA et AFCP; 2020:36-48. Accessed January 31, 2026. https://aclanthology.org/2020.jeptalnrecital-deft.4/

53. Grabar N, Dalloux C, Claveau V. CAS: corpus of clinical cases in French. *J Biomed Semantics*. 2020;11(1):7.

    doi: 10.1186/s13326-020-00225-x

54. Minh N, Tran VH, Hoang V, Ta HD, Bui TH, Truong SQH. ViHealthBERT: pre-trained language models for Vietnamese in health text mining. In: *Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resources Association; 2022:328-337. Accessed January 31, 2026. https://aclanthology.org/2022.lrec-1.35/

55. Byun S, Hong J, Park S, *et al*. Korean Bio-Medical Corpus (KBMC) for Medical Named Entity Recognition. *arXiv*. Published online 2024.

    doi: 10.48550/arXiv.2403.16158

56. Wang Y, Lin M, Hu Q, Bao L, Bai S, Li Y. Large and small models for collaborative cross-lingual data augmentation in entity relationship extraction for low-resource languages. *J King Saud Univ Comput Inf Sci*. 2025;37(4):56.

doi: 10.1007/s44443-025-00055-w

57. Sazzed S. BanglaBioMed: a biomedical named-entity annotated corpus for Bangla (Bengali). In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2022:323-329.

    doi: 10.18653/v1/2022.bionlp-1.31

58. Lu F, Qi Q, Qin H. Joint extraction of Uygur medicine knowledge with edge computing. *Tsinghua Sci Technol*. 2025;30(2):782-795.

    doi: 10.26599/TST.2024.9010006

59. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. *BMC Med Inform Decis Mak*. 2020;20(S7):242.

    doi: 10.1186/s12911-020-01241-8

60. Sousa H, Pasquali A, Jorge A, Santos CS, Lopes MA. A biomedical entity extraction pipeline for oncology health records in Portuguese. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. ACM; 2023:950-956.

    doi: 10.1145/3555776.3578577

61. Schneider ETR, de Souza JVA, Knafou J, *et al*. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2020:65-72.

    doi: 10.18653/v1/2020.clinicalnlp-1.7

62. Šuvalov H, Laur S, Kolde R. Information extraction from medical texts with BERT using human-in-the-loop labeling. In: *Studies in Health Technology and Informatics*. IOS Press; 2023.

    doi: 10.3233/SHTI230281

63. Šuvalov H, Lepson M, Kukk V, *et al*. Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study. *J Med Internet Res*. 2025;27:e66279.

    doi: 10.2196/66279

64. Coelho Da Silva T, Fernandes De Macêdo J, Magalhães R. Tracking the evolution of Covid-19 symptoms through clinical conversations. In: *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2023:41-47.

    doi: 10.18653/v1/2023.clinicalnlp-1.6

65. Vakili T, Hansson M, Henriksson A. SweClinEval: a benchmark for Swedish clinical natural language processing. In: *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*. University of Tartu Library; 2025:767-775. Accessed January 31, 2026. https://aclanthology.org/2025.nodalida-1.76/

66. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records. *IEEE Access*. 2021;9:19097-19110.

    doi: 10.1109/access.2021.3054479

67. Bondarenko DA, Ferrod R, Di Caro L. Combining Contrastive Learning and Knowledge Graph Embeddings to develop medical word embeddings for the Italian language. *arXiv*. Preprint posted online 2022.

    doi: 10.48550/arXiv.2211.05035

68. Postiglione M. Towards an Italian healthcare knowledge graph. In: *Lecture Notes in Computer Science*. Springer International Publishing; 2021:387-394.

    doi: 10.1007/978-3-030-89657-7_29

69. Vassileva S, Todorova G, Ivanova K, *et al*. Automatic Transformation of Clinical Narratives into Structured Format. In: *Proceedings of the Student Research Workshop Associated with RANLP 2021*. INCOMA Ltd. Shoumen, Bulgaria; 2021:219-227.

    doi: 10.26615/issn.2603-2821.2021_030

Appendix A

Table A1. Publications reviewed

Publication	Reference	Target language(s)	NER subtask	Base model	Data augmentation used?	Best reported F1 (%)	Statistical significance reported?	Availability of code and data
The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study	3	English, Spanish	NER	mBERT; BiLSTM-CRF	No	91.7	No	Code only
ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition	22	Arabic	NER	AraBERT	No	85.65	No	Yes
Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization	26	Spanish, English	NER; named entity normalization	mBERT	No	88.8	No	Yes
Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts	27	Spanish (Catalan)	NER; de-identification PHI removal	mBERT	No	97.2	No	Yes
Data augmentation and transfer learning for cross-lingual Named Entity Recognition in the biomedical domain	28	Spanish, English	NER	mBERT; XLM-RoBERTa	Yes	86.39	No	Na
Data-Constrained Synthesis of Training Data for De-Identification	29	Swedish, Spanish	NER; de-identification PHI removal	BERT	Yes	92.6	No	Data only
ClinLinker: Medical Entity Linking of Clinical Concept Mentions in Spanish	30	Spanish	Named entity normalization	SapBERT	No	NA	No	Code only
Transformers for Clinical Coding in Spanish	31	Spanish	NER; named entity normalization	mBERT; BETO; XLM-RoBERTa	No	NA	No	Yes
NLP modeling recommendations for restricted data availability in clinical settings	32	Spanish	NER	mBERT; BETO; RoBERTa; GPT-3.5	No	37.25	No	No
Deep learning with language models improves named entity recognition for PharmaCoNER	33	Spanish	NER offset detection and entity classification	mBERT; BETO; BioBERT; PubMedBERT	No	92.01	No	Code only
Recognition and normalization of multilingual symptom entities using in-domain-adapted BERT models and classification layers	34	Spanish, Portuguese, Swedish	NER	SapBERT	Yes	74.8	No	Yes

(*Cont'd...*)

Table A1. (*Continued*)

Publication	Reference	Target language(s)	NER subtask	Base model	Data augmentation used?	Best reported F1 (%)	Statistical significance reported?	Availability of code and data
Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives	35	Spanish	NER	mBERT; Flair; FastText	No	90.84	No	Yes
Transfer Learning for Low-Resource Clinical Named Entity Recognition	36	Spanish, French, English	NER	BioBERT; RoBERTa-ES; Google Cloud Translate	Yes	94.7	No	Na
Understanding and Improving Disability Identification in Medical Documents	37	Spanish, English	NER	BiLSTM-CRF; GloVe/character embeddings	No	83	No	No
Intent Identification and Entity Extraction for Healthcare Queries in Indic Languages	38	Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati, English	Intent identification; NER	mBERT; XLM-RoBERTa; IndicBERT	No	74.94	No	Data only
Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study	39	Estonian	NER	GPT; Llamas; BERT	Yes	69	No	Data only
Simple Semantic-based Data Augmentation for Named Entity Recognition in Biomedical Texts	40	English, Vietnamese	NER	BioBERT, PhoBERT	Yes	87.73	No	Na
On Biomedical Named Entity Recognition: Experiments in Interlingual Transfer for Clinical and Social Media Texts	41	Russian, English	NER	mBERT, XLM-RoBERTa	No	85.0	No	Na
On Creating an English-Thai Code-switched Machine Translation in Medical Domain	42	Thai, English	NER	NLLB; GPT	No	NA	No	Yes
Terminologies augmented recurrent neural network model for clinical named entity recognition	43	French, English	NER	BiGRU-CRF	No	92.2	No	Yes
Comprehensive Study on German Language Models for Clinical and Biomedical Text Understanding	44	German	NER; multi-label classification	BERT; mBERT; RoBERTa	Yes	85.5	No	Data only
Multilingual Clinical NER: Translation or Cross-lingual Transfer?	45	French, German	NER	mBERT XLM-RoBERTa	Yes	79.1	No	Code only

(*Cont'd...*)

Table A1. (*Continued*)

Publication	Reference	Target language(s)	NER subtask	Base model	Data augmentation used?	Best reported F1 (%)	Statistical significance reported?	Availability of code and data
Annotated dataset creation through large language models for non-English medical NLP	46	German	NER; named entity normalization; text classification	GPT; BERT; GPTNERMED	No	84.7	No	Yes
Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language	47	German	NER	mBERT; BERT aligner; FastAlign	Yes	95.4	No	No
GERNERMED: An open German medical NER model	48	German	NER	SpaCy NLP pipeline	No	81.54	No	Yes
Information extraction from German radiological reports for general clinical text and language understanding	49	German	NER relation extraction	German-MedBERT; R-BERT	No	~83	No	No
Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization	50	German	NER; named entity normalization	GPT-4; SapBERT	No	75.4	No	No
FRASIMED: a Clinical French Annotated Resource Produced through Crosslingual BERT-Based Annotation Projection	51	French	NER	mBERT; CamemBERT	Yes	96.4	No	No
Contextualized French Language Models for Biomedical Named Entity Recognition	52	French	NER	CamemBERT	No	72.62	No	Yes
CAS: corpus of clinical cases in French	53	French	NER information extraction	CRF; BiLSTM-CRF; CRF + active learning	No	91.3	No	Partial
ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining	54	Vietnamese	NER; acronym disambiguation; FAQ summarization	PhoBERT; viBERT; mBERT; ViHealthBERT	No	96.77	No	Yes
Korean Bio-Medical Corpus (KBMC) for Medical Named Entity Recognition	55	Korean	NER	GPT-3.5-turbo; KOSTOM; KM-BERT	Yes	88.86	No	No
Large and Small models for collaborative cross-lingual data augmentation in entity relationship extraction for low-resource languages	56	Mongolian, classical Chinese	Joint NER relation extraction	GPT-4; TPLinker,; OneRel; HugIE; SPKE	Yes	65.61	No	Data only

(*Cont'd...*)

Table A1. (*Continued*)

Publication	Reference	Target language(s)	NER subtask	Base model	Data augmentation used?	Best reported F1 (%)	Statistical significance reported?	Availability of code and data
BanglaBioMed: A Biomedical Named-Entity Annotated Corpus for Bangla (Bengali)	57	Bangla (Bengali)	NER	NA	No	NA	No	Code only
Joint Extraction of Uygur Medicine Knowledge with Edge Computing	58	Uyghur	Joint NER relation extraction	mBERT	No	91.54	Yes	No
Korean clinical entity recognition from diagnosis text using BERT	59	Korean	NER	mBERT	No	84	No	Yes
A Biomedical Entity Extraction Pipeline for Oncology Health Records in Portuguese	60	Portuguese (European)	NER; named entity normalization	BERTimbau	No	95.0	No	No
BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition	61	Portuguese (Brazilian)	NER	mBERT; BioBERTpt	No	92.6	No	Yes
Information Extraction from Medical Texts with BERT Using Human-in-the-Loop Labeling	62	Estonian	NER	BERT	No	NA	No	No
Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task	63	Japanese, English	NER; case identification	BERT; BioBERT; ClinicalBERT; PubMedBERT; UTH-BERT; mBERT; XLM-RoBERTa	Yes	89.26	No	Code only
Tracking the Evolution of Covid-19 Symptoms through Clinical Conversations	64	Portuguese (Brazilian)	NER	BERT; mBERT	No	85.66	No	No
SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing	65	Swedish	NER; document-level classification	BERT; mBERT	No	96.5	No	Data only
A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records	66	Italian	NER; de-identification PHI removal	mBERT; Flair; FastText	No	83.08	No	No
Combining Contrastive Learning and Knowledge Graph Embeddings to Develop Medical Word Embeddings for the Italian Language	67	Italian	NER	mBERT; KGE models	No	NA	No	Yes

(*Cont'd...*)

Table A1. (*Continued*)

Publication	Reference	Target language(s)	NER subtask	Base model	Data augmentation used?	Best reported F1 (%)	Statistical significance reported?	Availability of code and data
Towards an Italian Healthcare Knowledge Graph	68	Italian	NER	BERT; mBERT	No	NA	No	Na
Automatic Transformation of Clinical Narratives into Structured Format	69	Bulgarian	NER text classification	ClinicalBERT; MBG-ClinicalBERT	No	NA	No	No

Abbreviations: BERT: Bidirectional encoder representations from transformers; CRF: Conditional random field; FAQ: Frequently asked questions; KGE: Knowledge graph embedding; LLM: Large language model; NA: Not available; NER: Named entity recognition; NLLB: No Language Left Behind; PHI: Protected health information.