

Ontology-Integrated Natural Language Processing for Symptom Extraction and Disease Severity Staging from Clinical Notes: A Systematic Review

***Job Inyangala¹, Fidelis Musena Mukudi¹, Ronald Ojino², Morris Senghor Shisanya³**

¹Department of Computer Science, School of Computing and Mathematics, The Co-operative University of Kenya

²Department of Computing and Informatics, School of Science & Technology, The Open University of Kenya

³Department of Community Health and Mental Health Nursing, School of Nursing, Kibabii University

***Corresponding E-mail: jobnyangala@gmail.com**

Abstract

African health systems are starting to digitise clinical records at a speed much faster than they can interpret the information. Even though most of the symptom evidence for disease staging is present in unstructured clinical narratives in the EMR, such evidence is often not accessible to automated decision support. Natural Language Processing (NLP) has advanced rapidly in clinical entity extraction, but the contribution of formal clinical ontologies to extraction and, crucially, to severity staging has not been systematically assessed. This review collates peer-reviewed evidence on ontology-integrated NLP for symptom extraction and severity classification. Six databases were searched using terms combining natural language processing, symptom extraction and severity staging (PubMed, Scopus, IEEE Xplore, ACM Digital Library, Web of Science, Google Scholar) for the period from January 2019 to June 2025. After applying pre-specified eligibility criteria, the records were screened and forty-one studies were included in the narrative synthesis, in accordance with PRISMA 2020 guidelines. Studies were characterised by clinical domain, corpus, ontology, integration strategy, architecture and evaluation design, and appraised using an adapted QUADAS-2 instrument. Three integration approaches were identified: dictionary and concept mapping, ontology-supervised distant labelling, and ontology-constrained neural architectures. Ontology integration consistently improved concept normalisation, cross-corpus transfer and annotation efficiency, with the largest gains in low-annotation settings. Only nine studies addressed severity staging; these were largely confined to cardiology and respiratory medicine and were seldom validated against clinical grading instruments. No included study used an African clinical corpus. The evidence base is therefore methodologically well developed for extraction, weak for staging and geographically narrow. For automated severity staging to support clinical decision-making at district-hospital level, African health systems need infrastructure as much as computation: de-identified African clinical corpora with local extensions to the terminology, ontology-guided staging validated against recognised grading instruments, and transparent reporting.

Keywords: natural language processing, clinical ontology, SNOMED CT, symptom extraction, severity staging

Introduction

In health systems globally, Electronic Medical Records (EMRs) have emerged as the primary source of clinical data, with most of the clinically meaningful information being entered as free text, rather than in coded fields. Typically, symptom descriptions are included in admission notes, ward rounds and discharge summaries, as are functional limitations, exacerbation histories and the clinician's own severity judgements. Because this information is unstructured, it is unavailable to automated decision support, quality reporting and disease surveillance, and can be retrieved only by manual chart review, which is time-consuming and inconsistent between reviewers (Song et al., 2022; Sarwar et al., 2022). One solution to this is the use of Natural Language Processing (NLP). The task of identifying and annotating clinical concepts in continuous text (known as Named Entity Recognition (NER)) has made significant strides since the introduction of contextual embeddings and transformer models, with performance on benchmark corpora approaching inter-annotator agreement (IAA) (Yang et al., 2020a; Durango et al., 2023). A criticism nevertheless recurs systems recognise a symptom when it is mentioned but do not situate it within the clinical knowledge structure that gives it meaning. It is not enough to record that a patient is dyspnoeic: under the New York Heart Association (NYHA) criteria, a note that the patient is breathless when dressing denotes a specific functional class. The critical transition in the process of clinical decision support is from detection to graded interpretation (Steinkamp et al., 2020; Adejumo et al., 2024).

Clinical ontologies are the obvious source of that structure. In addition to preferred terms and synonyms, resources like the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) and the Unified Medical Language System (UMLS) encode hierarchical and attribute relationships between concepts, enabling a finding to be resolved to a normalised identifier and located in relation to its parent, qualifiers and severity attributes (Vuokko et al., 2023; Schulz et al., 2023). Ontologies have therefore been integrated into NLP pipelines in three ways: as a lexical resource for candidate generation, as a source of supervision where manual annotation is limited, and as a constraint on the output space of a neural model (Arbabi et al., 2019; Slater et al., 2021; Dong et al., 2023). Individually these studies report encouraging results. They have not, however, been evaluated against the question a system designer needs answered: Does the integration of ontologies provide a measurable improvement in symptom extraction, and can it be extended to severity staging?

No existing secondary literature addresses this question directly. Methodological reviews analyse architectures and corpora but treat terminology as an implementation detail, and domain-scoped reviews of dentistry, nursing and post-acute care do not isolate the contribution of ontologies, nor do terminology-scoped reviews evaluate staging performance, except for emerging use with LLMs (Durango et al., 2023; Ahmad et al., 2023; Bose et al., 2021; Vuokko et al., 2023; Chang & Sung, 2024). Ontology integration is thus widely recommended but poorly documented. The gap is felt most sharply in African health systems. Where clinician-to-population ratios are low and documentation is narrative-only, the potential value of automated symptom staging for triage, referral and district-level surveillance is high, but the annotated corpora on which supervised clinical NLP depends are essentially non-existent. Ontology-driven and weakly supervised approaches are appealing in this context because they reduce the need for annotation, yet they have not been examined on African clinical text.

This review therefore synthesises peer-reviewed evidence from clinical notes on symptom extraction and disease severity staging over the period 2019–2025, using ontology-integrated NLP. It answers four questions: (i) what resources and integration strategies have been used in relation to ontologies, (ii) what model architectures have been used, (iii) what performance has been reported for extraction and staging, and under which evaluation designs, and (iv) which methodological and geographical gaps limit the deployment in low-resource health systems. The article sets out where ontology integration demonstrably helps, what this implies for African settings, and the research agenda that follows.

Materials and Methods

The review was carried out as a systematic review of methodological studies and reported in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 statement. The protocol was adapted to a narrative synthesis design because the included studies were on computational systems, not on clinical interventions, and because differences in corpora, entity definitions and evaluation metrics preclude statistical pooling.

Six bibliographic sources were searched: PubMed, Scopus, IEEE Xplore, the ACM Digital Library, Web of Science and Google Scholar. Publications were searched from 1 January 2019 to 30 June 2025, reflecting the recent methodological generation from the time when transformer architectures began to be widely used for clinical text mining. The search string that was used consisted of four concept blocks combined with Boolean operators: (a) natural language processing OR text mining OR named entity recognition OR information extraction; (b) ontology OR terminology OR SNOMED CT OR UMLS OR knowledge graph OR controlled vocabulary; (c) symptom OR clinical finding OR sign OR functional status; and (d) severity OR staging OR grading OR classification OR stratification. Field tags were modified for each of the databases. Reference lists of all included studies were hand-searched, and forward citation tracking was performed in Google Scholar for relevant reviews.

Inclusion and exclusion criteria were pre-defined and are summarised in Table 1. Studies were included if they used an NLP method to analyse clinical free text, explicitly referenced a biomedical ontology or controlled terminology in at least one step of the pipeline, focused on symptoms or clinical findings in the free text, and at least one quantitative performance measure was reported. Studies were excluded if they lacked either a clinical text source or an operational terminology component, described a system without evaluating it, or reported no quantitative results.

Table 1. Eligibility criteria applied during screening

Domain	Inclusion	Exclusion
Text source	Clinical notes, discharge summaries, radiology or pathology reports from EMR/EHR systems	Biomedical abstracts, social media, synthetic text with no clinical provenance
Ontology use	Ontology or controlled terminology used for lexicon construction, normalisation, supervision or output constraint	Terminology mentioned only in discussion; no operational role in the pipeline
Target concept	Symptoms, signs, clinical findings, functional status, or their severity	Medication, procedure or laboratory extraction only
Outcome	At least one reported metric (precision, recall, F1, accuracy, AUC, kappa)	Architecture description or position paper with no evaluation
Publication	Peer-reviewed journal or conference paper in English, 2019–2025	Preprints, theses, editorials, non-English publications

Screening and Selection

Records were downloaded into a reference manager and deduplicated. Titles and abstracts were scanned to find items that met the criteria in Table 1, and the full text of the items that passed the initial screen was then evaluated. The screening was done twice by the author with a two-week interval between passes, and any disagreement between the passes was resolved by re-reading the full text in relation to the criteria, and intra-rater agreement between the two passes was good (Cohen's kappa = 0.84). The search returned 1,412 records, of which 386 were duplicates. The remaining 1,026 were screened at title and abstract, 902 were excluded and 124 were assessed in full text. Eighty-three full texts were excluded because the ontology played no operational role in the pipeline ($n = 31$), the target concepts were not symptoms or findings ($n = 22$), no quantitative evaluation was reported ($n = 14$), the text source was not clinical ($n = 11$) or the paper duplicated a system reported elsewhere ($n = 5$). Forty-one studies were included in the synthesis. The selection process is summarised in Table 2.

Table 2. Study identification, screening and inclusion flow (PRISMA 2020)

Stage	Records
Identified through database searching	1,412
Duplicates removed	386
Screened at title and abstract	1,026
Excluded at screening	902
Assessed in full text	124
Excluded at full text (with reasons)	83
Included in narrative synthesis	41

Data Extraction

A structured extraction form was developed, piloted on five studies and revised before full extraction. The following fields were extracted: bibliographic information, country of data source, clinical domain, corpus and corpus size, ontology or terminology used, point in the pipeline where the ontology was applied, model architecture, external or cross-corpus validation conducted, evaluation design (whether or not severity staging was attempted and against which grading instrument), and all reported performance metrics with their definitions. Where multiple model variants were reported, the best-performing variant was recorded, and the condition without ontology integration (as a comparator) was recorded separately so that the incremental contribution of ontology integration could be assessed.

Quality Appraisal

The methodological quality was evaluated by an instrument adapted for computational studies from QUADAS-2, which includes the following five domains: adequacy of reference standard, adequacy of the reference standard's provenance, representativeness of the corpus relative to the intended deployment setting, completeness of metric



reporting, including per-entity results and variance, and independence of the test partition from model development, including availability of code, corpus and ontology version. Each domain, and the study overall, was rated as low, unclear or high concern. Appraisal results were not used to exclude studies but to weigh confidence in the synthesis.

Synthesis

Results were summarised narratively. Studies were first grouped by ontology-integration strategy, then cross-tabulated with clinical domain, architecture and evaluation design. Reported performance was compared within integration strategies and only when entity definitions and metric conventions were similar. Because clinical NER evaluation is heterogeneous, performance figures are reported as observed ranges with the corresponding corpus named, and no pooled figure is given.

Results and Discussion

Characteristics of Included Studies

The 41 included studies were unevenly distributed across the review period, concentrated in the final three years, reflecting the diffusion of transformer-based clinical models. Most studies drew on data from North America and Western Europe. A substantial minority used publicly available de-identified corpora, mainly the MIMIC family and the eICU Collaborative Research Database, while the remainder used single-institution corpora that were not released. The clinical domains included substantial groupings in cardiology, oncology, respiratory medicine and general internal medicine, and smaller groupings in nursing documentation, nephrology, dentistry, and post-acute care (Chan et al., 2020; Mitha et al., 2023; Pethani & Dunn, 2023; Scharp et al., 2024). The most common terminology resources were SNOMED CT and the UMLS Metathesaurus, while a smaller set employed disease-specific or phenotype ontologies, and a few constructed custom lexicons anchored to a standard terminology. The distribution of studies in terms of main extraction dimensions is summarised in Table 3.

Table 3. Distribution of included studies across extraction dimensions

Dimension	Category	Studies (n = 41)
Ontology resource	SNOMED CT	17
	UMLS Metathesaurus	13
	Domain or phenotype ontology	7
	Bespoke lexicon anchored to a standard terminology	4
Integration strategy	Dictionary and concept mapping	16
	Ontology-supervised distant labelling	12
	Ontology-constrained neural architecture	13
Model architecture	Rule-based or dictionary pipeline	9
	Bi-LSTM-CRF or comparable sequence labeller	11
	Transformer encoder (BERT variants, clinical pretraining)	16
	Large language model, zero- or few-shot	5
Severity staging	Attempted	9
	Validated against a recognised grading instrument	4
Validation design	Internal split only	28
	External or cross-corpus validation	13
Geography of corpus	North America	22
	Europe	13
	Asia	6
	Africa	0

Ontology Integration Strategies

The synthesis identified three strategies. The first, dictionary and concept mapping, is a lexical resource approach: candidate spans are aligned with preferred terms and synonyms and disambiguated to concept identifiers, often supplemented with negation and context. Established clinical annotators follow this approach, and comparative testing of four such systems (MetaMap, cTAKES, CLAMP and commercial services) found that performance depends less on the matching algorithm than on how much of the target vocabulary the lexicon covers (Bai et al., 2021). Slater et al. (2021) demonstrated the corollary: increasing the number of terms attached to ontology classes improved characterisation of clinical text without any modification to the underlying model, and most residual errors traced to missing terms in the ontology rather than to the learner.

The second approach uses ontology as a source of supervision. Instead of manually labelled spans, concept labels are projected on unlabelled text to produce "noisy" training data that is then used to fit a supervised model. Dong et al. (2023) applied this to rare-disease identification in clinical notes, and Kersloot et al. (2019) showed that SNOMED CT concept and attribute relationships can be detected automatically at scale. The appeal of this family is economic: what constrains most clinical NLP projects is not model capacity but the cost of annotation. Distant supervision alleviates this constraint. The price is label noise, and the studies that reported error analysis found that most of the error was due to mismatches in the granularity of the ontology, where the projected concept is either more general or more specific than the clinical intent of the note.

The third strategy builds ontological structure into the model itself. Arbabi et al. (2019) constrained a neural encoder to the ontology's concept space so that unseen synonyms resolve to the correct class; Thukral et al. (2023) coupled NER output with biomedical ontologies, so that extraction and structuring could be a single process. Alahmar et al. (2022) standardised clinical pathways using SNOMED CT semantic relatedness, demonstrating that the ontology's relational information, rather than just the list of terms, has useful signal. Berge et al. (2023) found that combining paradigms — rule-based, supervised and unsupervised — outperformed any single paradigm, suggesting that the useful question is how to combine them rather than which to choose.

Across the three strategies, the common conclusion is that ontology integration leads to better normalisation and generalisation than just raw span detection. The studies that reported both span-level and concept-level data tended to have small gains at the span level and larger gains at the concept level, as the ontology can account for synonym and abbreviation variation that is not addressed by span detection. This distinction is often lost in reporting, and it is the qualification that matters most when interpreting claims that ontologies improve extraction.

Model Architectures and Their Interaction with Ontologies

Regarding architecture, the studies that are included chart the recent history of the field. Where interpretability and auditability are needed, rule-based and dictionary pipelines are still used, and Topaz et al. (2019) found rule-based and embedding pipelines comparable for mining fall-related information, while Prakash et al. (2024) found them similarly accurate for extracting dementia severity. Sequence labellers from the Bi-LSTM-CRF family remain appealing where training data are scarce and computational resources are limited (Dash et al., 2024). In later studies the modal architecture is the transformer (Yang et al., 2020a; Yang et al., 2022; Rohanian et al., 2024), with clinical-domain pretraining providing a certain advantage over general-domain models.

Two cautions recur. First, transformers are not uniformly effective: Gao et al. (2021) identified label distribution and long documents as persistent difficulties in clinical text classification. Second, the newest models are not always the strongest. An optimised BERT-based pipeline outperformed zero-shot approaches (Ochoa et al., 2025), and a scoping review concluded that the use of SNOMED CT with LLMs remains at an early stage (Chang & Sung, 2024). For symptom staging, which requires mapping evidence to an ordinal class, the evidence is more in favour of smaller supervised models and an explicit ontology than generative models. Rohanian et al. (2024) also demonstrated that lightweight transformers can maintain the accuracy of larger models, at a fraction of the inference cost, which is directly material where computing resources are limited.

Severity Staging and Its Validation

Of the 41 studies, only nine assigned a severity grade rather than simply detecting concepts, and only four validated that grade against a recognised clinical grading instrument. The best evidence is in cardiology. Solomon et al. (2021) used NLP to identify aortic stenosis and its severity from echocardiogram reports in a large, integrated health system, and found positive and negative predictive values greater than 95 per cent; Adejumo et al. (2024) used NLP to classify NYHA functional status from narrative text documentation, which had an AUROC of 0.99 at the development site and 0.98 at two external sites. These two studies represent the methodological benchmark in this literature: both used an ordinal target derived from a validated instrument and a clinically adjudicated reference standard, and one was externally validated. In respiratory medicine, Quint et al. (2021) offer a partial analogue, deriving and validating a Global Initiative for Chronic Obstructive Lung Disease severity estimate across multiple datasets and Prakash et al. (2024) show that a rule-based approach can recover dementia severity from clinicians' notes where the grading construct is clearly documented.

The remaining staging studies fell short in one of three ways. Some used a proxy for the grading instrument, such as billing codes or a coarse severe/not-severe dichotomy, neither of which corresponds to the clinical decision threshold; for grading haemodialysis symptom burden, Chan et al. (2020) demonstrated that codes and notes vary significantly and code-based severity labels were not a reliable ground truth. Others reported classification metrics that ignore the ordering of severity, so that an error one grade away is penalised as heavily as an error at the opposite end of the scale. A third group inferred severity from ontology attributes without demonstrating correspondence with clinicians' assessments; according to Cazzaniga et al. (2023) the accuracy of ontology-based labelling of pathology reports was 0.84–0.87, which is plausible for concept assignment but not for validation of severity grading. Progress in the field has therefore been concentrated in concept detection, with little work on graded interpretation, which is the tighter constraint on clinical deployment.

Methodological Quality and Reporting

A consistent pattern was identified when the domains of the adapted QUADAS-2 were used for appraisal. The concern about the reference standard was low when clinician adjudication was used and high when the codes and/or ontology attributes were used instead of expert judgement. Corpus representativeness was a concern in most studies, which used a single institution's corpus without describing its case mix. Test partitions were internal in all but the thirteen externally validated studies, so most reported figures describe within-corpus rather than deployed behaviour; and the performance loss associated with institutional domain shift is well documented (Saigaonkar & Narawade, 2024). Reporting was the weakest domain: per-entity results, confidence intervals and run-to-run variance were frequently missing, and the span-level/concept-level distinction was often not stated. Reproducibility was weaker still: code and corpus releases were rarely paired with the ontology release version, even though that version determines which concepts the results can cover. The apparent convergence of reported F1 values in the high 80s and low 90s therefore overstates how comparable the underlying evidence is.

Geographical Concentration and Relevance to African Health Systems

None of the included studies used an African clinical corpus. Twenty-two used North American data, thirteen European and six Asian; the corpora on which much of the field is based are from a small number of intensive care and academic settings across North America. This concentration has three consequences for African deployment. First, documentation conventions vary: abbreviation practices, switching between English and local languages, and the balance of narrative against structured entry are institution-specific, and models trained on one convention do not easily transfer to another. Second, vocabulary coverage of locally salient presentations is incomplete, so ontology-derived pipelines inherit the coverage gaps of the source terminology; the terminology coverage expansion result of Slater et al. (2021) suggests that this is the most tractable point of intervention, since coverage can be extended without retraining the model. Third, the annotation burden facing African projects is precisely the constraint the distant-supervision family is designed to relieve (Dong et al., 2023), which makes lexicon-anchored approaches a more plausible starting point than fully supervised transformer training.

This is echoed by a broader body of literature in the low-resource literature. Knowledge-driven and transfer-based methods dominate knowledge-based information extraction, while a hybrid terminology-plus-model approach has

proved a viable route to semantic annotation beyond English where labelled data are limited (Ariyanto et al., 2024; Campillos-Llanos et al., 2025). On the infrastructure side, Rohanian et al. (2024) is also relevant: lightweight clinical transformers retain almost the accuracy of larger models at substantially lower inference cost. The evidence therefore supports a design rationale rather than a demonstrated result: ontologically guided, annotation-light staging pipelines are the appropriate starting point for African clinical text, but the published record contains no corpus on which the approach could be tested (Dyers et al., 2025).

Limitations of This Review

Three limitations qualify these conclusions. Two time-separated passes with a single reviewer were made for screening and extraction, which could introduce selection bias. Grey literature, preprints and non-English publications were not included, potentially limiting the representation of literature from low- and middle-income countries. Pooled estimates are not provided and should not be inferred because the definitions of entities and conventions of the metrics were not uniform.

Conclusion and Recommendations

Ontology integration is well evidenced for symptom extraction but poorly evidenced for severity staging. Three strategies were identified across 41 studies, and all three improved concept normalisation, cross-corpus transfer and annotation efficiency, with the largest gains where manual annotation was scarcest. The gains are at concept level, not span level, and reporting the former as the latter exaggerates the role that ontologies play in detection. Nine studies attempted severity staging and four validated against a recognised grading instrument; this work was confined largely to cardiovascular and respiratory medicine, and none of it used an African corpus. The evidence base is therefore methodologically well developed for extraction, poorly developed for staging and geographically narrow.

Four recommendations follow. First, staging should be treated as an explicit, ordered task with a clear reference standard: studies should name the grading instrument used, state how narrative evidence was mapped to each grade, and use metrics that penalise misclassification in proportion to the distance between the predicted and the true grade. Second, descriptions of ontology-constrained architectures should report the terminology release version, the concept subset used and any vocabulary-expansion procedure, since coverage accounts for much of the residual error. Third, external or cross-corpus validation should become the default rather than the exception, since institutional domain shift is known to cause substantial performance loss. Fourth, and most fundamental for the region, the creation of de-identified African clinical corpora with local terminology extensions should be treated as research infrastructure, without which ontology-guided staging cannot be evaluated in an African health system. A defensible near-term programme is a lightweight clinical transformer constrained to a locally extended subset of SNOMED CT, distantly supervised from that subset, and evaluated in a single district hospital against clinician-assigned grades before any claim of transferability is made.

Acknowledgements

The authors thank the School of Computing and Mathematics, The Co-operative University of Kenya, for institutional support, and the anonymous reviewers whose comments improved the manuscript.

References

- Adejumo, P., Thangaraj, P. M., Dhingra, L. S., Aminorroaya, A., Zhou, X., Brandt, C., & Khera, R. (2024). Natural language processing of clinical documentation to assess functional status in patients with heart failure. *JAMA Network Open*, 7(11), e2443925. <https://doi.org/10.1001/jamanetworkopen.2024.43925>
- Ahmad, P. N., Shah, A. M., & Lee, K. (2023). A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. *Healthcare*, 11(9), 1268. <https://doi.org/10.3390/healthcare11091268>
- Alahmar, A., AlMousa, M., & Benlamri, R. (2022). Automated clinical pathway standardization using SNOMED CT-based semantic relatedness. *Digital Health*, 8, 20552076221089796. <https://doi.org/10.1177/20552076221089796>
- Arbabi, A., Adams, D. R., Fidler, S., & Brudno, M. (2019). Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR Medical Informatics*, 7(2), e12596. <https://doi.org/10.2196/12596>
- Ariyanto, A. D. P., Purwitasari, D., & Fatichah, C. (2024). A systematic review on semantic role labeling for information extraction in low-resource data. *IEEE Access*, 12, 57917-57946. <https://doi.org/10.1109/ACCESS.2024.3392370>



- Bai, L., Mulvenna, M. D., Wang, Z., & Bond, R. (2021). Clinical entity extraction: Comparison between MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical. In 2021 32nd Irish Signals and Systems Conference (pp. 1-6). IEEE. <https://doi.org/10.1109/ISSC52156.2021.9467856>
- Berge, G. T., Granmo, O. C., Tveit, T. O., Ruthjersen, A. L., & Sharma, J. (2023). Combining unsupervised, supervised and rule-based learning: The case of detecting patient allergies in electronic health records. *BMC Medical Informatics and Decision Making*, 23(1), 188. <https://doi.org/10.1186/s12911-023-02271-8>
- Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18), 8319. <https://doi.org/10.3390/app11188319>
- Campillos-Llanos, L., Valverde-Mateos, A., & Capllonch-Carrion, A. (2025). Hybrid natural language processing tool for semantic annotation of medical texts in Spanish. *BMC Bioinformatics*, 26(1), 7. <https://doi.org/10.1186/s12859-024-05949-6>
- Cazzaniga, G., Eccher, A., Munari, E., Marletta, S., Bonoldi, E., Della Mea, V., & L'Imperio, V. (2023). Natural language processing to extract SNOMED-CT codes from pathological reports. *Pathologica*, 115(6), 318. <https://doi.org/10.32074/1591-951X-952>
- Chan, L., Beers, K., Yau, A. A., Chauhan, K., Duffy, A., Chaudhary, K., & Nadkarni, G. N. (2020). Natural language processing of electronic health records is superior to billing codes to identify symptom burden in haemodialysis patients. *Kidney International*, 97(2), 383-392. <https://doi.org/10.1016/j.kint.2019.10.023>
- Chang, E., & Sung, S. (2024). Use of SNOMED CT in large language models: Scoping review. *JMIR Medical Informatics*, 12(1), e62924. <https://doi.org/10.2196/62924>
- Dash, A., Darshana, S., Yadav, D. K., & Gupta, V. (2024). A clinical named entity recognition model using pretrained word embedding and deep neural networks. *Decision Analytics Journal*, 10, 100426. <https://doi.org/10.1016/j.dajour.2024.100426>
- Dong, H., Suarez-Paniagua, V., Zhang, H., Wang, M., Casey, A., Davidson, E., & Wu, H. (2023). Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Medical Informatics and Decision Making*, 23(1), 86. <https://doi.org/10.1186/s12911-023-02181-9>
- Durango, M. C., Torres-Silva, E. A., & Orozco-Duque, A. (2023). Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research*, 29(4), 286-300. <https://doi.org/10.4258/hir.2023.29.4.286>
- Dyers, R. E., Mahomed, H., & Van Greunen, D. (2025). Design science research in quality improvement: Embedding rigour in digital health innovation. *African Journal of Primary Health Care & Family Medicine*, 17(2), 5194. <https://doi.org/10.4102/phcfm.v17i2.5194>
- Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., & Tourassi, G. (2021). Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 25(9), 3596-3607. <https://doi.org/10.1109/jbhi.2021.3062322>
- Kersloot, M. G., Lau, F., Abu-Hanna, A., Arts, D. L., & Cornet, R. (2019). Automated SNOMED CT concept and attribute relationship detection through a web-based implementation of cTAKES. *Journal of Biomedical Semantics*, 10(1), 14. <https://doi.org/10.1186/s13326-019-0207-3>
- Mitha, S., Schwartz, J., Hobensack, M., Cato, K., Woo, K., Smaldone, A., & Topaz, M. (2023). Natural language processing of nursing notes: An integrative review. *CIN: Computers, Informatics, Nursing*, 41(6), 377-384. <https://doi.org/10.1097/cin.0000000000000967>
- Ochoa, J. G. D., Layer, N., Mahr, J., Mustafa, F. E., Menzel, C. U., Muller, M., & Krohn, A. (2025). Optimized BERT-based NLP outperforms zero-shot methods for automated symptom detection in clinical practice. *medRxiv*, 2025-04. <https://doi.org/10.1101/2025.04.21.25326037>
- Pethani, F., & Dunn, A. G. (2023). Natural language processing for clinical notes in dentistry: A systematic review. *Journal of Biomedical Informatics*, 138, 104282. <https://doi.org/10.1016/j.jbi.2023.104282>
- Prakash, R., Dupre, M. E., Ostbye, T., & Xu, H. (2024). Extracting critical information from unstructured clinicians' notes data to identify dementia severity using a rule-based approach: Feasibility study. *JMIR Aging*, 7, e57926. <https://doi.org/10.2196/57926>
- Quint, J. K., O'Leary, C., Venerus, A., Holmgren, U., Varghese, P., & Cabrera, C. (2021). Development and validation of a method to estimate COPD severity in multiple datasets: A retrospective study. *Pulmonary Therapy*, 7(1), 119-132. <https://doi.org/10.1007/s41030-020-00139-0>
- Rohanian, O., Nouriborji, M., Jauncey, H., Kouchaki, S., Nooralahzadeh, F., & Clifton, L. (2024). Lightweight transformers for clinical natural language processing. *Natural Language Engineering*, 30(5), 887-914. <https://doi.org/10.1017/S1351324923000542>
- Saigaonkar, S., & Narawade, V. (2024). Domain adaptation of transformer-based neural network model for clinical note classification in Indian healthcare. *International Journal of Information Technology*, 1-19. <https://doi.org/10.1007/s41870-024-02053-z>
- Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., & Cavedon, L. (2022). The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys*, 55(2), 1-40. <https://doi.org/10.1145/3490234>
- Scharp, D., Hobensack, M., Davoudi, A., & Topaz, M. (2024). Natural language processing applied to clinical documentation in post-acute care settings: A scoping review. *Journal of the American Medical Directors Association*, 25(1), 69-83. <https://doi.org/10.1016/j.jamda.2023.09.006>
- Schulz, S., Case, J. T., Hendler, P., Karlsson, D., Lawley, M., Cornet, R., & Gao, Y. (2023). SNOMED CT and Basic Formal Ontology - convergence or contradiction between standards? The case of "clinical finding". *Applied Ontology*, 18(3), 207-237. <https://doi.org/10.3233/AO-230018>
- Slater, L. T., Bradlow, W., Ball, S., Hoehndorf, R., & Gkoutos, G. V. (2021). Improved characterisation of clinical text through ontology-based vocabulary expansion. *Journal of Biomedical Semantics*, 12(1), 7. <https://doi.org/10.1186/s13326-021-00241-5>
- Solomon, M. D., Tabada, G., Allen, A., Sung, S. H., & Go, A. S. (2021). Large-scale identification of aortic stenosis and its severity using natural language processing on electronic health records. *Cardiovascular Digital Health Journal*, 2(3), 156-163. <https://doi.org/10.1016/j.cvdhj.2021.03.003>
- Song, J., Hobensack, M., Bowles, K. H., McDonald, M. V., Cato, K., Rossetti, S. C., & Topaz, M. (2022). Clinical notes: An untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care. *Journal of Biomedical Informatics*, 128, 104039. <https://doi.org/10.1016/j.jbi.2022.104039>
- Steinkamp, J. M., Bala, W., Sharma, A., & Kantrowitz, J. J. (2020). Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. *Journal of Biomedical Informatics*, 102, 103354. <https://doi.org/10.1016/j.jbi.2019.103354>
- Thukral, A., Dhiman, S., Meher, R., & Bedi, P. (2023). Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications. *International Journal of Information Technology*, 15(1), 53-65. <https://doi.org/10.1007/s41870-022-01145-y>
- Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., & Bowles, K. H. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, 90, 103103. <https://doi.org/10.1016/j.jbi.2019.103103>



- Vuokko, R., Vakkuri, A., & Palojoki, S. (2023). Systematized Nomenclature of Medicine - Clinical Terminology (SNOMED CT) clinical use cases in the context of electronic health record systems: Systematic literature review. *JMIR Medical Informatics*, 11, e43750. <https://doi.org/10.2196/43750>
- Yang, X., Bian, J., Hogan, W. R., & Wu, Y. (2020a). Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12), 1935-1942. <https://doi.org/10.1093/jamia/ocaa189>
- Yang, X., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., & Wu, Y. (2022). GatorTron: A large language model for clinical natural language processing. *medRxiv*, 2022-02. <https://doi.org/10.2196/19735>