Critical Care and Resuscitation 26 (2024) 210–216

Contents lists available at ScienceDirect

Critical Care and Resuscitation

journal homepage: www.elsevier.com/locate/ccrj

Review Article

# Natural language processing in the intensive care unit: A scoping review

Julia K. Pilowsky, RN, PhD [a, b, c, *], Jae-Won Choi, MBiomedE, BE (Comp), BE-Health (HI) (ProfHons) [a, d], Aldo Saavedra, PhD [a, b], Maysaa Daher, BPsych, MAppStats [a], Nhi Nguyen, MBBS, FCICM [a, b, e], Linda Williams, RN, MHealthManagement [a], Sarah L. Jones, RN, Grad Dip Ed (Nursing), Grad Cert (ICU) [f]

[a] *Agency for Clinical Innovation, NSW Health, Australia;* [b] *University of Sydney, Australia;* [c] *Royal North Shore Hospital, NSW, Australia;* [d] *eHealth, NSW Health, Australia;* [e] *Nepean Hospital, NSW, Australia;* [f] *St George Hospital, NSW, Australia*

ARTICLE INFORMATION

*Article history:*
Received 30 May 2024
Received in revised form
30 June 2024
Accepted 30 June 2024

*Keywords:*
Intensive care medicine
Natural language processing
Artificial intelligence
Scoping review

ABSTRACT

***Objectives:*** Natural language processing (NLP) is a branch of artificial intelligence focused on enabling computers to interpret and analyse text-based data. The intensive care specialty is known to generate large volumes of data, including free-text, however, NLP applications are not commonly used either in critical care clinical research or quality improvement projects. This review aims to provide an overview of how NLP has been used in the intensive care specialty and promote an understanding of NLP's potential future clinical applications.

***Design:*** Scoping review.

***Data sources:*** A systematic search was developed with an information specialist and deployed on the PubMed electronic journal database. Results were restricted to the last 10 years to ensure currency.

***Review methods:*** Screening and data extraction were undertaken by two independent reviewers, with any disagreements resolved by a third. Given the heterogeneity of the eligible articles, a narrative synthesis was conducted.

***Results:*** Eighty-seven eligible articles were included in the review. The most common type (n = 24) were studies that used NLP-derived features to predict clinical outcomes, most commonly mortality (n = 16). Next were articles that used NLP to identify a specific concept (n = 23), including sepsis, family visitation and mental health disorders. Most studies only described the development and internal validation of their algorithm (n = 79), and only one reported the implementation of an algorithm in a clinical setting.

***Conclusions:*** Natural language processing has been used for a variety of purposes in the ICU context. Increasing awareness of these techniques amongst clinicians may lead to more clinically relevant algorithms being developed and implemented.



## 1. Introduction

The intensive care environment produces large amounts of data due to the multiple devices and continuous monitoring systems required to deliver care to critically ill patients. While much of this data is numerical, or structured data, an even greater proportion is unstructured, such as text-based documentation like the clinical progress notes.[1] However, while structured data can be analysed using traditional statistical techniques, interrogation of text-based data requires the use of a specialised group of techniques known as natural language processing (NLP). Public awareness of systems which use NLP has grown recently, with the advent of large language models such as ChatGPT. Despite this however, the use of NLP remains relatively uncommon both in critical care research and in clinical applications designed for use in the ICU.[2]

Natural language processing algorithms have been developed for a variety of tasks in other clinical specialties, including screening for clinical trial eligibility, automatically interpreting medical imaging reports and estimating the risk of disease progression.[3,4] An overview of the typical pipeline used to develop an algorithm

* Corresponding author at: 1 Reserve Road, St Leonards, NSW 2065, Australia.
*E-mail address:* Julia.Pilowsky@health.nsw.gov.au (J.K. Pilowsky).

https://doi.org/10.1016/j.ccrj.2024.06.008

involving NLP is shown in Fig. 1. Literature reviews have been performed to provide an overview of how NLP has been used in other clinical specialties including cardiology, radiology and mental health but has yet to be performed for critical care.[3–5] Given that NLP does not form part of traditional biostatistical methods, these types of reviews are important for increasing awareness amongst clinicians of the potential applications that NLP has both in the delivery of patient care as well as the conduct of clinical research. Therefore, the aim of the review was to provide an overview of how NLP has previously been used in the ICU clinical context and identify opportunities for how these techniques may be used in the future.

## 2. Methods

This review is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for Scoping Reviews.[6] A scoping review methodology was chosen to match the broad aim of the review and the need for narrative synthesis given the likely heterogeneity of relevant articles.[7]

### 2.1. Eligibility criteria and search strategy

To be included in the review, studies had to describe the use of a type of NLP technique on data generated in an ICU environment. This included studies whose primary aim was to develop or validate an NLP algorithm as well as studies which used NLP techniques as part of their methods to address a different research question. All studies had to be written in English, although may describe the use of NLP on text generated in other languages.

A search strategy was developed in collaboration with an information retrieval expert which combined the concepts 'natural language processing' and 'intensive care' and was performed on the PubMed database on the 9th of March 2023 (see Appendix 1). Only studies published in the last 10 years were included in the review. The NLP field has seen rapid development over the last decade, and any algorithm or application described prior to this period would likely have been superseded by more recent technologies. References were assessed for eligibility using the Covidence software. Title and abstract screening was performed by two independent reviewers (JP and JC), with disagreements resolved by a third (AS). Full-text review was similarly performed by two independent reviewers (JP and JC or AS), with disagreements resolved either through discussion or by a third reviewer (JC or AS).

### 2.2. Data extraction and synthesis

Data from each included article was independently extracted by two reviewers (JP, SJ, JC, AS, MD) with discrepancies resolved by discussion until consensus was reached in the review team. Data extraction and consensus were also performed using the Covidence Software, and exported to Microsoft Excel for analysis. Journal subject areas and categories were obtained from SCImago.[8] Extracted data included the aim of the study, the type and purpose of the NLP algorithm, what type of clinical notes were used and how large the dataset was.

A narrative synthesis was performed using a framework based on the broad categories of study type identified in the review, as has been done previously.[3] Five categories of study were determined based on the overall aim of each study and the purpose of NLP algorithm described within it: 1) studies that used NLP to predict a clinical outcome, 2) studies that presented an algorithm capable of classifying a specific clinical concept, 3) studies which used NLP to identify a patient cohort or other clinical concept as part of a broader research aim, 4) studies which used NLP to develop an algorithm intended for use in clinical surveillance or 5) studies which evaluated various technical aspects of NLP such as preprocessing techniques. Each study was also evaluated to determine the described algorithms' level of readiness for implementation in a clinical setting.[2] The four levels of readiness were defined as 1) development and internal validation only, 2) external

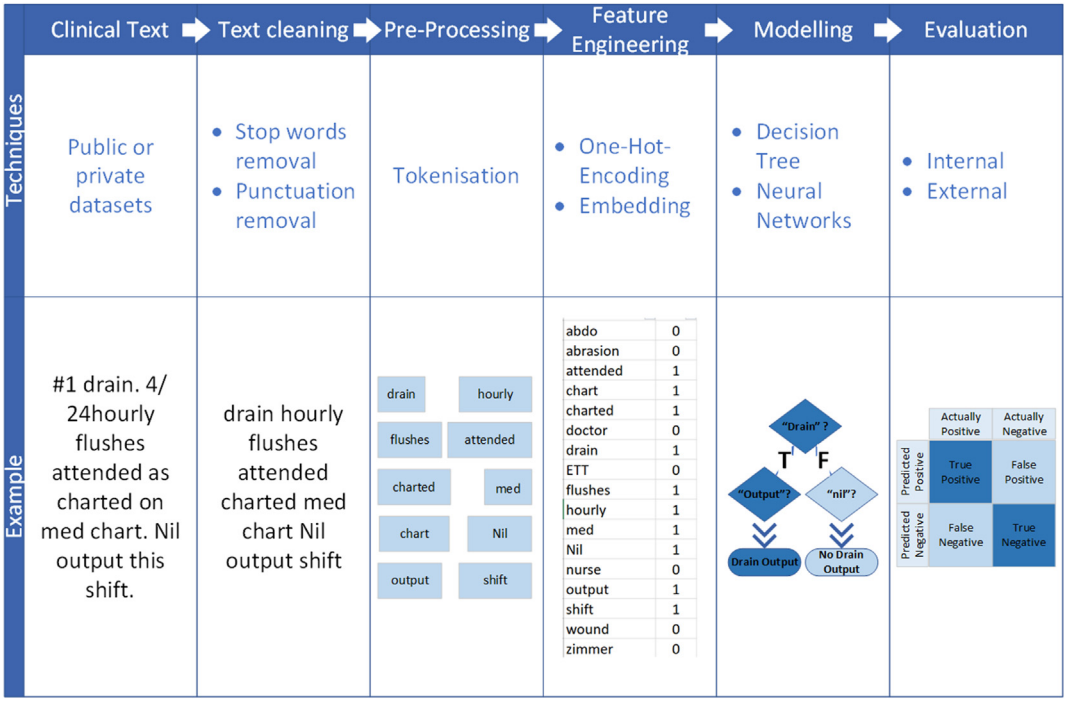


**Fig. 1.** Overview of a typical NLP algorithm development pipeline.

validation on a separate dataset from either a different location or time period, 3) prospective testing of the algorithm, 4) implementation and evaluation of clinical utility or clinical impact on patient outcomes.

## 3. Results

A total of 368 studies were retrieved in the initial search, 213 of which were excluded as they did not describe the use of NLP techniques or use data from an ICU. The full text of the remaining 155 articles was then assessed, with 87 determined to be eligible for inclusion in the review. The study screening and selection process is presented in Fig. 2. A glossary of commonly used NLP terms is presented in Table 1.

### 3.1. Study characteristics

Of the 87 studies included in the review, the majority used a version of the Medical Information Mart for Intensive Care (MIMIC) database (67.8%, n = 59) and only 17.2% (n = 15) reported extracting data from more than one hospital to create their dataset.[9] Sample sizes varied widely, with only nine participants in the smallest study and 101,196 in the largest. Types of performance metrics used also varied, with 18.4% (n-16) of studies not reporting any. Almost one-third of studies were published in journals with a subject area of 'Computer Science' (31.0%, n = 27) and despite all studies using data generated in an ICU, only 9 (10.3%) were published in a journal with the subject category 'Critical Care and Intensive Care Medicine'. The number of included studies published each year increased over the 10-year search period, with only 4 (4.6%) studies published in 2014 and 20 (23.0%) in 2022. Full characteristics of the 87 included studies are detailed in the Supplementary Appendix.

### 3.2. Study categories- prediction of a clinical outcome

Studies that aimed to predict a clinical outcome were the most common type of study included in the review (27.6%, n = 24), with the majority of these studies predicting mortality (66.7%, n = 16). Other clinical outcomes included development of acute kidney injury (16.7%, n = 4), sepsis or septic shock (n = 2) and post-operative acute respiratory failure (n = 1). Most studies used all clinical notes available in the study dataset and combined features derived from both structured and unstructured data to develop the predictive algorithm. Two studies used NLP to generate sentiment scores from nursing notes to use as a feature in their mortality prediction algorithms. Sentiment scores are a measure of how positive or negative the language in a particular text is, and both studies found that these scores were predictive of mortality.[10,11] Only two studies described the external validation of a predictive algorithm, and none had performed prospective testing or discussed potential strategies for implementing their algorithm into a clinical setting.[12,13] Both studies which described external validation of their algorithm reported strong performance with area under the receiver operating curve (AUC-ROC) measures of 0.92 to 0.83.

### 3.3. Study categories- clinical concept classification

Next most common (26.4%, n = 23) were those studies which aimed to develop an algorithm capable of extracting information about a particular clinical concept or condition from the progress notes. Several of these studies focussed on topics that are not captured in structured data fields such as interactions between patient family members and healthcare providers and social determinants of health. Several other studies focussed on radiology reports with the aim of automatically extracting information such

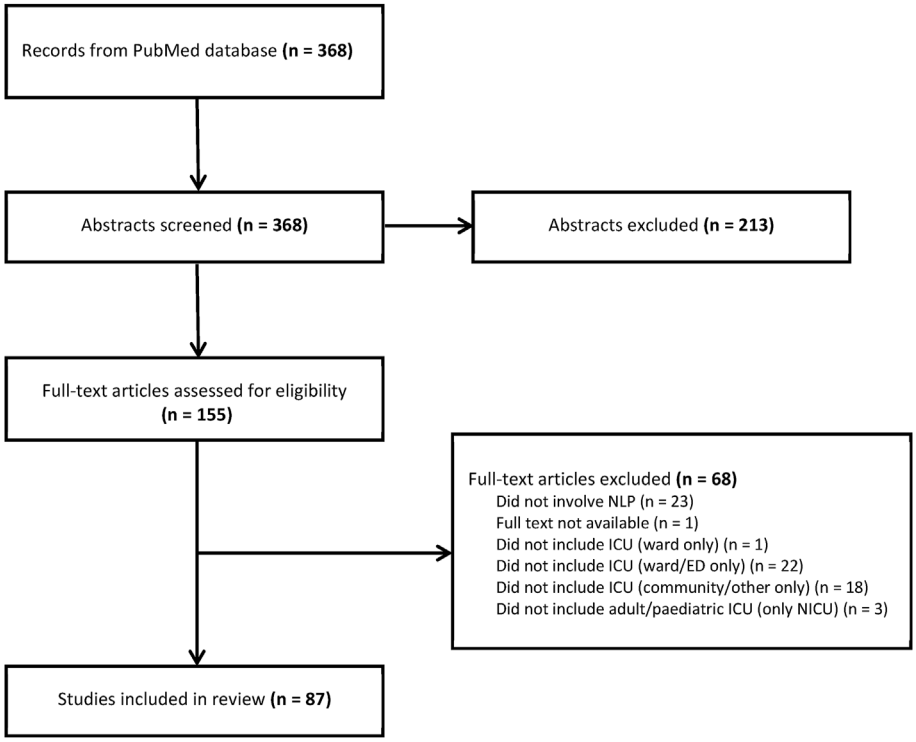


**Fig. 2.** PRSMA flow diagram of the study selection process.

**Table 1**
Glossary of terms relating to NLP.

| Term | Definition |
|---|---|
| Open-source software | Software for which the source code has been made openly available; typically no cost associated with its use unlike proprietary software |
| MIMIC | Medical Information Mart for Intensive Care (MIMIC) is an openly available dataset derived from patients admitted to the ICU of the Beth Israel Deaconess Medical Center in Boston, Massachusetts. |
| Manual text annotation | Process by which humans label various concepts and relationships present in a text, typically to enable the training or validation of an algorithm. |
| Large language model | These models are typically built on a transformer model architecture and pre-trained on large text datasets to allow human-like understanding and production of text. |
| Rule-based algorithm | Algorithms which depend on a series of pre-specified 'if then' rules to determine their output. Commonly used in combination with regular expressions. |
| Regular expressions | A series of characters used to find and match patterns of text. |
| Machine learning algorithm | Algorithms which learn from patterns in a dataset to determine their output. Examples include decision trees, random forests, support-vector machines and artificial neural networks. |
| Deep learning | A type of machine learning which uses multiple layers to process the input, typically as part of a neural network. |
| Structured data | Data which exists in a fixed format, such as numerical or categorical data. As opposed to unstructured data which can include free text, audio or image data. |
| Information extraction | A type of NLP task which involves obtaining structured data from unstructured text. May include identifying specific concepts, events or relationships in the text. |
| Named entity recognition | A type of information extraction that focusses on obtaining data about a specific entity such as a disease or medication |
| Embedding | Also known as vectorisation, this is process by which text-based data are converted to numerical data prior to being used as the input to an algorithm or model. A range of embedding methods have been developed, including embedding at the word and sentence level. |
| Tokenisation | The process of breaking up a piece of text into smaller units known as tokens. Typically performed at the word level so each token is a single word. |
| Internal validation | Internal validation tests how well a model or algorithm performs on the data from which it was derived. This is commonly done by splitting the available data into a training and testing set. |
| External validation | External validation is when a model is tested on an independently derived dataset that was not available at the time the model was being developed. Tests how well the model generalises to other datasets. |
| Pre-processing | Text pre-processing is the series of tasks performed prior to analysis. These may include converting text to lower case, removing digits, tokenisation, removing stopwords, stemming and lemmatisation. |
| Stemming | Part of text pre-processing which removes prefixes and suffixes to retain the core meaning of the word. |
| Stopwords | Words which do not contribute to the meaning of a text such as 'the' or 'and'. |
| Transformer model | A type of deep learning model which use a mechanism known as attention combined with a feed-forward neural network to increase their performance. Examples of transformer models include the various types of BERT (Bidirectional Encoder Representations from Transformers) models. |

as the location of a central venous catheter or the presence of acute respiratory distress syndrome in chest X-ray report. These algorithms could then be used to alert clinicians to any unexpected findings, such as an incorrectly positioned central line, thereby enabling early intervention to occur.[14] Similar to the prediction study category, only three studies in this category described external validation.[15–17] The three studies reported varying success in the generalisation of their algorithms to these external datasets with reported precision ranging from 0.77 to 0.98 and recall from 0.42 to 0.99.

### 3.4. Study categories- study cohort identification

Twenty studies (23.0%) used an NLP algorithm to identify a clinical concept or patient cohort as part of a broader observational research study. Algorithms in this group generally extracted information about concepts that are not recorded elsewhere in the medical record and would otherwise have required a manual chart review to obtain, such as patient mobilisation within the ICU, financial constraints and considerations and family involvement in care planning. Two studies used meta-data associated with clinical text to measure the amount of redundancy in clinical documentation and to investigate whether NLP can be used to provide an objective evaluation of intensive care medical trainees' oral case presentations.[18,19] Unlike the previous two study categories, reporting of the various algorithms' performance metrics was relatively uncommon, with only half (n = 10) reporting any performance metrics.

### 3.5. Study categories- technical NLP evaluation

The fourth most common type of study (17.2%, n = 15) was that which focussed on the technical aspects of NLP algorithm development. These studies generally aimed to identify which type of algorithm or text pre-processing technique demonstrated the best performance for a nominated task and were all published in computer science or health informatics journals. The tasks which the various algorithms performed in this study category tended to be broader than those in previous categories, such as creating a daily summary of clinical problems or identifying the section headings within a clinical progress note. Almost all of the studies in this category (80.0%, n = 12) used a version of the MIMIC database or another publicly available data source such as those made available through the n2c2 data challenge.[9,20]

### 3.6. Study categories- clinical surveillance

Clinical surveillance studies were the least common type of study identified (5.7%, n = 5). These studies developed algorithms that were intended to provide real-time alerts to clinicians about hospital-acquired complications such as healthcare-associated infections and adverse drug events. Two studies in this category described the prospective evaluation of their algorithm. The first described the implementation of an algorithm to identify indwelling catheter urinary tract infections.[21] The algorithm demonstrated poor sensitivity (33%) which was attributed to the design of the clinical information system which stored vital sign data separately to progress note data and therefore important variables, such as patient temperature, were not accessible to the algorithm. The second study described an algorithm that developed a phenotypic profile of critically ill children to facilitate the interpretation of whole genome sequencing results.[22] The algorithm was tested prospectively on seven critically ill children and was found to reduce the median time to diagnosis by an estimated 28 h.

**Table 2**
List of open source NLP resources.

| Name | Description | Resource Location |
|---|---|---|
| *Software for NLP Tasks* | | |
| tm | Text mining toolkit (R) | https://CRAN.R-project.org/package=tm |
| Stanford NLP Group | Range of NLP tools | https://nlp.stanford.edu/software/ |
| Gensim | Topic modelling and other NLP tasks (Python) | https://pypi.org/project/gensim/ |
| Natural Language Toolkit | Suite of tools for performing NLP (Python) | https://www.nltk.org/ |
| text2vec | Suite of tools for performing NLP (R) | https://text2vec.org/index.html |
| fastText | Library for generating word embeddings and classifying text (Python) | https://fasttext.cc/ |
| NeuroNER | Program which performs named entity recognition (Python) | https://github.com/Franck-Dernoncourt/NeuroNER |
| NimbleMiner | Text mining system (R) | http://github.com/mtopaz/NimbleMiner |
| spaCy | Framework for performing NLP (Python) | https://spacy.io/ |
| MedSpaCy | Clinical library for use with spaCy | https://github.com/medspacy/medspacy |
| re | Library for performing regular expressions (Python) | https://docs.python.org/3/library/re.html |
| regex | | https://pypi.org/project/regex/ |
| sense2vec | Tool used to generate word vectors (Python) | https://github.com/explosion/sense2vec |
| TextBlob | Suite of tools for performing NLP (Python) | https://pypi.org/project/textblob/0.9.0/ |
| Transformers | Toolkits of pretrained models for a range of NLP tasks (Python) | https://pypi.org/project/transformers/ |
| Simple Transformers | | https://pypi.org/project/simpletransformers/ |
| SemEHR | Open source clinical information extraction tool | https://github.com/CogStack/CogStack-SemEHR |
| MedCAT | | https://github.com/CogStack/MedCAT |
| quanteda | Package for quantitative text analysis (R) | https://quanteda.io/ |
| Tokenizers | Tokenization package (R and Python) | https://CRAN.R-project.org/package=tokenizers<br>https://pypi.org/project/tokenizers/ |
| *Software for Text Annotation* | | |
| brat rapid annotation tool | Open source text annotation tool | https://brat.nlplab.org/index.html |
| ClinicalRegex | Open source text annotation program | https://lindvalllab.github.io/clinical-regex/ |
| *Software for Machine Learning Tasks* | | |
| scikit-learn | Machine learning library for Python | https://pypi.org/project/scikit-learn/ |
| TensorFlow | Open source machine learning platform | https://www.tensorflow.org/ |
| Keras | Python library for developing deep learning models | https://keras.io/ |
| XGBoost | Library for performing gradient boosting algorithms (R and Python) | https://github.com/dmlc/xgboost |
| PyTorch | Framework for building deep learning models in Python | https://pytorch.org/ |

### 3.7. Types of algorithms and analysis tools

The majority of included studies (74.7%, n = 65) reported using a type of machine learning, with deep learning models being the most popular (n = 40). Pre-trained transformer models were also used in 15 studies, most commonly a version of the Bidirectional Encoder Representations from Transformers (BERT) model. The original BERT model was trained on the BookCorpus and English Wikipedia; however, domain-specific versions such as BioBERT and ClinicalBERT have since been made available.[23–25] These versions have further trained the original BERT model on PubMed abstracts and full-text articles available on PubMed Central and on the progress notes and other free text available in the MIMIC database. Rules-based algorithms were used by 44.8% (n = 39) of included studies. These algorithms generally involved searching for a group of keywords related to the specified phenomenon of interest to determine whether it was present or not. This was often paired with techniques such as regular expressions to account for variations in how words were written such as spelling mistakes or differences due to tense or plurality. Python was the most commonly used analysis platform (n = 46), followed by R (n = 18). Several proprietary software tools were also identified. However, the majority of studies used an open-source platform. A list of tools and other NLP resources identified in the review is presented is Table 2.

## 4. Discussion

This scoping review has provided an overview of the wide variety of studies which have used NLP techniques on data generated in an ICU context. The most common use of NLP was as part of a prediction study, with NLP-derived features generally used in combination with structured data to develop a predictive algorithm. An almost equally large number of studies described NLP algorithms capable of extracting information about a clinical concept or used NLP to identify a specific patient cohort as part of a broader study. Both machine learning and rules-based methods were described in the included studies, as well as deep learning models. The vast majority of studies were written for an audience with a computer science or health informatics background, with relatively few published in journals intended for intensive care clinicians. Very few studies externally validated their algorithm and none described the evaluation of an algorithm which had been implemented in a clinical setting. Finally, more than two-thirds of studies utilised the MIMIC database which is derived from a single centre in the United States. While this was presumably due to the open-access nature of that database, it highlights the need for future studies to be conducted on more diverse datasets. Furthermore, the dominance of this specific patient cohort should be considered when conducting future reviews in this area, particularly those which include a quantitative synthesis.

The generally low level of readiness for clinical implementation found in the studies included in this review is not unique to NLP algorithms developed in the critical care context. Similar reviews which sought to provide an overview of NLP in other clinical specialties such as cardiology and diabetes care also found that studies in these areas predominantly described algorithm development with very few evaluating their clinical implementation.[4,26] This lack of research on clinical implementation was also described in a systematic review into the use of all types of AI in the ICU and has been suggested as a key reason for the slow uptake of AI in the Australian healthcare system more broadly.[2,27] While a variety of studies in this review have demonstrated that it is feasible to use NLP on clinical data, there is a clear need for more studies that are able to demonstrate both the safety and efficacy of NLP, and other AI-based algorithms, when implemented in clinical settings.

Algorithms which utilise NLP have the potential to benefit both patients and staff if they can eventually be implemented in a clinical setting. For example, many surveillance systems currently either rely on clinical coding data or place an additional burden on clinical staff. Automating these processes based on progress note documentation would enable real-time surveillance data to be obtained whilst simultaneously reducing the administrative burden for clinical staff. The performance of clinical risk-prediction models has also been shown to improve with the addition of NLP-derived features, potentially by enabling access to many important clinical concepts which are not well captured in structured data sources.[28] Previous studies have highlighted inaccuracies in clinical concepts such as mental health disorders, drug and alcohol use and other social determinants of health, which limits the quality of research conducted about these vulnerable populations.[29–31] By utilizing NLP, critical care researchers can incorporate more accurate data about these understudied patient groups without having to spend the human resources required to obtain the data manually.

### 4.1. Limitations

Due to the heterogenous nature of the articles identified, a quantitative synthesis was unable to be performed. This heterogeneity was not unexpected given the broad nature of the review question and was why a scoping review methodology was utilised. Quantitative comparisons of the performance metrics reported in the various studies could also not be made due to the range of different NLP methods, outcome measures and evaluation strategies used, and this should be considered for future reviews with more specific questions and aims. Resource limitations and the variety of study designs also meant a formal critical appraisal of the methodological quality of the included studies was not performed. However, the 'level of readiness' assessment was performed instead to flag to clinicians which algorithms may be appropriate to consider for clinical implementation. Finally, given that included studies were limited to those published within the last 10 years and written in English, this review cannot draw conclusions regarding older or non-English applications of NLP in the ICU.

## 5. Conclusion

This scoping review has provided an overview of the variety of ways in which NLP techniques have been used on data generated in an ICU context. Relatively few studies were intended for a clinical audience or described the clinical impact of their findings. Future research in this area should ensure clinicians are involved in the development of algorithms intended for clinical use and include implementation and evaluation of their algorithms' clinical impact.

## CRediT author statement

**Julia K Pilowsky:** Conceptualization, Methodology, Investigation, Writing- Original Draft, **Jae-Won Choi:** Investigation, Visualisation, Writing- review & editing, **Aldo Saavedra:** Investigation, Writing- review & editing, **Maysaa Daher:** Investigation, Writing- review & editing, **Linda Williams:** Conceptualization, Writing- review & editing, **Nhi Nguyen:** Conceptualization, Writing- review & editing, **Sarah L Jones:** Investigation, Project administration, Writing- review & editing.

## Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Nil.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ccrj.2024.06.008.

## References

[1] Kong H-J. Managing unstructured big data in healthcare system. Healthc Inform Res 2019;25(1):1–2. https://doi.org/10.4258/hir.2019.25.1.1.
[2] van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021;47:750–60.
[3] Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language processing in radiology: a systematic review. Radiology 2016;279(2):329–43. https://doi.org/10.1148/radiol.16142770.
[4] Turchioe MR, Volodarskiy A, Pathak J, Wright DN, Tcheng JE, Slotwiner D. Systematic review of current natural language processing methods and applications in cardiology. Heart 2022;108(12):909–16.
[5] Le Glaz A, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, et al. Machine learning and Natural Language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708. https://doi.org/10.2196/15708.
[6] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467–73.
[7] Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19–32.
[8] SCImago. Sjr — SCImago journal & country rank. http://www.scimagojr.com; 2023.
[9] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Epub 20160524 Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35. PubMed PMID: 27219127; PubMed Central PMCID: PMC4878278.
[10] Waudby-Smith IER, Nam T, Dubin JA, Lee J. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. PLoS One 2018;13(6):e0198687. https://doi.org/10.1371/journal.pone.0198687.
[11] Liu Z, Yang Y, Song H, Luo J. A prediction model with measured sentiment scores for the risk of in-hospital mortality in acute pancreatitis: a retrospective cohort study. Ann Transl Med 2022;10(12):676. https://doi.org/10.21037/atm-22-1613.
[12] Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of prediction models for critical care outcomes using Natural Language processing of electronic health record data. JAMA Netw Open 2018;1(8):e185097. https://doi.org/10.1001/jamanetworkopen.2018.5097. e.
[13] Mahendra M, Luo Y, Mills H, Schenk G, Butte AJ, Dudley RA. Impact of different approaches to preparing notes for analysis with natural language processing on the performance of prediction models in intensive care. Critical care explorations 2021;3(6).
[14] Shah M, Shu D, Prasath VBS, Ni Y, Schapiro AH, Dufendach KR. Machine learning for detection of correct peripherally inserted central catheter tip position from radiology reports in infants. Appl Clin Inf 2021;12(4):856–63.
[15] Mayampurath A, Churpek MM, Su X, Shah S, Munroe E, Patel B, et al. External validation of an acute respiratory distress syndrome prediction model using radiology reports. Crit Care Med 2020;48(9):e791.
[16] Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: a transferable clinical natural language processing model for electronic health records. Artif Intell Med 2021;118:102086.
[17] Miller MI, Orfanoudaki A, Cronin M, Saglam H, So Yeon Kim I, Balogun O, et al. Natural language processing of radiology reports to detect complications of ischemic stroke. Neurocritical Care 2022;37(Suppl 2):291–302.
[18] Searle T, Ibrahim Z, Teo J, Dobson R. Estimating redundancy in clinical text. J Biomed Inf 2021;124:103938.
[19] King AJ, Kahn JM, Brant EB, Cooper GF, Mowery DL. Initial development of an automated platform for assessing trainee performance on case presentations. ATS scholar 2022;3(4):548–60.
[20] National NLP clinical challenges (n2c2) [cited 2024 07/02/2024]. Available from: https://n2c2.dbmi.hms.harvard.edu/.
[21] Branch-Elliman W, Strymish J, Kudesia V, Rosen AK, Gupta K. Natural Language processing for real-time catheter-associated urinary tract infection surveillance: results of a pilot implementation trial. Infect Control Hosp Epidemiol 2015;36(9):1004–10. https://doi.org/10.1017/ice.2015.122. Epub 2015/05/26.
[22] Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med 2019;11(489):eaat6177.

[23] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
[24] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.
[25] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019. arXiv preprint arXiv:190403323.
[26] Turchin A, Florez Builes LF. Using Natural Language processing to measure and improve quality of diabetes care: a systematic review. J Diabetes Sci Technol 2021;15(3):553–60. https://doi.org/10.1177/19322968211000831. PubMed PMID: 33736486.
[27] van der Vegt A, Campbell V, Zuccon G. Why clinical artificial intelligence is (almost) non-existent in Australian hospitals and how to fix it. Med J Aust 2023;220(4):172–5. https://doi.org/10.5694/mja2.52195.
[28] Vagliano I, Dormosh N, Rios M, Luik TT, Buonocore TM, Elbers PWG, et al. Prognostic models of in-hospital mortality of intensive care patients using neural representation of unstructured text: a systematic review and critical appraisal. J Biomed Inf 2023;146:104504. https://doi.org/10.1016/j.jbi.2023.104504.
[29] Nguyen TQ, Simpson PM, Braaf SC, Cameron PA, Judson R, Gabbe BJ. Level of agreement between medical record and ICD-10-AM coding of mental health, alcohol and drug conditions in trauma patients. Health Inf Manag J 2019;48(3):127–34.
[30] Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. Medicine 2020;99(52).
[31] Pilowsky JK, Elliott R, Roche MA. Association between preexisting mental health disorders and adverse outcomes in adult intensive care patients: a data linkage study. Crit Care Med 2023;51(4):513–24.