Use of unstructured text in prognostic clinical prediction models: a systematic review

Tom M. Seinen[1], Egill Fridgeirsson[1], Solomon Ioannou[1], Daniel Jeannetot[1], Luis H. John[1], Jan A. Kors[1], Aniek F. Markus[1], Victor Pera[1], Alexandros Rekkas[1], Ross D. Williams[1], Cynthia Yang[1], Erik van Mulligen[1], Peter R. Rijnbeek[1]

Corresponding author:
Tom M. Seinen (t.seinen@erasmusmc.nl), Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands.

Author affiliations:
[1] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Keywords:
Clinical prediction model, Prognostic prediction, Natural language processing, Machine learning

Word count:
4016

ABSTRACT

Objective

This systematic review aims to assess how information from unstructured clinical text is used to develop and validate prognostic risk prediction models. We summarize the prediction problems and methodological landscape and assess whether using unstructured clinical text data in addition to more commonly used structured data improves the prediction performance.

Materials and Methods

We searched Embase, MEDLINE, Web of Science, and Google Scholar to identify studies that developed prognostic risk prediction models using unstructured clinical text data published in the period from January 2005 to March 2021. Data items were extracted, analyzed, and a meta-analysis of the model performance was carried out to assess the added value of text to structured-data models.

Results

We identified 126 studies that described 145 clinical prediction problems. Combining text and structured data improved model performance, compared to using only text or only structured data. In these studies, a wide variety of dense and sparse numeric text representations were combined with both deep learning and more traditional machine learning methods. External validation, public availability, and explainability of the developed models was limited.

Conclusion

Overall, the use of unstructured clinical text data in the development of prognostic prediction models has been found beneficial in addition to structured data in most studies. The EHR text data is a source of valuable information for prediction model development and should not be neglected. We suggest a future focus on explainability and external validation of the developed models, promoting robust and trustworthy prediction models in clinical practice.

INTRODUCTION

Prognostic prediction models are increasingly common in clinical research and practice [1,2]. A prognostic model predicts which patients, among a target population of patients, will experience some clinical outcome during a prediction horizon, using predictors measured during an observation window prior to the time of prediction. The growing availability of observational data in electronic health records (EHRs) forms a rich source to develop prediction models in a data-driven manner [2,3]. Although most clinical risk prediction research is centered on the use of structured EHR data, such as coded conditions, measurements, and drug prescriptions, the majority of information in EHRs is typically stored in vast quantities of unstructured text, for example, nursing notes, discharge letters, or radiology reports [4]. Compared to structured data, clinical free text lacks an organized structure or terminology, is large in terms of file size, and contains patient-sensitive information, which complicates its use for the construction of prediction models. However, information captured in text can be more detailed and extensive than in the structured data, as it is not limited to specific code systems or input fields. Therefore, incorporating this information could potentially improve prediction model performance.

The growing availability of unstructured text in EHR data, increased computational power, and progress in natural language processing (NLP) techniques are now enabling the use of text data for the development of prediction models. Textual data has already been used in the clinical domain for a variety of tasks. Several reviews have elucidated the adoption of clinical NLP, focusing primarily on the general task of extracting information from clinical notes [5-11] or diagnostic classification of patients, including case detection, patient identification, and phenotyping [4,9,12]. A recent systematic review by Yan et al. [13] studied the use of clinical text in early sepsis prediction. However, to our knowledge, no broad systematic review has been conducted on the development of prognostic prediction models incorporating unstructured clinical text. As text-based prognostic prediction models start to be developed, it becomes increasingly important to reflect on the work that has been done, summarize the methodological landscape, and discover whether text data has value supplementing coded data.

Consequently, the objective of this review is to assess how information extracted from unstructured EHR text data is utilized to develop and validate prognostic prediction models. We evaluated the studies on the study settings and populations, text processing methods and representations, machine-learning methods and feature sets combinations, performance evaluation and external validation, and model explainability and availability. Furthermore, we determined the value of text in addition to structured data by comparing the performance between models using these different feature sets within the studies.

MATERIALS AND METHODS

Review protocol

The review protocol for this study was registered on June 17, 2021, and is publicly available at the Open Science Framework Registries (https://osf.io/gw628).

Eligibility criteria

This review targeted studies from the last 15 years (January 2005 to March 2021) describing the development and evaluation of prognostic clinical prediction models that incorporate information extracted in a data-driven manner from unstructured EHR text data. The three inclusion criteria are defined as follows. (1) The study described the development and evaluation of a prognostic clinical prediction model. (2) The model predictors were based on information extracted from unstructured EHR text data. (3) Information was automatically extracted from the unstructured text in a data-driven manner. Data-driven implies that the extraction of information from the text was exploratory and not restricted to features that were expected to be important. An extensive description of the inclusion criteria is provided in the supplementary material.

Literature search

Four databases were used for the literature search: Embase, MEDLINE, the Web of Science core collection, and Google Scholar. The database choice and the search strategy creation was aided by a medical librarian. The search strategy consisted of four clauses that incrementally limited the search results: (1) Prediction models; (2) The medical domain or electronic health records; (3) A notion of text data, clinical notes, or NLP methods; (4) The period from January 2005 till March 2021, studies in the English language, and excluding conference abstracts and animal research. The full search strategy can be found in Table S1.

Screening

All studies found in the literature search were first screened for fulfilling the eligibility criteria based on the title and abstract. Those that were found relevant underwent a second screening for inclusion based on the full text. In both screening phases, one reviewer (TS) screened all studies and ten other reviewers (EF, SI, DJ, LJ, JK, AM, VP, AR, RW, CY) independently screened one-tenth of the total number of studies. This resulted in each study being screened by two independent reviewers. Any discrepancies between them, in both screening phases, were resolved in a consensus meeting.

Data extraction and synthesis

Data were extracted from the included studies by one reviewer (TS) using a pre-defined set of data items. The set is outlined in Table 1, including the type of input. Some items are based on clinical prediction item sets from the *critical appraisal and data extraction for systematic reviews of prediction modelling studies* (CHARMS) checklist [14] and the *transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD) statement [15]. Ten data item topics were distinguished: (1) the general publication information, (2) the study setting, (3) the study population, (4) unstructured text data predictors, (5) structured data predictors, (6) machine-learning methods and feature sets, (7) the internal and (8) external validation, (9) model explainability, and (10) model availability.

The input for text representation methods consisted of a list of both sparse and dense numeric vector representations. Sparse representations included Bag-of-Words (*BoW*), a Bag-of-Words variant: Term Frequency – Inverse Document Frequency (*TFIDF*), and clinical concept extraction (*CE*). Dense vector representations included topic models (*TM*), word and document embeddings (*WE*, *DE*), and summarizing scores (*SS*), such as a sentiment score. Combinations of representations were

possible. The machine learning methods were of varying complexity and interpretability, including methods ranging from simple linear or logistic regression (*LinR*, *LogR*), random forests or other tree-based methods (*RFTB*), gradient boosting (*GB*), and Support Vector Machines (*SVM*), to complex deep neural networks (*DNN*).

If a study reported on multiple prediction problems, the data items were extracted for each reported problem. The model and validation data items were only extracted for the – self-reported – best performing structured-data, text-data, and combined-data models in each problem. For data items with free text input, the results were manually categorized after data extraction to enable analysis. We performed a meta-analysis on the reported model performance, comparing the structured-data, text-data, and combined-data models, for each prediction problem. The differences in the area under the receiver operator curve (AUC) metric were calculated for each reported feature-set comparison: text and structured data ($\Delta AUC_{TS} = AUC_T - AUC_S$), combined and structured data ($\Delta AUC_{CS} = AUC_C - AUC_S$), and combined and text data ($\Delta AUC_{CT} = AUC_C - AUC_T$). The AUC differences indicate the relative performance difference between the use of the three feature sets within each prediction problem and are suitable to be compared across studies.

Table 1. List of data items for data extraction, by topic. Data item sources indicated by A: CHARMS and B: TRIPOD; an asterisk (*) indicates data items added to the review protocol.

Item topic	Data Item	Input type
1. General information	Publication year	Year
	Journal	Free text
2. Study setting	Dataset	Free text
	Country of data	Country
	Clinical setting	Free text
	Study dates	Range of years
3. Population	Type of study*	Cohort, Case-control
	Target population[AB]	Free text
	Prediction outcome[A]	Free text
	Prediction time horizon[A]	Hours, Days, Years, Relative time, Timepoint
	Prediction outcome type	Binary, Multiclass, Continues
4. Unstructured text data predictors	Type of text content	Free text
	Language of text	Language
	Observation window[B]	Hours, Days, Years, Relative time, Timepoint
	Pre-processing methods	Free text
	Text representation methods	Bag-of-Words (*BoW*), TFIDF, Concept Extraction (*CE*), Word Embedding (*WE*), Document Embedding (*DE*), Topic model (*TM*), Summarizing Score (*SS*). (multiple possible)
	Used ontologies/vocabularies	Free text
	Used software/program/package	Free text
	Number of predictors[A]	Number
5. Structured data predictors	Types of structured data	Free text
	Observation window[B]	Hours/Days/Years/Relative time/Timepoint
	Number of predictors[A]	Number
6. Model	Machine-learning method[A]	Logistic regression (*LogR*), linear regression (*LinR*), cox proportional hazards regression (*Cox*), Naïve Bayes (*NB*), Random forests or other tree based methods (*RFTB*), Gradient Boosting (*GB*), Support Vector Machines (*SVM*), (simple) Neural Networks (*NN*), Recurrent Neural Networks (*RNN*), Convolutional Neural Networks (*CNN*), Transformer, (other) Deep Neural Networks (*DNN*), Ensembles, Other.
	Feature-set	Structured/Text/Combined
7. Internal validation	Number of subjects/observations[A]	Number
	Number of cases[A]	Number
	AUC, AUPRC, F1-score[A]	Values
	Accuracy, sensitivity (or recall), specificity, and positive predictive value (or precision) reported?[A]	Yes/No
	MSE/MAE reported?[A]	Yes/No
	ROC/PR curves presented?[A]	Yes/No
	Calibration plot or metrics presented?[AB]	Yes/No
8. External validation	Type of external validation[A]	Same/another department/center/country
	Same items as internal validation	
9. Explainability	Global feature importance presented?*	Yes/No
	Single patient (local) feature importance presented?*	Yes/No
10. Final model availability	Is the final model directly available to apply to different data?[A]	Yes/No
	Is the study code available to reproduce the methods?*	Yes/No

RESULTS

Search and data-extraction results

The literature search, performed in March 2021, resulted in a total of 5043 studies. The PRISMA flow diagram is presented in Figure 1. After deduplication, removing 2030 studies, a set of 3013 studies was screened on title and abstract. We excluded 2783 studies that violated one of the inclusion criteria. Full-text screening of the remaining 230 studies resulted in 126 relevant studies to be included in the review. The 104 studies that were excluded based on their full text consisted of 5 duplicate studies, 52 studies not performing prognostic modelling or a performance evaluation, 13 studies with no use of text data in the prediction model, 28 studies without data-driven information extraction, and 6 studies with other reasons for exclusion (no full-text available, not peer-reviewed, reviews).

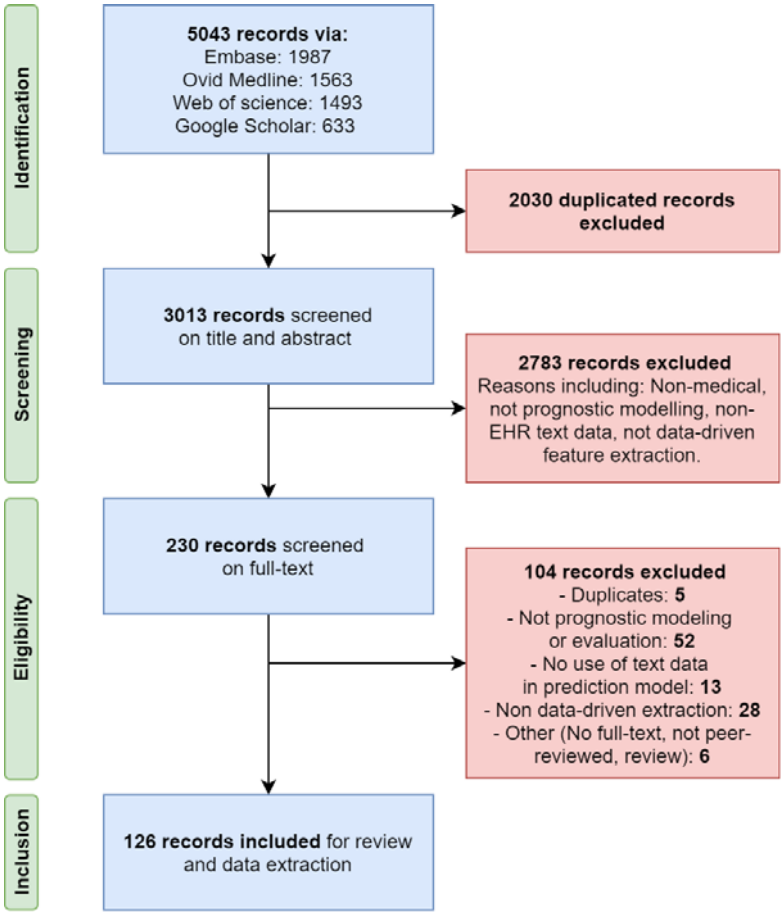

Figure 1. PRISMA flow diagram with the search and screening results of the systematic review.

We extracted the different data items for each study and its reported prediction problems. Fourteen of the 126 studies reported multiple prediction problems, resulting in a total of 145 prediction problems. A list of characteristics of the studies ordered by publication year is presented in Table S2 and the extracted data items are available as supplementary data. The large majority of the reviewed studies (79%) was published in the period from January 2018 to March 2021 (Table 2). No eligible studies were found in the period 2005 to 2011. The studies were published in a variety of journals and conference proceedings. The journals with the highest number of studies were the Journal of Biomedical Informatics (9%), PLoS ONE (6%), BMC Medical Informatics and Decision

Making (5%), JMIR Medical Informatics (5%), and the Journal of the American Medical Informatics Association (5%).

Table 2. Number of included studies by publication year.

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021 (until March)
Number of included studies	4	0	5	4	9	5	19	30	41	9

Around half of the reviewed studies compared models that used structured data (S), text data (T), or a combination of structured and text data (C). A comparison between all three feature sets (S:T:C) was reported for 23% of the prediction problems. In 27% of the problems two feature sets were compared (S:T 4%; S:C 19%; T:C 3%) and in 50% no comparison was made and the use of only one feature set was reported (T 33%; C 17%).

Clinical settings and prediction problems

Most prediction problems focused on general hospital care settings (47%), followed by intensive care (18%), emergency care (14%), surgical care (8%), and psychiatric or mental health care (7%). Only a few prediction problems (6%) were set in outpatient, or radiology settings. Almost half (47%) of the prediction problems used a local proprietary dataset, 13% used a collection of two or more local datasets, and 7% used registry, claims, or survey datasets. One third (33%) of the prediction problems was developed on a publicly available dataset, specifically, in 47 out of 48 cases a *Medical Information Mart for Intensive Care* (MIMIC-II/III) database [16,17] was used.

Figure 2 visualizes the different categories of target populations, clinical outcomes, and prediction horizons that make up each prediction problem. The three largest target populations were patients with general hospital admissions (22%), intensive care (ICU) admissions (18%), and emergency department (ED) visits (14%). The largest outcome events were mortality (29%), diagnosis of a specific disease or condition (19%), and hospital or ICU readmission (12%). Most prediction problems had as prediction horizon a period of months (30%) or the period of hospital/ICU admission (27%). There were 82 unique combinations of which 58 only occurred once. The prediction of mortality during admission in ICU patients occurred most often (7%), followed by the prediction of admission or transfer at ED discharge for patients visiting the ED (6%). The observation window prior to the time of prediction was not reported in 15% of the problems. The most-reported observation windows were the first hours of a hospital or ICU admission or ED visit (20%), the entire time of admission or the visit (15%), and during triage (10%).

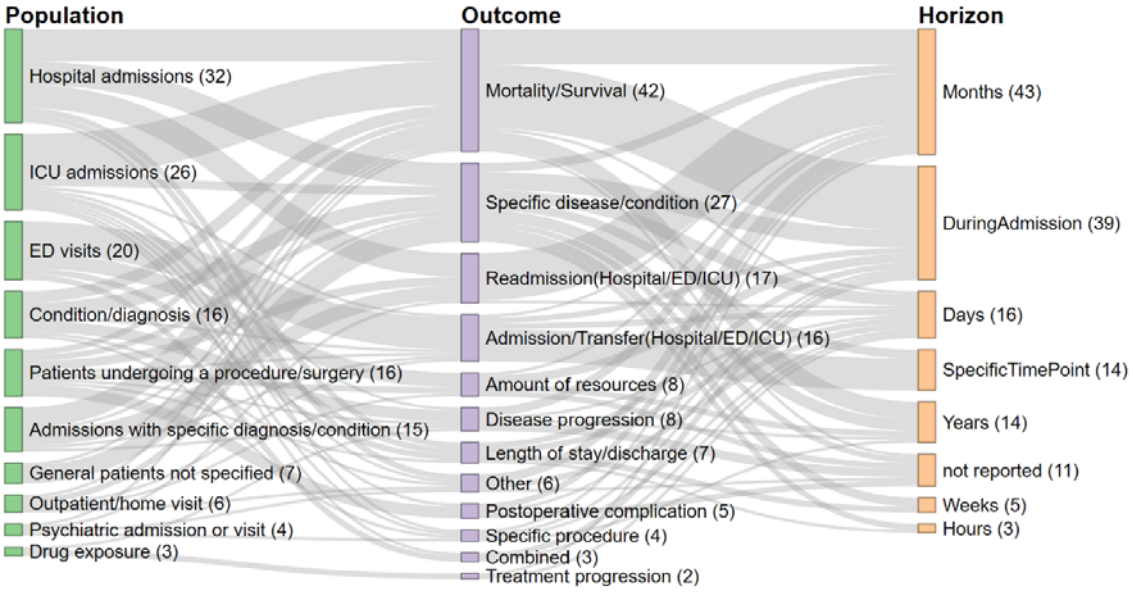

Figure 2. Sankey diagram of the different categories of target populations and clinical outcomes, and clinical outcomes and prediction horizons, ordered by size. The number in parentheses indicates the number of prediction problems with these categories and the width of the connection between two categories represents the number of prediction problems with this combination of categories.

The distribution of the number of observations and outcome cases is depicted in Figure 3A together with the distribution of their ratio in Figure 3B. The number of observations differed much between studies, from only a few hundred observations to a few million, with an mean of 87,016 and a median of 17,973 observations. Observations and outcome cases had an average ratio of 0.20 and a median of 0.14. In only 0.2% of prediction problems, the number of observations was not reported and in 16% the number of outcome cases was missing.

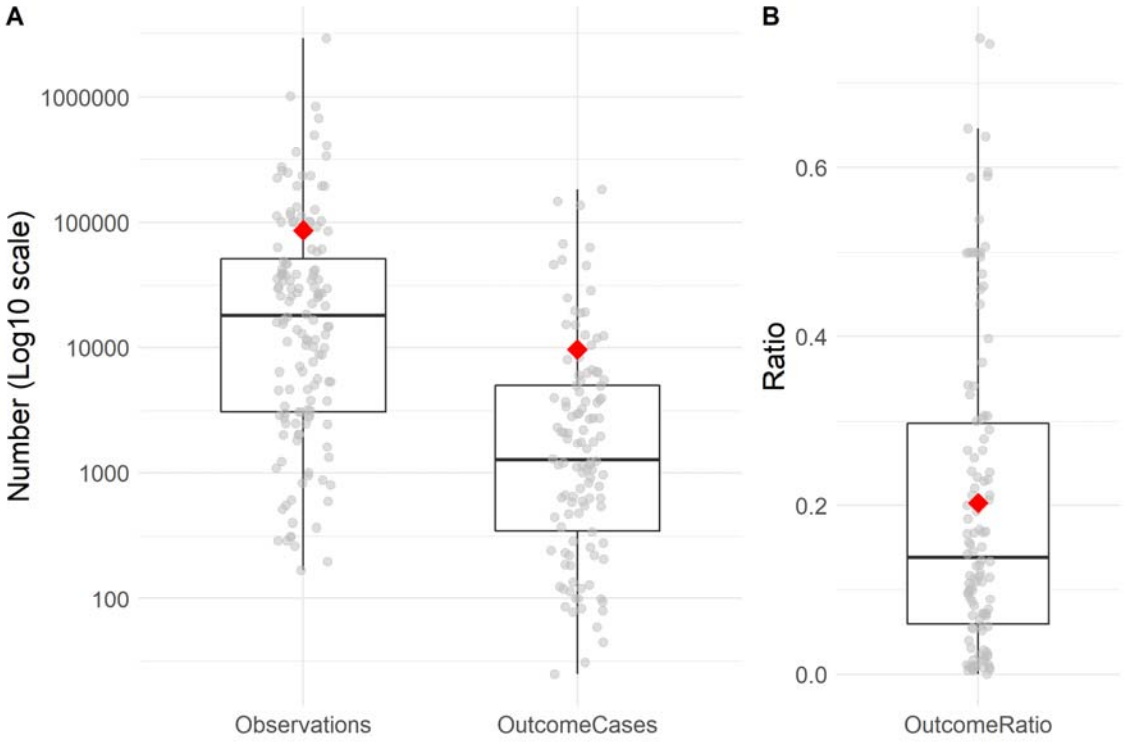

Figure 3. A: Boxplots of the number of observations (left) and outcome cases (right) of 145 prediction problems. B: Boxplot of the ratios between the number of observations and outcome cases. In both A and B the mean is indicated by the red diamond and the grey points represent the underlying data.

Pre-processing methods and text representations

Text pre-processing methods, applied before the text representation creation, were well-reported in 74% of prediction problems. The pre-processing of text commonly included methods such as sentence splitting, tokenization, lemmatization, the removal of stop words, punctuation, or numbers, abbreviation disambiguation, and the filtering of tokens based on frequencies. The text data was written in English for the majority (79%) of the prediction problems, followed by Chinese (6%) and Portuguese (3%). In total 12 different languages were reported.

Bag-of-Words text representations (including TFIDF) were used most often in 36% of the text and combined-data models, followed by word embeddings (18%) and concept extraction (14%). In some cases, multiple representations were combined by concatenation (*Combination*) (6%) or dense representation were generated from extracted concepts (*CE/WE, CE/TM*) (3%). Dense representations had a median dimension of 200 features against 6985 features in sparse representations (Figure S1). Pre-processing methods were more frequently reported together with the use of Bag-of-Words (85%) and TFIDF (92%) compared to concept extraction (54%), document embeddings (58%), and word embedding (75%) methods, which often used out-of-the-box tools or software. *MetaMap* [18] was the most common software used for extracting clinical concepts from text data, in 11 out of the 30 models using extracted concepts. Concept extraction also requires a clinical ontology or vocabulary as a reference for coded concepts. The full set or a part of the UMLS vocabularies [19] was used most, for 26 models. Five models made only use of the SNOMED Clinical Terms [20] and for another four no ontology was reported.

Machine learning methods

The most used methods for training text and combined-data models were logistic regression (27%), recurrent neural networks (13%), and random forest or other tree-based methods (10%). For the structured-data models, logistic regression was also the most prevalent method (30%), followed by gradient boosting (20%) and recurrent neural networks (12%). In the large majority (89%) of prediction problems, a binary prediction model was trained, followed by multi-class prediction (7%) and the prediction of a continuous outcome (4%).

Figure 4A depicts the use of both text representations (abstracted as dense, sparse, or combined representations) and machine-learning methods (traditional, (deep) neural networks, or ensemble methods) in both text and combined-data models over the years. It can be observed that until 2017 the use of sparse representations and traditional models was most common, but after 2018 the use of both dense text representations and (deep) neural networks took off. To understand their joined rise we examined the relationship between the model's machine-learning method and its textual input representation in Figure 4B. It shows that the dense word and document embeddings served primarily as the input for deep learning methods, while the Bag-of-Words representations were commonly used by more traditional machine learning methods. The text representations and machine-learning methods are significantly associated, $X^2(4, N=183)=36.1$, $p<.001$.

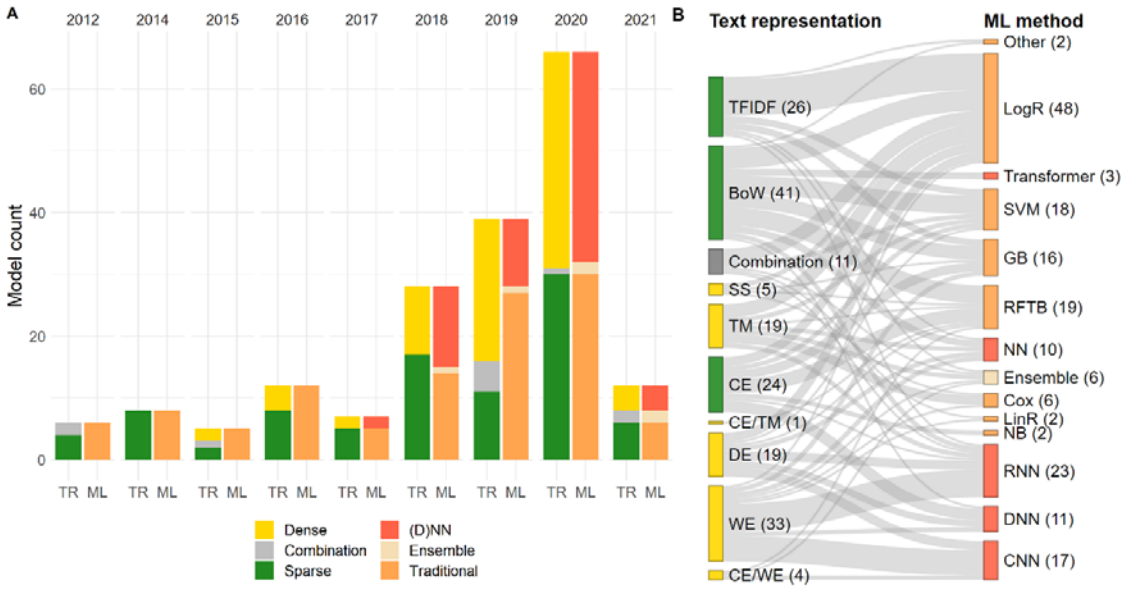

Figure 4. A: The use of different text representations (TR) and machine-learning (ML) methods in text-based or combined-data prediction models over time. No eligible studies in 2013. B: The combinations of text representations (left) and machine-learning methods (right) in text-based or combined-data prediction models. The number in parentheses indicates the number of prediction problems with these categories and the width of the connection between two categories represents the number of prediction problems with this combination of categories. Abbreviations can be found in Table 1.

Model performance evaluation and comparison

The internal validation model performance was reported using the AUC (or c-statistic/index) for the majority (83%) of prediction models. For the other models only metrics that are based on dichotomized outcomes, such as accuracy, sensitivity (or recall), specificity, and positive predictive value (or precision), were reported. The mean squared error or mean absolute error were reported

for regression models predicting continuous outcomes. The F1-score was reported for 31% of the models and the area under the precision-recall curve (AUPRC) for 14%. The combined reporting of metrics is visualized in Figure S2. A receiver operator curve or precision-recall curve was presented for 39% of problems, but for only 12% a calibration plot or calibration metric (such as the brier-score or calibration intercept and slope) was presented.

Figure 5A depicts the distributions of AUC differences (ΔAUC) between the structured-data, text-data, and combined-data models within each prediction problem. The combined-data models had a visibly higher performance than the text or structured-data models and the average AUC differences, for both ΔAUC_{CS} and ΔAUC_{CT}, were significantly larger than zero, $t(53)=6.76$, $p<.001$ and $t(33)=5.49$, $p<.001$ respectively. Text-data models did not perform significantly better or worse than the structured-data models. Their AUC difference, ΔAUC_{TS}, showed a large variation across prediction problems. We investigated whether there was a relationship between these AUC differences and the clinical settings. Figure 5B shows how the text and structured-data model performance differences vary between four clinical settings: emergency, hospital, intensive, and surgical care. Psychiatric care is not presented as it only had one observation. Following a full pairwise comparison of the distributions, we found that the AUC difference means of the intensive and surgery care prediction problems were different $t(8.55)=-3.95$, $p=.024$ (Bonferroni adjusted). This implies that models using text data in the surgical care setting had on average a higher performance (compared to structured data models) than in the intensive care setting.

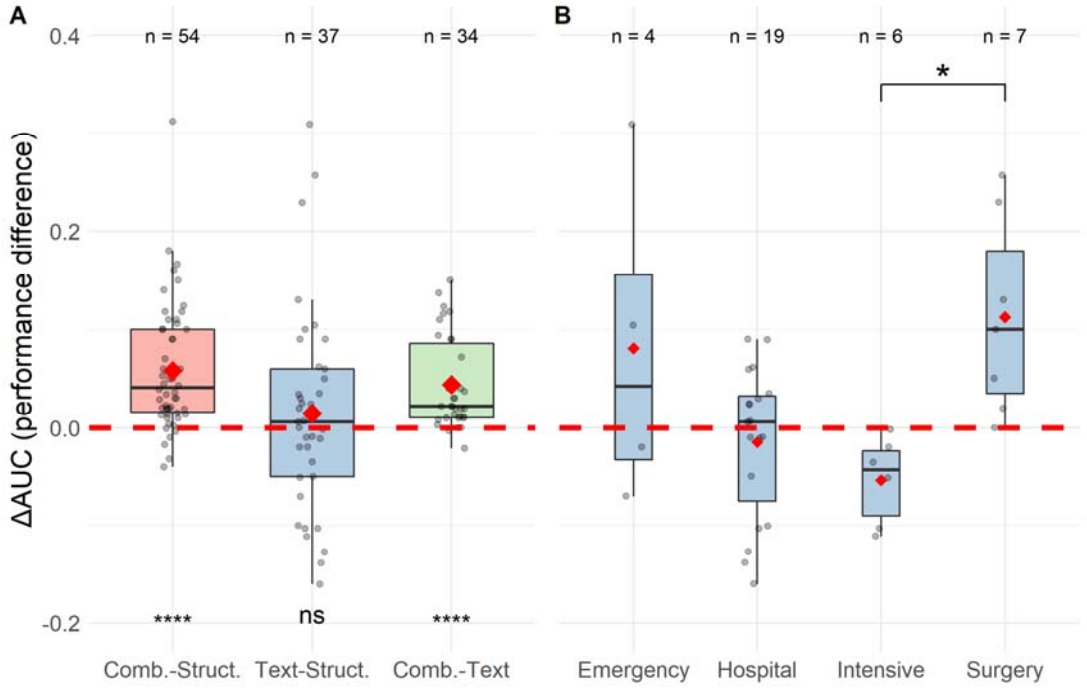

*Figure 5. A: AUC difference distribution boxplots of the combined and structured-data models (ΔAUC_{CS}), the text and structured-data models (ΔAUC_{TS}), and combined and text-data models (ΔAUC_{CT}). B: Text and structured-data model AUC difference (ΔAUC_{TS}) boxplots for four different clinical settings. In both A and B, the means are indicated by a red diamond, the grey points represent the underlying data, sample sizes are shown on top, and the red dotted line indicates the AUC difference of zero. ns=not significant, *=p<.05, ****=p<.001.*

Notably, prediction models were externally validated in only two studies. Marafino et al. [21] externally validated their in-hospital mortality prediction model within three medical centers and Menger et al. [22] externally validated their in-hospital patient violence prediction model at two sites. Both studies reported a small to moderate decrease in external validation model performance, between 0.023 and 0.079 AUC difference.

Model explainability and availability

Global feature importance, the feature importance over all predictions, was reported for 44% of the prediction problems and a local feature importance, the contribution of features for a specific prediction, was presented in 6% of problems. The final model was presented or made available in only 5% of prediction problems, while the code used for training the model was in 21% directly available online.

DISCUSSION

Model performance

We found that in the 126 studies developing prognostic clinical prediction models using unstructured EHR text data, published in the last 15 years, the performance of combined-data models on average outperformed the text and structured-data models. This demonstrates that the available EHR text data is a source of valuable information able to improve model performance in addition to structured data. Yan et al. [13] found comparable results in a review of nine studies predicting sepsis. While on average text-data models did not outperform structured-data models, we did see an interesting difference between clinical settings. In intensive care prediction problems, the structured-data models had a higher performance than the text-data models when compared to surgical care. This may be explained by the inherent differences in clinical documentation and code registration between the clinical settings. Intensive care is generally a structured-data rich setting, while in surgical care the information is contained in surgical reports. It shows that the distribution of information over text and structured data is likely dependent on various factors.

Clinical settings, datasets, and language

Hospital care (including intensive, surgical, and emergency care) was the most common clinical setting. While the combinations of different target populations, outcome events, and prediction horizons varied much between prediction problems, common themes could be observed, such as the prediction of mortality in the ICU or ED discharge disposition. Almost half of the studies used a proprietary dataset and most studies using public datasets (almost a third of the reviewed studies) relied on a MIMIC dataset [16]. Being one of the few large public clinical datasets containing anonymized unstructured data, MIMIC enables transparent and reproducible research on clinical text and serves as a benchmark for clinical NLP tasks. Almost 90% of the reviewed studies were performed on English clinical text. This suggests that opportunities still exist for studying model development using unstructured text in other languages [23,24].

Text processing and machine learning methods

The techniques used for pre-processing text and creating numeric text representations were generally well described. The impact of pre-processing methods on the model performance can be significant and those methods are therefore essential to report [25]. The sparse Bag-of-Words and TFIDF representations and the dense word and document embeddings were most frequently used and we found an association between the types of text representation and machine learning methods. The (deep) neural network methods generally used a dense text representation, while regularized logistic regression methods, random forests, or SVMs largely took sparse representations as input.

Model explainability and External validation

Global or local feature importance was presented for less than half of the prediction problems, which may be considered rather limited given the importance of model explainability and trustworthiness in clinical risk prediction [26]. Compared to simple logistic regression models, deep learning models need additional effort to be explained. Deep learning is well-suited for handling and combining structured and unstructured data [27], but the high complexity and dense input features impede direct explainability without post-hoc explanation techniques [26].

Furthermore, only two studies presented external validation results and relatively few studies shared their trained model or code. Externally validating prediction models using text data might be challenging, due to the differences in (sub)language and EHR systems or the fear of sharing identifiable patient information captured by the model. However, assessing generalizability and

external validity remains important in model development [28]. Frameworks exist, such as the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [29], that deal with the lack of syntactic and semantic interoperability in health data. The OMOP CDM enables external validation by evaluating trained prediction models on other databases, only reporting back the aggregated and anonymized results, and it allows research to meet the challenges of validating text-based models between databases using different languages [1,3].

We suggest a future research focus on model explainability and external validation, and advocate the sharing of code or trained models for external validation by others. These steps will not only expand the model's generalizability but will also promote the use of robust and trustworthy prediction models in clinical practice [28].

Strengths and limitations

There were likely some published studies eligible for inclusion that we did not find. For example, studies that incorporated text data in a prediction model but did not mention it in the title or abstract would have been missed. Nonetheless, the search query was designed to capture a wide variety of terms that would indicate the use of unstructured text data. Furthermore, the level of granularity for predefined categories for text representations or machine-learning methods was high. More granular categories on, for example, the different deep learning architectures could have been collected for more detailed and in-depth information. However, this would also have resulted in decreasing numbers per category, complicating interpretation. To the best of our knowledge, this is the first systematic review on the development of prognostic clinical prediction models using unstructured text. We performed a broad literature search over a long period of time, resulting in a large set of eligible studies in a wide variety of clinical settings, not focused on one specific prediction problem. The comparison of the relative performance between text, structured, and combined feature sets within each study allowed us to assess the value of text data in prediction model development. Finally, we made the extracted data available to provide transparency and reproducibility.

CONCLUSION

In this systematic review, we found that the use of unstructured clinical text data in the development of prognostic prediction models was beneficial in most of the studies. Combining unstructured text with structured data in prognostic prediction model development generally improved model performance, while the performance of text-data models compared to structured-data models varied. Overall, unstructured text in EHR data should not be neglected, as it is a source of valuable information that can improve prediction model performance in addition to structured data. But the information available in both structured and unstructured data is likely dependent on multiple factors, such as the clinical setting. Models were generally developed in hospital care settings using a variety of text representations and machine-learning methods and we found a relationship between the types of text representation and machine-learning methods used. Furthermore, it is a cause for concern that only two studies externally validated their developed prediction models and that many studies had little attention for model explainability. Therefore, we suggest a focus on external validation and model explainability in future research. Additionally, we emphasize the importance of studying the use of text in non-English languages in prediction model development.

FUNDING

This work has received support from the European Health Data & Evidence Network (EHDEN) project. EHDEN has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

ACKNOWLEDGEMENTS

The authors would like to thank Christa Niehot from the Erasmus MC medical library for her input on the search strategy.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The extracted data used for generating the results are available as supplementary data.

REFERENCES

1. Reps JM, Schuemie MJ, Suchard MA, et al. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018;25(8):969-75.
2. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24(1):198-208.
3. Khalid S, Yang C, Blacketer C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. *Comput Methods Programs Biomed* 2021 doi: 10.1016/j.cmpb.2021.106394.
4. Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;23(5):1007-15.
5. Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. *Yearb Med Inform* 2020;29(1):208.
6. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Med Inform* 2020;8(3):e17984.
7. Assale M, Dui LG, Cina A, et al. The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. *Front Med (Lausanne)* 2019;6:66 doi: 10.3389/fmed.2019.00066.
8. Velupillai S, Suominen H, Liakata M, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *J Biomed Inform* 2018;88:11-19 doi: 10.1016/j.jbi.2018.10.005.
9. Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform* 2019;7(2):e12239 doi: 10.2196/12239.
10. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26(4):364-79.
11. Fu S, Chen D, He H, et al. Clinical Concept Extraction: a Methodology Review. *J Biomed Inform* 2020:103526.
12. Mujtaba G, Shuib L, Idris N, et al. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst Appl* 2019;116:494-520.
13. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. *J Am Med Inform Assoc* 2021.
14. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
15. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation* 2015;131(2):211-19.
16. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3(1):1-9.
17. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Crit Care Med* 2011;39(5):952.
18. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annu Symp Proc; 2001. American Medical Informatics Association.
19. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(suppl_1):D267-D70.
20. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform* 2006;121:279.

21. Marafino BJ, Park M, Davies JM, 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 doi: 10.1001/jamanetworkopen.2018.5097.
22. Menger V, Spruit M, Van Est R, et al. Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Netw Open* 2019;2(7) doi: 10.1001/jamanetworkopen.2019.6709.
23. AlShuweihi M, Salloum SA, Shaalan K. Biomedical corpora and natural language processing on clinical text in languages other than English: a systematic review. Recent Advances in Intelligent Systems and Smart Applications, 2021:491-509.
24. Névéol A, Dalianis H, Velupillai S, et al. Clinical natural language processing in languages other than english: opportunities and challenges. *J Biomed Semant* 2018;9(1):1-13.
25. Mahendra M, Luo Y, Mills H, et al. Impact of Different Approaches to Preparing Notes for Analysis With Natural Language Processing on the Performance of Prediction Models in Intensive Care. *Crit Care Explor* 2021;3(6).
26. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform* 2020:103655.
27. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2018;19(6):1236-46.
28. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245.
29. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;216:574.
30. Halpern Y, Horng S, Nathanson LA. A comparison of dimensionality reduction techniques for unstructured clinical text. *ICML 2012 Workshop on Clinical Data Analysis* 2012.
31. Karnik S, Tan SL, Berg B, et al. Predicting atrial fibrillation and flutter using electronic health records. *Annu Int Conf IEEE Eng Med Biol Soc* 2012;2012:5562-65.
32. Lehman LW, Saeed M, Long W, et al. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012;2012:505-11.
33. Li J, Guo L, Handly N, et al. Semantic-enhanced models to support timely admission prediction at emergency departments. *Netw Model Anal Health Inform Bioinform* 2012;1(4):161-72 doi: 10.1007/s13721-012-0014-6.
34. Huang SH, LePendu P, Iyer SV, et al. Toward personalizing treatment for depression: Predicting diagnosis and severity. *J Am Med Inform Assoc* 2014;21(6):1069-75 doi: 10.1136/amiajnl-2014-002733.
35. Huddar V, Rajan V, Bhattacharya S, et al. Predicting postoperative acute respiratory failure in critical care using nursing notes and physiological signals. *Annu Int Conf IEEE Eng Med Biol Soc* 2014;2014:2702-05 doi: 10.1109/EMBC.2014.6944180.
36. Kontio E, Airola A, Pahikkala T, et al. Predicting patient acuity from electronic patient records. *J Biomed Inform* 2014;51:35-40 doi: 10.1016/j.jbi.2014.04.001.
37. Poulin C, Shiner B, Thompson P, et al. Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS One* 2014;9(1) doi: 10.1371/journal.pone.0085733.
38. Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. *J Biomed Inform* 2014 doi: 10.1016/j.jbi.2014.08.006.
39. Caballero K, Akella R. Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records. *AMIA Annu Symp Proc* 2015;2015:1831-40.
40. Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform* 2015;54:114-20 doi: 10.1016/j.jbi.2015.02.003.

41. Perotte A, Ranganath R, Hirsch JS, et al. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *J Am Med Inform Assoc* 2015;22(4):872-80 doi: 10.1093/jamia/ocv024.
42. Roysden N, Wright A. Predicting Health Care Utilization After Behavioral Health Referral Using Natural Language Processing and Machine Learning. *AMIA Annu Symp Proc* 2015;2015:2063-72.
43. Cohen KB, Glass B, Greiner HM, et al. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning. *Biomed Inform Insights* 2016;8:11-8 doi: 10.4137/BII.S38308.
44. Hassanpour S, Langlotz CP. Predicting High Imaging Utilization Based on Initial Radiology Reports: A Feasibility Study of Machine Learning. *Acad Radiol* 2016;23(1):84-89 doi: 10.1016/j.acra.2015.09.014.
45. Hu D, Huang Z, Chan TM, et al. Utilizing Chinese admission records for MACE prediction of acute coronary syndrome. *Int J Environ Res Public Health* 2016;13(9) doi: 10.3390/ijerph13090912.
46. Luo YF, Rumshisky A. Interpretable Topic Features for Post-ICU Mortality Prediction. *AMIA Annu Symp Proc* 2016;2016:827-36.
47. McCoy TH, Castro VM, Roberson AM, et al. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* 2016;73(10):1064-71 doi: 10.1001/jamapsychiatry.2016.2172.
48. Miotto R, Li L, Kidd BA, et al. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016;6:26094 doi: 10.1038/srep26094.
49. Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry* 2016;6(10):e921 doi: 10.1038/tp.2015.182.
50. Soguero-Ruiz C, Hindberg K, Mora-Jiménez I, et al. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *J Biomed Inform* 2016;61:87-96 doi: 10.1016/j.jbi.2016.03.008.
51. Temple MW, Lehmann CU, Fabbri D. Natural Language Processing for Cohort Discovery in a Discharge Prediction Model for the Neonatal ICU. *Appl Clin Inform* 2016;7(1):101-15 doi: 10.4338/ACI-2015-09-RA-0114.
52. Buchan K, Filannino M, Uzuner Ö. Automatic prediction of coronary artery disease from clinical narratives. *J Biomed Inform* 2017;72:23-32 doi: 10.1016/j.jbi.2017.06.019.
53. Frost DW, Vembu S, Wang J, et al. Using the Electronic Medical Record to Identify Patients at High Risk for Frequent Emergency Department Visits and High System Costs. *Am J Med* 2017;130(5):601.e17-01.e22 doi: 10.1016/j.amjmed.2016.12.008.
54. Hong SN, Son HJ, Choi SK, et al. A prediction model for advanced colorectal neoplasia in an asymptomatic screening population. *PLoS One* 2017;12(8) doi: 10.1371/journal.pone.0181040.
55. Lucini FR, Fogliatto FS, da Silveira GJC, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. *Int J Med Inform* 2017;100:1-8 doi: 10.1016/j.ijmedinf.2017.01.001.
56. Zhang XY, Kim J, Patzer RE, et al. Prediction of Emergency Department Hospital Admission Based on Natural Language Processing and Neural Networks. *Methods Inf Med* 2017;56(5):377-89 doi: 10.3414/ME17-01-0024.
57. Adamou M, Antoniou G, Greasidou E, et al. Toward automatic risk assessment to support suicide prevention. *Crisis* 2018.
58. Bahl M, Barzilay R, Yedidia AB, et al. High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision. *Radiology* 2018;286(3):810-18 doi: 10.1148/radiol.2017170549.

59. Banerjee I, Gensheimer MF, Wood DJ, et al. Probabilistic Prognostic Estimates of Survival in Metastatic Cancer Patients (PPES-Met) Utilizing Free-Text Clinical Narratives. *Sci Rep* 2018;8(1):10037 doi: 10.1038/s41598-018-27946-5.
60. Boag W, Doss D, Naumann T, et al. What's in a Note? Unpacking Predictive Value in Clinical Note Representations. *AMIA Jt Summits Transl Sci Proc* 2018;2017:26-34.
61. Coulet A, Shah NH, Wack M, et al. Predicting the need for a reduced drug dose, at first prescription. *Sci Rep* 2018;8(1):15558 doi: 10.1038/s41598-018-33980-0.
62. Gligorijevic D, Stojanovic J, Satz W, et al. Deep attention model for triage of emergency department patients. *Proc SIAM Int Conf Data Min* 2018.
63. Golas SB, Shibahara T. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data: bmcmedinformdecismak ..., 2018.
64. Huang YX, Lee J, Wang S, et al. Privacy-Preserving Predictive Modeling: Harmonization of Contextual Embeddings From Different Sources. *JMIR Med Inform* 2018;6(2):278-91 doi: 10.2196/medinform.9455.
65. Ian ERWS, Tran N, Dubin JA, et al. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PLoS One* 2018;13(6) doi: 10.1371/journal.pone.0198687.
66. Krishnan GS, Kamath SS. A supervised learning approach for ICU mortality prediction based on unstructured electrocardiogram text reports. *Nat Lang Process Inf Syst* 2018.
67. Li Y, Yao L, Mao C, et al. Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes. *Proceedings (IEEE Int Conf Bioinformatics Biomed)* 2018;2018:683-86 doi: 10.1109/bibm.2018.8621574.
68. Menger V, Scheepers F, Spruit M. Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text. *Applied Sciences* 2018;8(6) doi: 10.3390/app8060981.
69. Parreco J, Hidalgo A, Kozol R, et al. Predicting mortality in the surgical intensive care unit using artificial intelligence and natural language processing of physician documentation. *Am Surg* 2018;84(7):1190-94 doi: 10.1177/000313481808400736.
70. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;1 doi: 10.1038/s41746-018-0029-1.
71. Sundararaman A, Ramanathan SV, Thati R. Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance. *Big Data Res* 2018;13:65-75 doi: 10.1016/j.bdr.2018.05.004.
72. Sushil M, Šuster S, Luyckx K, et al. Patient representation learning and interpretable evaluation using clinical notes. *J Biomed Inform* 2018;84:103-13 doi: 10.1016/j.jbi.2018.06.016.
73. Weissman GE, Hubbard RA, Ungar LH, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Crit Care Med* 2018;46(7):1125-32 doi: 10.1097/CCM.0000000000003148.
74. Yang Y, Wang X, Huang Y, et al. Ontology-based venous thromboembolism risk factors mining and model developing from medical records. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018. IEEE.
75. Afshar M, Dligach D, Sharma B, et al. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. *J Am Med Inform Assoc* 2019;26(11):1364-69 doi: 10.1093/jamia/ocz068.
76. Akbilgic O, Homayouni R, Heinrich K, et al. Unstructured Text in EMR Improves Prediction of Death after Surgery in Children. *Informatics (MDPI)* 2019;6(1) doi: 10.3390/informatics6010004.
77. Alvarez-Mellado E, Holderness E, Miller N. Assessing the Efficacy of Clinical Sentiment Analysis and Topic Extraction in Psychiatric Readmission Risk Prediction. *EMNLP (2019)* 2019.

78. Apostolova E, Uppal A, Galarraga JE, et al. Towards Reliable ARDS Clinical Decision Support: ARDS Patient Analytics with Free-text and Structured EMR Data. *AMIA Annu Symp Proc* 2019;2019:228-37.
79. Beeksma M, Verberne S, van den Bosch A, et al. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Med Inform Decis Mak* 2019;19(1):36 doi: 10.1186/s12911-019-0775-2.
80. Brown AD, Kachura JR. Natural Language Processing of Radiology Reports in Patients With Hepatocellular Carcinoma to Predict Radiology Resource Utilization. *J Am Coll Radiol* 2019;16(6):840-44 doi: 10.1016/j.jacr.2018.12.004.
81. Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA J Ethics* 2019;21(2):E167-E79 doi: 10.1001/amajethics.2019.167.
82. da Silva DA, ten Caten CS, dos Santos RP, et al. Predicting the occurrence of surgical site infections using text mining and machine learning. *PLoS One* 2019;14(12) doi: 10.1371/journal.pone.0226272.
83. Danielsen AA, Fenger MHJ, Østergaard SD, et al. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. *Acta Psychiatr Scand* 2019;140(2):147-57 doi: 10.1111/acps.13061.
84. Danilov G, Kotik K, Shifrin M, et al. Prediction of Postoperative Hospital Stay with Deep Learning Based on 101 654 Operative Reports in Neurosurgery. *Stud Health Technol Inform* 2019;258:125-29.
85. Gong J, Bai X, Li DA, et al. Prognosis Analysis of Heart Failure Based on Recurrent Attention Model. *Ing Rech Biomed* 2020;41(2):71-79 doi: 10.1016/j.irbm.2019.08.002.
86. Khadanga S, Aggarwal K, Joty S. Using clinical notes with time series data for ICU management. *EMNLP (2019)* 2019.
87. Kongburan W, Chignell M, Charoenkitkarn N, et al. Enhancing Predictive Power of Cluster-Boosted Regression With Text-Based Indexing. *IEEE Access* 2019;7:43394-405 doi: 10.1109/ACCESS.2019.2908032.
88. Korach ZT, Cato KD, Collins SA, et al. Unsupervised Machine Learning of Topics Documented by Nurses about Hospitalized Patients Prior to a Rapid-Response Event. *Appl Clin Inform* 2019;10(5):952-63 doi: 10.1055/s-0039-3401814.
89. Krishnan GS. Evaluating the quality of word representation models for unstructured clinical text based ICU mortality prediction. *Proc ICDCN* 2019.
90. Liu R, Greenstein JL, Sarma SV, et al. Natural Language Processing of Clinical Notes for Improved Early Prediction of Septic Shock in the ICU. *Annu Int Conf IEEE Eng Med Biol Soc* 2019;2019:6103-08 doi: 10.1109/EMBC.2019.8857819.
91. Mahajan SM, Ghani R. Combining Structured and Unstructured Data for Predicting Risk of Readmission for Heart Failure Patients. *Stud Health Technol Inform* 2019;264:238-42 doi: 10.3233/SHTI190219.
92. Makino M, Yoshimoto R, Ono M, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci Rep* 2019;9(1):11862 doi: 10.1038/s41598-019-48263-5.
93. Nakayama JY, Hertzberg V, Ho JC. Making sense of abbreviations in nursing notes: A case study on mortality prediction. *AMIA Jt Summits Transl Sci Proc* 2019;2019:275-84.
94. Payrovnaziri SN, Barrett LA, Bis D, et al. Enhancing Prediction Models for One-Year Mortality in Patients with Acute Myocardial Infarction and Post Myocardial Infarction Syndrome. *Stud Health Technol Inform* 2019;264:273-77 doi: 10.3233/SHTI190226.
95. Ross EG, Jung K, Dudley JT, et al. Predicting Future Cardiovascular Events in Patients with Peripheral Artery Disease Using Electronic Health Record Data. *Circ Cardiovasc Qual Outcomes* 2019;12(3) doi: 10.1161/CIRCOUTCOMES.118.004741.
96. Shin B, Hogan J, Adams AB, et al. Multimodal ensemble approach to incorporate various types of clinical notes for predicting readmission. *IEEE EMBS Int Conf Biomed Health Inform* 2019.

97. Si Y, Roberts K. Deep Patient Representation of Clinical Notes via Multi-Task Learning for Mortality Prediction. *AMIA Jt Summits Transl Sci Proc* 2019;2019:779-88.
98. Sterling NW, Patzer RE, Di MY, et al. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int J Med Inform* 2019;129:184-88 doi: 10.1016/j.ijmedinf.2019.06.008.
99. Sun M, Baron J, Dighe A, et al. Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes and Structured Multivariate Physiological Measurements. *Stud Health Technol Inform* 2019;264:368-72 doi: 10.3233/SHTI190245.
100. Wang LQ, Sha L, Lakin JR, et al. Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients With Dementia for Earlier Palliative Care Interventions. *JAMA Netw Open* 2019;2(7) doi: 10.1001/jamanetworkopen.2019.6972.
101. Weissman GE, Ungar LH, Harhay MO, et al. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *J Biomed Inform* 2019;89:114-21 doi: 10.1016/j.jbi.2018.12.001.
102. Yang YQ, Wang X, Huang Y, et al. Ontology-based venous thromboembolism risk assessment model developing from medical records. *BMC Med Inform Decis Mak* 2019;19 doi: 10.1186/s12911-019-0856-2.
103. Zhang X, Bellolio MF, Medrano-Gracia P, et al. Use of natural language processing to improve predictive models for imaging utilization in children presenting to the emergency department. *BMC Med Inform Decis Mak* 2019;19(1):287 doi: 10.1186/s12911-019-1006-6.
104. Bacchi S, Gluck S, Tan YR, et al. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. *Intern Emerg Med* 2020;15(6):989-95 doi: 10.1007/s11739-019-02265-3.
105. Barash Y, Soffer S, Grossman E, et al. Alerting on mortality among patients discharged from the emergency department: A machine learning model. *Postgrad Med J* 2020 doi: 10.1136/postgradmedj-2020-138899.
106. Barber EL, Garg R, Persenaire C, et al. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecol Oncol* 2021;160(1):182-86 doi: 10.1016/j.ygyno.2020.10.004.
107. Baxter SL, Klie AR, Saseendrakumar BR, et al. Predicting Mortality in Critical Care Patients with Fungemia Using Structured and Unstructured Data. *Annu Int Conf IEEE Eng Med Biol Soc* 2020;2020:5459-63 doi: 10.1109/EMBC44109.2020.9175287.
108. Ben Miled Z, Haas K, Black CM, et al. Predicting dementia with routine care EMR data. *Artif Intell Med* 2020;102 doi: 10.1016/j.artmed.2019.101771.
109. Chen YW, Zhang LG, Zhang J, et al. Preoperative Risk Prediction of Heart Failure with Numerical and Textual Attributes. *Int J Innov Comp Inf Control* 2020;16(6):2035-46 doi: 10.24507/ijicic.16.06.2035.
110. Chen WJ, Lu ZJ, You LJ, et al. Artificial Intelligence-Based Multimodal Risk Assessment Model for Surgical Site Infection (AMRAMS): Development and Validation Study. *JMIR Med Inform* 2020;8(6) doi: 10.2196/18186.
111. Chen CH, Hsieh JG, Cheng SL, et al. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. *Int J Med Inform* 2020;139 doi: 10.1016/j.ijmedinf.2020.104146.
112. Chen CH, Hsieh JG, Cheng SL, et al. Early short-term prediction of emergency department length of stay using natural language processing for low-acuity outpatients. *Am J Emerg Med* 2020;38(11):2368-73 doi: 10.1016/j.ajem.2020.03.019.
113. Danilov G, Kotik K, Shifrin M, et al. Predicting Postoperative Hospital Stay in Neurosurgery with Recurrent Neural Networks Based on Operative Reports. *Stud Health Technol Inform* 2020;270:382-86 doi: 10.3233/SHTI200187.

114. Fernandes M, Mendes R, Vieira SM, et al. Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PLoS One* 2020;15(3) doi: 10.1371/journal.pone.0229331.
115. Fernandes M, Mendes R, Vieira SM, et al. Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing. *PLoS One* 2020;15(4) doi: 10.1371/journal.pone.0230876.
116. Gensheimer MF, Aggarwal S, Benson KRK, et al. Automated model versus treating physician for predicting survival time of patients with metastatic cancer. *J Am Med Inform Assoc* 2020 doi: 10.1093/jamia/ocaa290.
117. Goodwin TR, Demner-Fushman D. A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. *J Am Med Inform Assoc* 2020;27(4):567-76 doi: 10.1093/jamia/ocaa004.
118. Guo WP, Xu ZM, Ye XJ, et al. A Time-Critical Topic Model for Predicting the Survival Time of Sepsis Patients. *Sci Program* 2020;2020 doi: 10.1155/2020/8884539.
119. Hane CA, Nori VS, Crown WH, et al. Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study. *JMIR Med Inform* 2020;8(6):e17819 doi: 10.2196/17819.
120. Hashir M, Sawhney R. Towards unstructured mortality prediction with free-text clinical notes. *J Biomed Inform* 2020;108 doi: 10.1016/j.jbi.2020.103489.
121. Heo TS, Kim YS, Choi JM, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *J Pers Med* 2020;10(4):1-11 doi: 10.3390/jpm10040286.
122. Hsu CC, Karnwal S, Mullainathan S. Characterizing the Value of Information in Medical Notes. *EMNLP (2020)* 2020.
123. Izquierdo JL, Ancochea J, Savana C-RG, et al. Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing. *J Med Internet Res* 2020;22(10):e21801 doi: 10.2196/21801.
124. Korach ZT, Yang J, Rossetti SC, et al. Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. *Int J Med Inform* 2020;135 doi: 10.1016/j.ijmedinf.2019.104053.
125. Le S, Allen A, Calvert J, et al. Development and Validation of a Convolutional Neural Network Model for ICU Acute Kidney Injury Prediction. *Kidney Int Rep* 2020 doi: 10.1016/j.ekir.2021.02.031.
126. Lee D, Jiang X, Yu H. Harmonized representation learning on dynamic EHR graphs. *J Biomed Inform* 2020;106 doi: 10.1016/j.jbi.2020.103426.
127. Levis M, Leonard Westgate C, Gui J, et al. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol Med* 2020:1-10 doi: 10.1017/S0033291720000173.
128. Li Y, Nair P, Lu XH, et al. Inferring multimodal latent topics from electronic health records. *Nat Commun* 2020;11(1) doi: 10.1038/s41467-020-16378-3.
129. Meng Y, Speier WF, Ong M, et al. HCET: Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Record for Predicting Depression. *IEEE J Biomed Health Inform* 2020;PP doi: 10.1109/JBHI.2020.3004072.
130. Mohammadi R, Jain S, Namin AT, et al. Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study. *JMIR Med Inform* 2020;8(11):e19761 doi: 10.2196/19761.
131. Mugisha C, Paik I. Pneumonia Outcome Prediction Using Structured And Unstructured Data From EHR. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2020. IEEE.

132. Nakatani H, Nakao M, Uchiyama H, et al. Predicting Inpatient Falls Using Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study. *JMIR Med Inform* 2020;8(4):e16970 doi: 10.2196/16970.
133. Obeid JS, Dahne J, Christensen S, et al. Identifying and Predicting Intentional Self-Harm in Electronic Health Record Clinical Notes: Deep Learning Approach. *JMIR Med Inform* 2020;8(7):e17784 doi: 10.2196/17784.
134. Roquette BP, Nagano H, Marujo EC, et al. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Netw* 2020;126:170-77 doi: 10.1016/j.neunet.2020.03.012.
135. Shukla SN, Marlin BM. Integrating Physiological Time Series and Clinical Notes with Deep Learning for Improved ICU Mortality Prediction. *Proc ACM Conf Health Inference Learn (2020)* 2020.
136. Sterckx L, Vandewiele G, Dehaene I, et al. Clinical information extraction for preterm birth risk prediction. *J Biomed Inform* 2020;110 doi: 10.1016/j.jbi.2020.103544.
137. Sterling NW, Brann F, Patzer RE, et al. Prediction of emergency department resource requirements during triage: An application of current natural language processing techniques. *J Am Coll Emerg Physicians Open* 2020;1(6):1676-83 doi: 10.1002/emp2.12253.
138. Tahayori B, Chini-Foroush N, Akhlaghi H. Advanced natural language processing technique to predict patient disposition based on emergency triage notes. *Emerg Med Australas* 2020 doi: 10.1111/1742-6723.13656.
139. Topaz M, Woo K, Ryvicker M, et al. Home Healthcare Clinical Notes Predict Patient Hospitalization and Emergency Department Visits. *Nurs Res* 2020;69(6):448-54 doi: 10.1097/NNR.0000000000000470.
140. Wang HY, Li YK, Khan SA, et al. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artif Intell Med* 2020;110 doi: 10.1016/j.artmed.2020.101977.
141. Weegar R, Sundström K. Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PLoS One* 2020;15(8 August 2020) doi: 10.1371/journal.pone.0237911.
142. Xu L, Hogan J, Patzer RE, et al. Noise Pollution in Hospital Readmission Prediction: Long Document Classification with Reinforcement Learning. *2020 BioNLP ACL Workshop on Biomedical Natural Language Processing* 2020.
143. Ye JC, Yao L, Shen JH, et al. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med Inform Decis Mak* 2020;20 doi: 10.1186/s12911-020-01318-4.
144. Zhang DD, Yin CC, Zeng JC, et al. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak* 2020;20(1) doi: 10.1186/s12911-020-01297-6.
145. Boag W, Kovaleva O, McCoy TH, et al. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Transl Psychiatry* 2021;11(1) doi: 10.1038/s41398-020-01104-w.
146. Chen YP, Lo YH, Lai F, et al. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: Algorithm development and validation study. *J Med Internet Res* 2021;23(1) doi: 10.2196/25113.
147. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun* 2021;12(1) doi: 10.1038/s41467-021-20910-4.
148. Klang E, Kummer BR, Dangayach NS, et al. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. *Sci Rep* 2021;11(1):1381 doi: 10.1038/s41598-021-80985-3.

149. Muhlestein WE, Monsour MA, Friedman GN, et al. Predicting Discharge Disposition Following Meningioma Resection Using a Multi-Institutional Natural Language Processing Model. *Neurosurgery* 2021 doi: 10.1093/neuros/nyaa585.
150. Oliwa T, Furner B, Schmitt J, et al. Development of a predictive model for retention in HIV care using natural language processing of clinical notes. *J Am Med Inform Assoc* 2021;28(1):104-12 doi: 10.1093/jamia/ocaa220.
151. Ribelles N, Jerez JM, Rodriguez-Brazzarola P, et al. Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients. *Eur J Cancer* 2021;144:224-31 doi: 10.1016/j.ejca.2020.11.030.
152. Tang C, Plasek JM, Shi X, et al. Estimating Time to Progression of Chronic Obstructive Pulmonary Disease with Tolerance. *IEEE J Biomed Health Inform* 2021;25(1):175-80 doi: 10.1109/JBHI.2020.2992259.
153. Yang HY, Kuang L, Xia FQ. Multimodal temporal-clinical note network for mortality prediction. *J Biomed Semant* 2021;12(1) doi: 10.1186/s13326-021-00235-3.