Journal of the American Medical Informatics Association, 30(3), 2023, 588–603
https://doi.org/10.1093/jamia/ocac240
Advance Access Publication Date: 13 December 2022
Review

Review

Combining text mining with clinical decision support in clinical practice: a scoping review

Britt W.M. van de Burgt[1,2], Arthur T.M. Wasylewicz[1], Bjorn Dullemond[3], Rene J.E. Grouls[4], Toine C.G. Egberts[5,6], Arthur Bouwman[2,7], and Erik M.M. Korsten[1,2]

[1]Department Healthcare Intelligence, Catharina Hospital Eindhoven, Eindhoven, The Netherlands, [2]Department of Electrical Engineering, Signal Processing Group, Technical University Eindhoven, Eindhoven, The Netherlands, [3]Department of Mathematics and Computer Science, Technical University of Eindhoven, Eindhoven, The Netherlands, [4]Department of Clinical Pharmacy, Catharina Hospital Eindhoven, Eindhoven, The Netherlands, [5]Department of Clinical Pharmacy, University Medical Centre Utrecht, Utrecht, the Netherlands, [6]Department of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands and [7]Department of Anesthesiology, Catharina Hospital Eindhoven, Eindhoven, The Netherlands

Corresponding Author: Britt W.M. van de Burgt, MSc, Department Healthcare Intelligence, Catharina Hospital Eindhoven, Michelangelolaan 2, 5623 EJ Eindhoven, The Netherlands; britt.vd.burgt@catharinaziekenhuis.nl

Received 25 August 2022; Revised 17 October 2022; Editorial Decision 6 November 2022; Accepted 1 December 2022

ABSTRACT

Objective: Combining text mining (TM) and clinical decision support (CDS) could improve diagnostic and therapeutic processes in clinical practice. This review summarizes current knowledge of the TM-CDS combination in clinical practice, including their intended purpose, implementation in clinical practice, and barriers to such implementation.
Materials and Methods: A search was conducted in PubMed, EMBASE, and Cochrane Library databases to identify full-text English language studies published before January 2022 with TM-CDS combination in clinical practice.
Results: Of 714 identified and screened unique publications, 39 were included. The majority of the included studies are related to diagnosis ($n=26$) or prognosis ($n=11$) and used a method that was developed for a specific clinical domain, document type, or application. Most of the studies selected text containing parts of the electronic health record (EHR), such as reports (41%, $n=16$) and free-text narratives (36%, $n=14$), and 23 studies utilized a tool that had software "developed for the study". In 15 studies, the software source was openly available. In 79% of studies, the tool was not implemented in clinical practice. Barriers to implement these tools included the complexity of natural language, EHR incompleteness, validation and performance of the tool, lack of input from an expert team, and the adoption rate among professionals.
Discussion/Conclusions: The available evidence indicates that the TM-CDS combination may improve diagnostic and therapeutic processes, contributing to increased patient safety. However, further research is needed to identify barriers to implementation and the impact of such tools in clinical practice.

Key words: text mining, CDS, NLP, electronic health record, free-text

INTRODUCTION

Medical errors remain common and each year patients are unnecessarily harmed due to such errors, despite efforts over the last 2 decades to improve the situation. To prevent medical errors, healthcare professionals must have the right information at the right time for the right patient without disruptions to their workflow.[1–4] In addition to addressing human factors and culture, information technology could significantly contribute to a reduction in the incidence of medical errors.[3,5–7] The rapid development of medical and information technology has led to an environment in which clinical data are digitally stored in patients' electronic health records (EHRs). Analysis of these data could contribute to a better, safer, and more efficient patient care.[8]

One technology to obtain these goals is clinical decision support (CDS).[9,10] These intend to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information.[9] CDS systems can be divided into basic and new CDS systems.[4,10] Basic CDS systems provide reminders to assist health care providers and implement evidence-based clinical guidelines at the point of care, but it cannot deal with different problems simultaneously: it assesses the clinical risk of a drug-drug interaction and that of renal insufficiency separately from each other.[11,12] The report by James described a new generation of CDS systems that "make it easy to do it right".[4] Beyond the use of reminders or digital checklists to increase compliance, these systems combine clinical data to help medical professionals manage an increasingly complex practice environment.[4,9,13–19] This sounds very promising, but these new generation CDS systems are not yet widely used in clinical practice, mainly due to 2 factors. First, clinician acceptance of CDS systems is low because most systems are complex and not well integrated into the clinical workflow. Second, CDS systems draw on a broad array of clinical information from many different information subsystems,[4] including structured and unstructured (free-text) data. An example of unstructured data that are still a crucial part of EHRs and the healthcare culture are free-text narratives (ie, descriptions of clinical observations, findings, and evaluations). Thus, the healthcare culture presents a barrier to implementing potentially useful computer applications. Even more, this unstructured data are not always accessible by CDS systems.[20]

Text mining (TM) could be a useful tool to extract information from unstructured data in EHRs.[8,21] TM is a variation of data mining that involves the detection of knowledge from textual data.[21–23] It is utilized worldwide in many settings. In healthcare, TM has been used to identify adverse drug events, help physicians make diagnoses, and informed treatment decisions.[24–28]

Combining TM with CDS systems could support professionals access the right information at the right time for the right patient without interruptions to the workflow. This, because TM can extract information from free-texts and CDS systems, can use this information to assist professionals in decision-making processes. The aim of this scoping review is to summarize the current knowledge on using TM combined with CDS systems in clinical practice. Specifically, the review addresses the questions: For what purposes are TM and CDS systems utilized? Are these tools implemented in clinical practice? If not, what are the barriers to such implementation?

MATERIALS AND METHODS

Registration and protocol

This scoping review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[29] The protocol was registered in PROSPERO (https://www.crd.york.ac.uk/prospero/, ID: CRD42022303470).

Data sources and searches

Medline, EMBASE, and Cochrane Library were searched utilizing the search criteria described in the Supplementary Material S1. Search criteria were broadly defined to capture all information that has been characterized as a TM-CDS combination. The review focused on studies published before January 1, 2022.

Medical subject headings terminology was utilized where possible (in PubMed and Cochrane), and keywords were utilized in EMBASE, a database that does not employ medical subject headings terminology. The search terms utilized were: text mining and Clinical Decision Support Systems (CDSS), text mining and CDSS, Natural Language Processing (NLP) and CDSS or NLP, and Clinical Decision Support Systems.

Study selection process

The electronic search results from the databases were merged using Mendeley (Mendeley Ltd., Version 1803), and duplicates removed. Next, the records were imported into the web-based tool Rayyan (https://rayyan.qcri.org/, Ouzzani 2016[30]) and independently reviewed for eligibility by 2 authors (BBT and EKN). The studies were screened by examining their titles, abstracts, and methods; after the screening, the full texts of the remaining publications were read. Disagreements on whether to include a study were resolved through discussion with a third team member (BDD).

Inclusion and exclusion criteria were determined *a priori*. The following inclusion criteria were applied: (1) Studies combined TM with CDS in a clinical practice. For our study, TM (also known as NLP) was defined as a process of deriving high-quality information from free-text or unstructured data drawn from a patient's medical record or parts thereof and converted into structured data. CDS was defined as a system that has the intention to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information, including the earlier described basic and new generation CDS systems. Therefore, the combination of TM and CDS contains these 2 definitions with the goal to combine unstructured and structured data. See Figure 1 for an overview of the described inclusion criteria. (2) Studies had an accessible abstract and full-text version in English. Relevant narrative reviews were evaluated for background information but excluded from the review. The reference lists of the included studies were cross-checked for additional studies.

Data extraction

The following information was extracted from the full text of the included articles: geographical location of the study, year of publication, application field (ie, etiology, diagnosis, prognosis, or therapy), clinical domain (eg, oncology), type of patient care (eg, inpatient), sample size validation, type of free-text used for TM (eg, radiology reports), TM-CDS tool (tool name and whether it already existed or was developed for the study), availability of the software source (yes, no), nature of comparison used in the study (eg, gold standard, tool), TM technique used (eg, annotation), CDS technique used (eg, Bayesian network), CDS way of advice (passive, active; as described in Kubben et al 2019),[10] quantitative outcome measures (eg, sensitivity) and reported estimates thereof, qualitative or additional findings, and barriers to implementation in clinical practice mentioned by the authors.

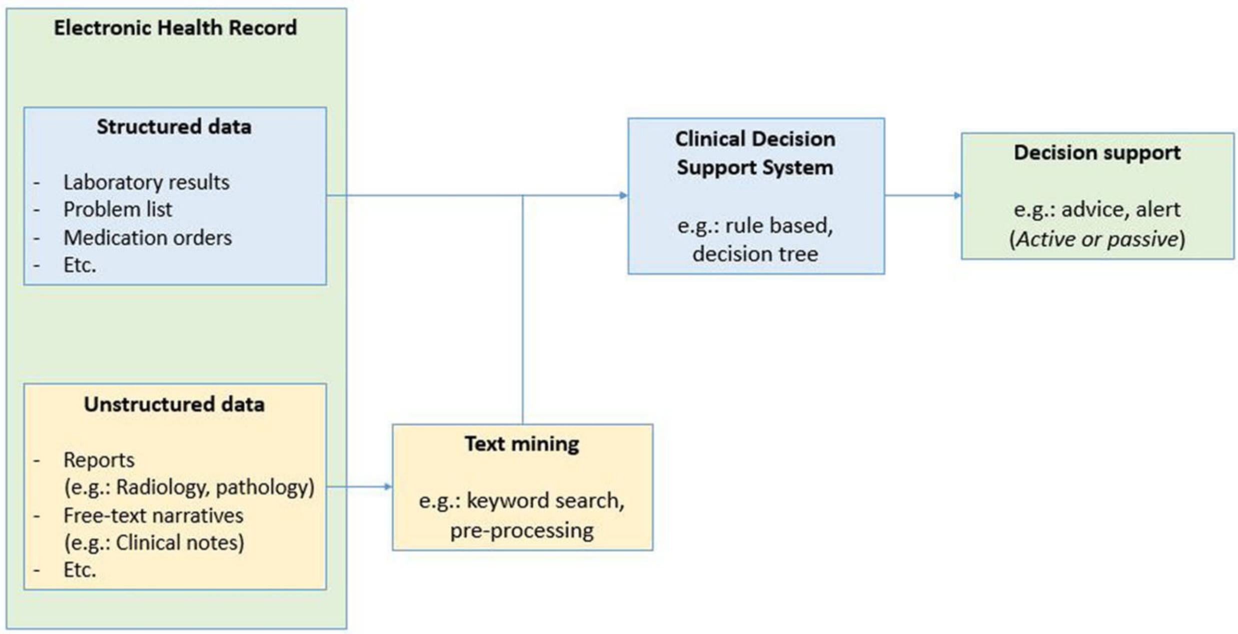

Figure 1. An overview of the inclusion criteria.

Data analysis

Quantitative outcome estimates were derived from the articles as reported or calculated by the current study team when the data were reported. Accuracy (ie, how closely the tool adheres to the gold standard) was defined as the sum of true positives and true negatives divided by the total number of cases. Sensitivity (ie, the tool's ability to identify true positive cases) was defined as the number of true positives divided by the sum of true positives and false negatives. Specificity (ie, the tool's ability to exclude false positive cases) was defined as the number of true negatives divided by the sum of true negatives and false positives. Positive predictive value (PPV) (ie, the likelihood that the tool corresponds to a true positive case) was defined as the number of true positives divided by the total number of positive cases identified by the tool. Negative predictive value (NPV) (ie, the likelihood that a record not coded for the condition is a true negative case) was defined as the number of true negatives divided by the total number of negative cases. The F-score (a measure of the test's accuracy) was calculated as PPV times sensitivity times 2 divided by the sum of sensitivity and PPV. The highest possible F-score value is 1.0 and the lowest possible value is 0. Cohen's kappa was used to measure inter-rater reliability (ie, concordance between recommendations). A Cohen's kappa of 1 is considered to indicate perfect agreement.

RESULTS

Electronic database searches yielded 850 studies, of which 714 were unique (see Figure 2). After screening, 115 studies were selected for full-text review, and 39 studies were included in the final analysis. Of the 76 excluded studies, 47 were excluded because they did not include CDS systems or TM, but data mining or another kind of mining. The remaining 29 studies were excluded because they were abstract-only ($n = 10$), reviews ($n = 5$), or did not combine TM with CDSS ($n = 14$). The majority of the included studies were used for diagnosis (67%; $n = 26$), used reports for TM/CDS (41%; $n = 16$), included data concerning inpatients (59%; $n = 22$), evaluated a tool that was developed for the study (59%; $n = 23$), were performed in an English-speaking country (United States and Australia; 90%; $n = 35$), and were conducted after 2011 (72%; $n = 28$), see Table 1.

Table 2 provides the data extracted from the included studies. The majority of the studies ($n = 33$) contained quantitative data, 6 studies included only qualitative information.

Application field

The majority of the studies were related to either the diagnostic process (67%; $n = 26$) or prognosis (28%; $n = 11$). No studies were related to etiology. One study, Nguyen et al,[31] was included in both the diagnosis and therapy categories. Most articles related to the diagnostic process concerned pulmonary diseases ($n = 7$) or cardiovascular diseases ($n = 6$). Each study focused on one disease to support diagnostic decision making. However, one study by Yang et al evaluated the misdiagnosis rate for several common diseases, for example hypertension and diabetes.[32]

The studies related to prognosis had 2 primary aims. The first was to assist clinicians by increasing patient follow-up and adherence to guidelines.[33–35] The second was to improve patient safety by extracting medical problems from electronic clinical documents to maintain a problem list that was as complete as possible.[36,37] The remaining studies in this category contributed to patient safety but had no common aim. Their purposes included reducing errors in eligibility criteria,[38] building a probabilistic topic model to predict clinical order patterns,[39] testing previously defined triggers,[40] enhancing protocol assignment,[41] and evaluating the impact of appropriate use criteria.[42]

Three of the 39 studies were in the therapy category. Two of these studies developed tools to support physicians in prescribing the correct antibiotic or dosage, the third aimed to reduce sedation order errors.[31,43,44]

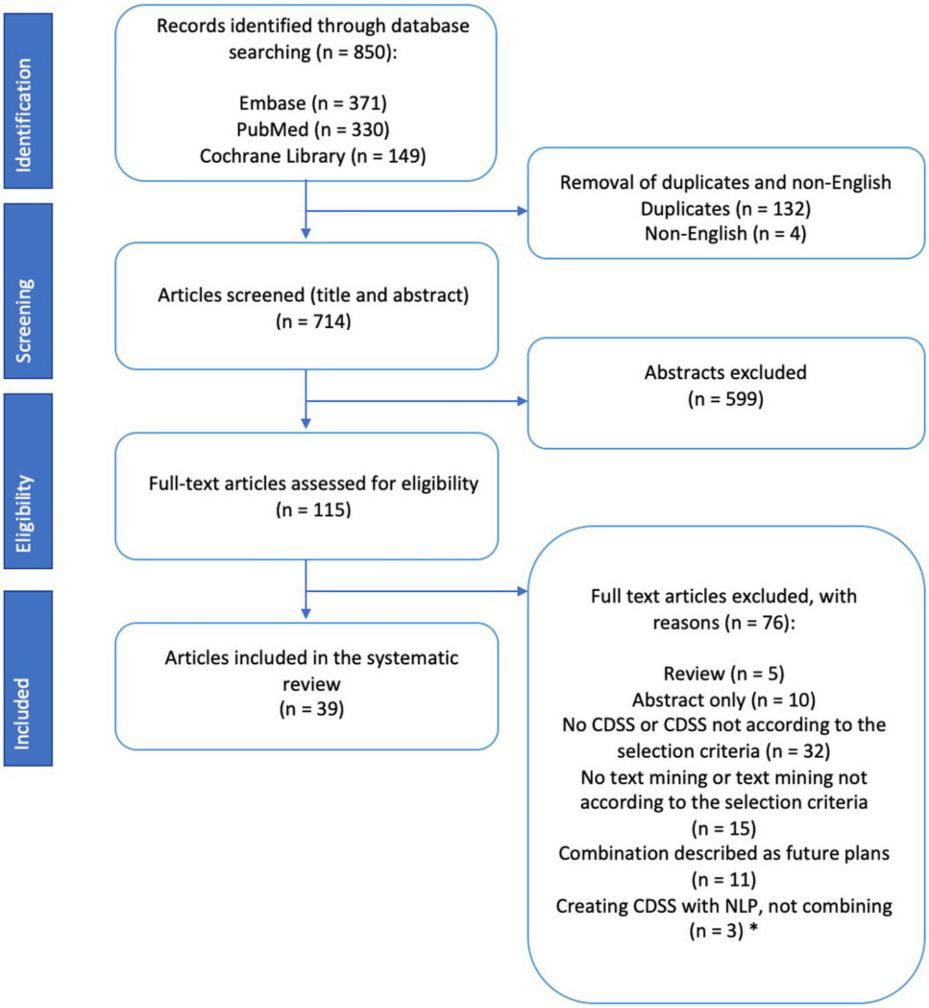

Figure 2. PRISMA flow diagram of the included studies. CDSS: clinical decision support systems; NLP: Natural Language Processing. *Creating CDSS with NLP, not combining; these studies used NLP to create a rule-based system, but did not use NLP to extract free-text or unstructured data and therefore could not combine unstructured data with structured data.

Free-text used for TM and tools

The majority of the studies utilized reports (41%; $n=16$), subdivided into radiology ($n=7$), pathology ($n=6$) and other ($n=1$), or free-text narratives (36%; $n=14$), and 23 studies (59%) utilized a tool that were the software was "developed for the study". Fifteen studies (38%) had a software source that was available for use by others. Two of the utilized tools, the REgenstrief eXtraction tool (REX) and Medical Language Extraction and Encoding System (MedLEE) were used to combine CDS and TM in 4 studies.[45–48] The REX tool uses pattern matching and a rule-based NLP system to extract patient information from admission notes, radiology reports, and pathology reports,[45,46] whereas MedLEE extracts, structures, and encodes clinical information drawn from textual patient reports.[49]

Quantitative and qualitative outcomes

The overall quantitative outcomes of the studies ($n=33$) varied. Specifically, PPV ranged from 7.5% to 100.0%, sensitivity ranged from 47% to 100%, specificity ranged from 63% to 100%, NPV ranged from 95.6% to 100.0%, F-score ranged from 25.00% to 99.89%, Cohen's kappa ranged from 58.3% to 90.0%, and accuracy ranged from 84.00% to 98.67%. No difference in the variation of quantitative outcomes was observed based on the application field. Nearly every study used the outcomes of PPV, sensitivity, and specificity. However, outcome accuracy was only used in diagnostic studies, Cohen's kappa was most commonly used in prognostic studies (66.67%; $n=2$), and neither accuracy nor Cohen's kappa were used in therapy studies. Notably, all quantitative outcomes from the "open-source available software" were higher than the outcomes from tools that did not have open-source available software. In addition, studies that used the REX tool had the highest quantitative outcome values.

All 6 studies that included qualitative findings reported outcomes that contributed to the study goal. The goals of these studies included adhering to clinical pathways with 100% compliance, decreasing the use of computed tomography scans, identifying trauma patients or children with suspected injuries, defining the status of epilepsy, and interpreting Papanicolaou test reports.

Use in clinical practice and barriers to implement

In 9 studies (23%), the tool was used in clinical practice, whereof 8 implemented a combination of TM-CDS in a hospital setting in the

Table 1. Characteristics of the included studies

Characteristics	Studies (n)	%
Geographical location of study		
United States	34	87
Europe (Spain, Finland)	2	5
Asian (Taiwan, China)	2	5
Australia	1	3
Year of publication		
Before 2001	3	8
2005–2010	8	21
2011–2015	11	28
2016–2021	17	44
Application field[a]		
Etiology	0	0
Diagnosis	26	67
Prognosis	11	28
Therapy	3	8
Clinical domain		
Oncology	4	10
Cardiovascular	6	15
Infectious diseases	4	10
Pulmonary diseases	7	18
Radiology	3	8
Adherence to clinical guideline[b]	4	10
Maintenance of a problem list	2	5
Other (eg, epilepsy, depression, hospital admission)	9	23
Type of patient care		
Inpatient	23	59
Outpatient	7	18
Both	6	15
Unknown	3	8
Sample size validation[c]		
<500	20	51
500–10 000	13	33
>10 000	5	13
Free-text used for text mining		
Reports (eg, radiology reports, pathology reports)	16	41
Orders (eg, clinical orders)	2	5
Discharge summaries	2	5
Clinical notes/free-text narratives (eg, free-text documents)	14	36
Other (eg, messages, scheduling data)	5	13
TM-CDS tool		
Software already developed[d]	16	41
Developed for the study[e]	23	59
Source available software[f]	15	38

CDS: clinical decision dupport; USMSTF: United States Multi-Society Task Force; TM: text mining.

[a]One study was included in 2 categories being, diagnosis and therapy.

[b]This includes studies that tried to improve adherence to clinical guidelines.

[c]This includes patients and free-text, Supplementary Appendix S2 shows which are patients and which are free-text.

[d]This includes tools that were developed before the study, meaning already existing tools that were used in this study.

[e]This includes tools that were developed in the study by the authors.

[f]Tools whereof the source was available for others, containing tools that were developed for the study or tools whereof the software was already developed.

United States. Four of these 8 studies implemented the tool in a hospital emergency department, a setting in which rapid diagnosis is preferable.[42,47,50–54] Only 8% of studies ($n=3$) implemented a real-time tool.[34,52,55] One of these, Cruz et al, implemented the first real-time TM-CDS combination tool outside the United States, in Spain.[34]

The primary reason that TM-CDS combination tools were not yet implemented in clinical practice was the complexity of natural language and low specificity and sensitivity. Natural language often includes text mistakes, abbreviations, and misspellings.[31,45,48,56–58] Friedlin et al (2008) and Matheny et al (2012) have described additional problems in tools understanding natural language. Word-sense disambiguation and negation detection were the main causes of NLP-related errors in these 2 studies, and these barriers made it difficult or impossible for CDS-TM tools to interpret free-text.[45,59]

Another obstacle to implement these tools was EHR incompleteness. Occasionally, clinical information was not documented in a patient's charts, which can result in an insufficient amount of information for meaningful processing and measurement.[31,32,34,60] Another potential source of error in electronic notes is introduced through the "cut-and-paste" feature. For example, relative references such as "five years ago" could be propagated over multiple years of notes and therefore lead to misdiagnosis of a patient.[61] An additional concern regarding the interpretability of the results and generalizability of the findings is that data from only one hospital were included in most of the studies, which may erase differences in workflows, domain-specific NLP methods, and EHRs between hospitals.[38,40,41,43,51,57,60,62]

Other barriers for implementation include the validation of the tool and the unfamiliar interface.[63–66] Mendonça et al found that even if the output of a natural language processor accurately extracts and structures the information in patient reports, it does not guarantee that the tool will be useful in a clinical practice. Many steps, like a testing phase, are required before such tools can be used.[48]

Furthermore, Matheny et al and Sung et al[38] found that the performance of a tool was directly related to the number of iterations (or sample size) performed on rule building in the training set. They did not measure the time spent on processing patient clinical notes. However, Sung et al[38] observed that long processing time is a weakness of MetaMap that renders the tool insufficient for real-time annotation of a large amount of clinical notes. They recommended building an information technology infrastructure that would be capable of processing a large volume of notes prior to implementation and usage of the tool.[38]

Lack of input from an expert team is another major barrier to usage of TM-CDS tools in clinical practice. Friedlin and McDonald (2008)[67] reported that the developer of the software also acted as a gold standard and evaluator of the data extraction process. Similarly, Jain et al and Wagholikar et al (2012 and 2013)[47,65,66] suggested that their results may be biased because the manual coding of one physician was being used as a gold standard. Based on this observation, Wagholikar et al concluded that it was necessary to consult other expert physicians to validate the tool.[66]

DISCUSSION

This review covers the field of TM-CDS combinations in clinical practice. Many studies mentioned a TM-CDS combination; however, only 39 studies were identified that combined TM with CDS and reported the results of this combination. The majority of the included studies are related to diagnosis and most of the studies used a method that was developed for a specific clinical domain, document type, or application. Most of the evaluated TM-CDS tools have not been implemented in clinical practice. The overall

Table 2. Outcomes and additional findings of the studies combining CDS and TM

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%)							Additional findings
				A	Sens	Spec	PPV	NPV	F	Co K	
Aronsky et al 2001[68]	Diagnosis	Comparing a clinically valid gold standard (3-step diagnostic evaluation process and 8 independent physicians) versus the model, of ED patients whose CXR report was available during the encounter to extract pneumonia diagnosis from unstructured data, detect and prompt initiation of antibiotic treatment to reduce disease severity and mortality	Sensitivity, specificity, PPV, NPV	n/a	96	63	14.2	99.5	25[a]	n/a	The area under the receiver operating characteristic curve was 0.881 (95% CI, 0.822–0.925) for the Bayesian network alone and 0.916 (95% CI, 0.869–0.949) combined (*P* = .01)
Bozkurt et al 2016[57]	Diagnosis	Comparing reference standard decision support system with the NLP decision support systems of patients with mammography reports, to provide decision support as part of the workflow of producing the radiology report.	Accuracy	97,58	n/a	n/a	n/a	n/a	n/a	n/a	The system performed extraction of imaging observations with their modifiers from text reports with precision = 94.9%, recall = 90.9%, and F-measure = 92%. They also compared the BI-RADS categories, accuracy rate of the Bayesian network outputs for each setting were calculated as 98.14% (history nodes included) and 98.15% (history nodes not included). The NLP-DSS and RS-DSS had closely matched probabilities, with a mean paired difference of 0.004 ± 0.025. The concordance correlation of these paired measures was 0.95.
Byrd et al 2012[58]	Diagnosis	Comparing the performance of the machine-learning and rule-based labelers of patients diagnosed with HF in primary care to identify heart failure with the Framingham diagnostic criteria to detect heart failure early.	Precision, recall, F-score	n/a	89.6	n/a	92.5	n/a	93.2	n/a	Detection with the Framingham criteria had an F-score of 0.910. Encounter labeling achieves an F-score of 0.932.
Chen et al 2017[39]	Prognosis	Comparing clinical order suggestions against the "correct" set of orders that actually occurred within a follow-up verification time for the patient with an encounter from their initial presentation until hospital discharge to build a probabilistic topic model representations of hospital admissions processes	Sensitivity and PPV	n/a	47	n/a	24	n/a	32[a]	n/a	Existing order sets predict clinical orders used within 24 h with area under the receiver operating characteristic curve 0.81, precision 16%, and recall 35%. This can be improved to 0.90, 24%, and 47% by using probabilistic topic models to summarize clinical data into up to 32 topics.

(continued)

Table 2. continued

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%) A	Sens	Spec	PPV	NPV	F	Co K	Additional findings
Cruz et al 2019[34b,c]	Prognosis	Comparing the same patients in the primary care area of SESCAM with recommendations before and after implementation of the CDSS to improve Adherence to Clinical Pathways and reducing clinical variability	Qualitative	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Adherence rates to clinical pathways improved in 8 out of 18 recommendations when the real-time CDSS was employed, achieving 100% compliance in some cases. The improvement was statistically significant in 3 cases ($P < .05$). Considering that this was a preliminary study, these results are promising.
Day et al 2007[54b]	Diagnosis	Reading documentation to determine whether registry inclusion criteria were met and printing admission lists versus the tool of trauma patients or patients who died to automate the process of identifying trauma patients	Qualitative	n/a	n/a	n/a	n/a	n/a	n/a	n/a	This program has made the job of identifying trauma patients much less complicated and time consuming. It improved the efficiency and reduced the amount of time wasted using multiple for-mats to ensure that all patients who qualify for inclusion are found. The program also stores relevant patient information to a permanent electronic database
Denny et al 2010[61]	Prognosis	Comparing the gold standard (expert physicians' manual review of EMR notes) versus the tool of patients whose colonoscopy statuses were unknown to detect colorectal cancer screening status	Recall and precision	n/a	TR: 91 CS : 82 CC : 93	n/a	TR : 95 CS : 95 CC : 95	n/a	TR : 93[a] CS : 88[a] CC : 94[a]	n/a	–
Evans et al 2016[53b]	Diagnosis	Comparing patients with HF treated using the new tool compared with HF patients who had received standard care at the same hospital before the tool was implemented to help identify high risk heart failure patients	Sensitivity, specificity, PPV	n/a	82.6–95.3	82.7–97.5	97.45	n/a	89–96[a]	n/a	–
Fiszman et al 2000[69]	Diagnosis	Comparing SymText against 4 physicians, 2 different keyword searches, and 3 lay persons of patients with a primary ICD-9 hospital discharge diagnosis of bacterial pneumonia to find cases of CAP to support diagnosis and treatment.	Recall, precision and specificity	n/a	95	85	78	n/a	86[a]	n/a	–
Friedlin and McDonald 2006[46]	Diagnosis	Comparing REX to the gold standard (specially trained human coders as well as an experienced physician) of patients who have had a chest x-ray with dictation reports to identify congestive heart failure	Sensitivity, specificity, PPV, NPV	n/a	100	100	95.4	100	98[a]	n/a	–

(continued)

Table 2. continued

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%)							Additional findings
				A	Sens	Spec	PPV	NPV	F	Co K	
Friedlin et al 2008[45]	Diagnosis	Comparing REX to the gold standard (human review) of patients with MRSA keywords to improve reporting notifiable diseases with automated electronic laboratory MRSA reporting.	Sensitivity, specificity, PPV, F-measure	n/a	99.96	99.71	99.81	99.93	99.89	n/a	REX identified over 2 times as many MRSA positive reports as the electronic lab system without NLP.
Garvin et al 2018[60]	Diagnosis	Comparing CHIEF versus reference standard and of External Peer Review Program cases involving HF patients discharged from 8 VA medical centers to accurately automate the quality measure for inpatients with HF.	Sensitivity and PPV	n/a	RS: 98.9 EPRP: 98.5	n/a	98.7	n/a	RS: 99[a] EPRP : 99[a]	n/a	Reference standard (RS) External Peer Review Program (EPRP). Of the 1083 patients available for the NLP system, the CHIEF evaluated and classified 100% of cases.
Hazlehurst et al 2005[56]	Diagnosis	Comparing MediClass versus the gold standard of patients who are smokers to detect clinical events in the medical record	Specificity and sensitivity	n/a	82	93	n/a	n/a	n/a	n/a	–
Imler et al 2013[63]	Diagnosis	Comparing annotation by gastroenterologists (reference standard) versus the NLP system of veterans who had an index colonoscopy to extract meaningful information from free-text gastroenterology reports for secondary use, to connect the pathologic record that is generally disconnected from the reports	Recall, precision, accuracy, and f-measure	L: 97 S: 96 *N*: 84	L > 82 S > 92 *N* > 66	n/a	L > 95 S > 95 *N* > 64	n/a	L : 96 S : 96 *N*: 62	n/a	
Imler et al 2014[64]	Diagnosis	Comparing NLP-based CDS surveillance intervals with those determined by paired, blinded, manual review of patients with an index colonoscopy for any indication except surveillance of a previous colorectal neoplasia to improve adherence to evidence-based practices and guidelines in endoscopy.	Kappa statistic	n/a	n/a	n/a	n/a	n/a	n/a	74	Fifty-five reports differed between manual review and CDS recommendations. Of these, NLP error accounted for 54.5%, incomplete resection of adenomatous tissue accounted for 25.5%, and masses observed without biopsy findings of cancer accounted for 7.2%. NLP based CDS surveillance intervals had higher levels of agreement with the standard than the level agreement between experts.
Jain et al 1996[47b]	Diagnosis	Comparing manual coding (gold standard) versus MedLee of patients with culture positive tuberculosis to find cases of tuberculosis in radiographic reports to identify eligible patients from the data.	Sensitivity	n/a	With L: 78.9 w/o L: 85.4	n/a	n/a	n/a	n/a	n/a	MedLEE agreed on the classification of 152/171 (88.9%) reports 129/142 (90.8%) suspicious for TB and 23/29 (79.3%) not suspicious for TB; and 1072/1197 (89.6%) terms indicative of TB. Analysis showed that most of the discrepancies were caused by MedLEE not finding the location of the infiltrate. By ignoring the location (L) of the infiltrate, the agreement became 157/171 (91.8%) reports and 946/1026 (92.2%) terms.

(continued)

Table 2. continued

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%)							Additional findings
				A	Sens	Spec	PPV	NPV	F	Co K	
Jones et al 2012[52b]	Diagnosis	Screening tool sensitivity and specificity as well as for ICD-9 plus radiographic confirmation were compared to physician review of ED patients with a chest x-ray of CT scan to identify patients with pneumonia	Sensitivity, PPV, specificity, NPV	n/a	61	96	52	97	56[a]	n/a	Among 41 true positive cases, ED physicians recognized and agreed with the tool in 39%. In only 6 cases did physicians proceed to complete the accompanying pneumonia decision support tool. Of the 39 false positive cases, the NLP incorrectly identified pneumonia in 74%. Of the 8 false negative cases, one was due to failure of the NLP to identify pneumonia
Kalra et al 2020[41]	Prognosis	Comparing manually review versus the models (kNN, RF, DNN) of older men and include both primary and tertiary care indications to enhance multispecialty CT and MRI protocol assignment quality and efficiency	Precision and recall	n/a	kNN: 83.4 RF: 92.2 DNN: 91.5	n/a	kNN: 77.5 RF: 81.7 DNN: 83.6	n/a	kNN : 80[a] RF : 86[a] DNN : 87[a]	n/a	Baseline protocol assignment performance achieved weighted precision of 0.757–0.824. Simulating real-world deployment using combined thresholding techniques, the optimized deep neural network model assigned 69% of protocols in automation mode with recall 95% =(weighted accuracy). In the remaining 31% of cases, the model achieved 92% accuracy in CDS mode.
Karwa et al 2020[33]	Prognosis	Comparison of Colonoscopy Follow-up Recommendations between CDS algorithm and endoscopists of patients with a colonoscopy to assist clinicians to generate colonoscopy follow-up intervals based on the USMSTF guidelines	Cohen's Kappa	n/a	n/a	n/a	69%	n/a	n/a	58.3	Discrepant recommendations by endoscopists were earlier than guidelines in 91% of cases.
Kim et al 2015[70]	Diagnosis	A 5-fold cross validation with the tool to measure the contribution of each of the 4 subset (lexial, concept position, related concept, section) of patients with LVEF or LVSF to classify the contextual use of both quantitative and qualitative LVEF assessments in clinical narrative documents.	Recall, precision and F-measure	n/a	QT: 95.6 QL: 94.2	n/a	QT: 95.6 QL: 94.2	n/a	QT: 95.6 QL: 94.2	n/a	The experimental results showed that the classifiers achieved good performance, reaching 95.6% F1-measure for quantitative (QT) assessments and 94.2% F1-measure for qualitative (QL) assessments in a 5-fold cross validation evaluation.
Kivekäs et al 2016[40c]	Prognosis	Comparing the review team versus SAS of adult epilepsy patients to test the functionality and validity of the previously defined triggers to describe the status of epilepsy patient's well-being.	Qualitative	n/a	n/a	n/a	n/a	n/a	n/a	n/a	In both medical and nursing data, the triggers described patients' well-being comprehensively. The narratives showed that there was overlapping in triggers.
	Diagnosis	Comparing manual reference versus algorithm of patients with at least		n/a	ASD: 84 SDA:	n/a	ASD: 91	n/a	ASD: 87	n/a	Among those instances in which the automated system matched the reference set

(continued)

Table 2. continued

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%)							Additional findings
				A	Sens	Spec	PPV	NPV	F	Co K	
Matheny et al 2012[59]		one surgical admission to identify infectious symptoms	F measure, Fleiss' Kappa, precision, recall,		62		SDA: 67		SDA: 64		determination for symptom, the system correctly detected 84.7% of positive assertions, 75.1% of negative assertions, and 0.7% of uncertain assertions.
Mendonça et al 2005[48]	Diagnosis	Comparing Clinicians' judgments versus the tool of infants admitted at the NICU to identify pneumonia in newborns, to reduce manual monitoring	Sensitivity, specificity and PPV	n/a	71	99	7.5	n/a	n/a	n/a	–
Meystre and Haug 2005[37]	Prognosis	Comparing NLP tools versus reference standard which was created with a chart review of patients admitted in a cardiovascular unit, stay of at least 48h, and with a discharge diagnosis in the list of the 80 selected diagnosis to extract medical problems from electronic clinical document to maintain, complete and up-to-date a problem list	Precision and recall Cohen's Kappa	n/a	74	n/a	75.6	n/a	75[a]	90	A custom data subset for MMTx was created, making it faster and significantly improving the recall to 0.896 with a non-significant reduction in precision.
Meystre and Haug 2008[36]	Prognosis	Comparing the control group (the standard electronic problem list) versus the intervention group (the Automated Problem List system) of inpatients of the 2 inpatients wards for at least 48h, > 18 years and not already enrolled in a previous phase of this study to improve the completeness and timeliness of an electronic problem list	Sensitivity, specificity, PPV and NPV	n/a	81.5	95.7	78.4	95.6	80[a]	n/a	–
Nguyen et al 2019[31c]	Diagnosis Therapy	Comparing the resultant number of test results identified by the system for clinical review to the full set of test results that would have been manually reviewed of patients with ED encounters, to ensure important diagnoses are recognized and correct antibiotics are prescribed.	PPV, sensitivity and F-measure	n/a	94,3	n/a	85,8	n/a	89,8	n/a	–
Raja et al 2012[50b]	Diagnosis	Pulmonary angiography were compared before and after CDS implementation of ED patients who underwent CT pulmonary angiography to decrease the use and increase in yield of CT for acute pulmonary embolism	Accuracy, sensitivity, PPV, NPV and specificity	97.8	91.3	98.7	91.3	98.7	91.3	n/a	Quarterly CT pulmonary angiography use increased 82.1% before CDSS implementation, from 14.5 to 26.4 examinations per 1000 patients ($P<.0001$). After CDSS implementation, quarterly use decreased 20.1%, from 26.4 to 21.1 examinations per 1000 patients ($P=.0379$).

(continued)

Table 2. continued

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%)							Additional findings
				A	Sens	Spec	PPV	NPV	F	Co K	
Raja et al 2019[42b]	Prognosis	Comparing research assistant Golden standard (3 physicians) versus the tool of patients < 50 years with a history of uncomplicated nephrolithiasis presenting to the ED to evaluate the impact of an appropriate use criterion for renal colic based on local best practice, implemented on the ED use of CT.	Qualitative	n/a	n/a	n/a	n/a	n/a	n/a	n/a	The final sample included 467 patients (194 study site) before and 306 (88 study site) after AUC implementation. The study site's CT of ureter rate decreased from 23.7% (46/194) to 14.8% (13/88) ($P = .03$) after implementation of the AUC. The rate at the control site remained unchanged, 49.8% (136/273) versus 48.2% (105/218) ($P = .3$).
Rosenthal et al 2019[51b]	Diagnosis	Comparing UPMC Children's Hospital of Pittsburgh earlier study results versus UPMC Hamot and Mercy of children < 2 years who triggered the EHR-based alert system to increase the number of young children identified as having injuries suspicious for physical abuse	Qualitative	n/a	n/a	n/a	n/a	n/a	n/a	n/a	A total of 242 children triggered the system, 86 during the pre-intervention and 156 during the intervention. The number of children identified with suspicious injuries increased 4-fold during the intervention ($P<.001$). Compliance was 70% (7 of 10) in the pre-intervention period versus 50% (22 of 44) in the intervention, a change that was not statistically different ($P=.55$).
Shen et al 2020[43]	Therapy	Comparing Pre-existing workflow versus pilot workflow of patients undergoing outpatient endoscopy to decrease sedation-type order errors	Precision, PPV, NPV, Sensitivity and Specificity.	n/a	89.1	99.2	28.5	99.9	43[a]	n/a	–
Smith et al 2021[55b]	Diagnosis	Comparing the algorithm to manual review, to expected performance of the model and to prior work in adults using CXR and other clinical data to recognize patients with pneumonia < 18 years.	Sensitivity, specificity, PPV, F-measure	n/a	89.9	94.9	78.1	n/a	83.5	n/a	–
Stultz et al 2019[44]	Therapy	Comparing different meningitis dosing alert triggers and dosing error rates between antimicrobials with and without meningitis order sentences of patients admitted to an inpatient pediatric service or the pediatric ED, to provide a meningitis specific dosing alert for detecting meningitis management	Sensitivity, PPV	n/a	67,5	n/a	80,9	n/a	74[a]	n/a	Antimicrobials with meningitis order sentences had fewer dosing errors (19.8% vs 43.2%, $P < .01$).
Sung et al 2018[38c]	Prognosis	Comparing Metamap to IVT eligibility criteria of adult ED patients with AIS who presented within 3h of onset but were not treated with IVT to errors in determining eligibility for IVT in stroke patients	Precision, recall and F-score	n/a	PL: 81.2 DL: 97.2	n/a	PL: 99.8 DL: 100	n/a	PL: 89.5 DL: 98.6	n/a	Users using the task-specific interface achieved a higher accuracy score than those using the current interface (91% vs 80%) in assessing the IVT eligibility criteria. The completion time between the interfaces was statistically similar (2.46 min vs 1.70 min).

(continued)

Table 2. continued

Study	Application field	Comparison	Quantitative outcome measure	Outcomes estimates (%)							Additional findings
				A	Sens	Spec	PPV	NPV	F	Co K	
Wadia et al 2017[35]	Prognosis	Comparing the gold standard (discussion between pathologists and oncologist) versus the tool of patients undergoing colonoscopy or surgery for colon lesions to identifying cases that required close clinical follow up	Recall, specificity, precision and F-score	n/a	100	98.5	95.2	n/a	97.5	n/a	–
Wagholikar et al 2012[65]	Diagnosis	Comparing the interpretation of free-text Pap reports of Physician versus CDSS of patients with Pap reports to develop a computerized clinical decision support system for cervical cancer screening that can interpret free-text Pap reports.	Qualitative	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Evaluation revealed that the CDSS outputs the optimal screening recommendations for 73 out of 74 test patients and it identified 2 cases for gynecology referral that were missed by the physician. The CDSS aided the physician to amend recommendations in 6 cases.
Wagholikar et al 2013[66]	Diagnosis	Comparing cervical cancer screening of care providers versus the CDSS of patients who had visited the Mayo clinic Rochester in March 2012 to ensure deployment readiness of the system.	Accuracy	87	n/a	n/a	n/a	n/a	n/a	n/a	When the deficiencies were rectified, the system generated optimal recommendations for all failure cases, except one with incomplete documentation.
Watson et al 2011[62]	Diagnosis	Comparing the model versus reading and evaluating patient characteristics in the EHR notes of patients who are discharged with a principal diagnosis of HF to examine psychosocial characteristics as a predictor to heart failure to reduce hospital readmissions	Sensitivity and specificity	n/a	>80	>80	n/a	n/a	n/a	n/a	Detection of 5 characteristics that were associated with an increased risk for hospital readmission
Yang et al 2018[32c]	Diagnosis	Comparing 4 machine learning algorithms, as well as our proposed model of patients with multiple diseases to assist diagnosis	Accuracy, recall, precision and F-score	98.67	96.02	n/a	95.94	n/a	95.96	n/a	–
Zhou et al 2015[71]	Diagnosis	Comparing the golden standard (manual review) versus tool of patients with an history of ischemic heart disease and hospitalized to identify patients with depression	Sensitivity, specificity and PPV	n/a	HC: 92.4 IC: 77.4	n/a	HC: 86.9 IC: 64.9	n/a	HC: 89.6 IC: 70.6	n/a	–

A: accuracy; AIS: acute ischemic stroke; ASD: automated symptom detection; AUC: Area Under the Curve; CAP: community acquired pneumonia; CC: completed colonoscopies; CDSS: clinical decision support system; Co k: Cohen's kappa; CS: colonoscopy status; CT: computed tomography; CXR: Chest X ray; DNN: deep neural network; DL: document level; ED: Emergency Department; EHR: Electronic Health Record; EMR: electronic medical record; ERPR: external peer review program; F: F-score; HC: high confidence; HF: heart failure; IC: intermediate confidence; ICD-9: International Statistical Classification of Diseases and Related Health Problems; IVT: intravenous thrombolytic therapy; kNN: k-nearest neighbor; L: location; LVEF: left ventricular ejection fraction; LVSF: left ventricular systolic function; MMTx: MetaMap Transfer; MRI: magnetic resonance imaging; MRSA: methicillin-resistant *Staphylococcus aureus*; *N*: number; NICU: Neonatal Intensive Care Unit; NLP: natural language processing; NPV: negative predictive value; Pap: Papanicolaou; PL: phrase level; PPV: positive predictive value; Sens: sensitivity; QL: qualitative; QT: quantitative; RF: random forest; RS: reference standard; S: size; SDA: symptom detection with assertion; SESCAM: Servicio de Salud de Castilla—La Mancha; Spec: specificity; TB: tuberculosis; TR: timing references; UPMC: University of Pittsburgh Medical Center; USMSTF: US Multisociety Task Force on Colorectal Cancer; VA: veterans affairs.

[a]These F-scores were calculated according to the formula: 2*(sensitivity * PPV)/(sensitivity + PPV).

[b]These studies were implemented in clinical practice.

[c]These studies were not performed in an English speaking country.

quantitative outcomes of the studies varied substantially. Overall, the studies indicate that TM-CDS combinations can increase patient safety, decrease time to diagnosis, and suggest the best therapy for a patient.

The lack of focus on medication errors ($n = 0$) and cancer diagnosis ($n = 4$) in TM-CDS studies is surprising due to the focus on these issues in TM literature and the fact that both are leading causes of death that are particularly complex and costly in many countries.[1,21] Similarly, a recent review by Jiang et al found that the primary disease concentration area for artificial intelligence (including NLP and other computational techniques) in health care was cancer, neurology, and cardiology. An opportunity exists to study the contribution of the TM-CDS combination in making a diagnosis in these fields.[72]

Only 53% of the studies included in diagnosis utilized reports. This was not in line with our expectations, because the usage of reports is logical due to the semi-structured data they consist of and their primary diagnostic purpose. This is substantiated with the number of publications in TM and the fact that radiologists are progressive in utilizing technological solutions (eg, automated dictation). Even more, the growing importance of structured data is reflected in radiologists' increasing embrace of structured reporting, standardized coding systems, ontologies, and common data elements.[73]

Barriers to implementation

A striking finding of this review is that, despite the benefits and local successes of the TM-CDS combination, research has not led to wide implementation and integration in clinical practice. A primary limitation of TM-CDS combinations mentioned is the complexity of natural language. The performance of any NLP system is constrained by the quality of the human-composed text.[70] Basic information is often inconsistently entered by humans. As clinical text repositories grow, these repositories will increasingly include conflicting data, which poses a challenge to any NLP system.[70,71]

The presence of a functioning system does not ensure it will be adopted by users. For example, Wagholikar et al (2013) concluded that use of an unfamiliar interface led to participants' mistakes, which in turn can lead to a low adoption rate despite the positive effects of technological advances, such as EHRs.[52,74] A 2016 review by Kruse et al found that physicians face a range of barriers to EHR implementation, including complexity of the system, which can lead to mistakes.[75]

A formal standard for TM techniques has not yet been established, leading to the utilization of diverse techniques at different levels and different performance outcomes, which makes these techniques hard to compare. For example, Stultz et al[44] uses keyword extraction, whereas Meystre and Haug[37] uses multiple preprocessing steps and extraction. In addition to different TM techniques, there are different CDS systems. These systems should be developed by the "5 rights", meaning to give the right information, to the right person, in the right format, through the right channel, at the right time in workflow.[76,77] However, it is technically difficult to actively provide the right data at the right time to the right person. Unlike most of the studies in this review, Rosenthal et al gave active advice using a pop-up alert and lightbulb icon to alert the professional. This increased the number of cases identified, but the compliance of the guidelines did not change.[51] One reason is that the system's advice often comes too late for professionals (eg, after the appointment with the patient), which has contributed to negative associations, compliance issues, and lack of acceptance of the TM-CDS combination.[78]

Fourteen studies in this review utilized free-text narrative documents. These documents contain all patient information that is of interest to professionals, but are complex because they contain medical terms, abbreviations, acronyms, local dialect, and lack of proper punctuation. This makes it difficult to extract data and interpret the free-text using the available TM-CDS tools. In addition, most of the studies used only one free-text document type. Utilizing the complete EHR or multiple types of free-text documents from the EHR leads to a more complex algorithm (eg, requiring multiple preprocessing steps). Future studies should include multiple types of free-text documents from the EHR or the complete EHR to represent all known information and develop algorithms that can process this information.

The specific language used by a tool presents another obstacle because the complexity of natural language differs between languages. Some languages present more difficulties in their semantic and morphological components than others. English is the dominant language of TM, but studies have also been conducted in Spanish, Dutch, and German. Therefore it is necessary to propose approaches to TM-CDS tools for clinical texts for languages other than English, as proposed by Reyes-Ortiz et al.[79]

Study strengths and limitations

This review was conducted in accordance with the PRISMA statement to ensure the use of appropriate methods. Several of the recurrent strengths and weaknesses of specific articles have already been discussed. Additional strengths include the evaluation of TM algorithms and CDS performance. Potential areas for bias in this review include the search process, development of exclusion criteria, assembling of the review, and publication. All efforts to minimize bias were made whenever possible. It proved challenging to assess the quality of the studies within this review because relevant formal standards and comparable outcomes have not been established for TM algorithms. Additional limitations include small samples of patients or texts, multiple synonyms of TM, and lack of a true comparative evaluation of the TM algorithm or CDS used in each study to other methods.

Directions for future research

The TM-CDS combination offers potential as an effective system that gives the right information to healthcare specialists at the right time. Therefore, a relevant formal standard of TM-CDS combinations that provides active advice should be created and TM should be integrated into CDS systems.

If a formal standard for TM-CDS combination that provide active advice were to be developed, it would still be difficult to implement such systems due to the low adoption rate. Boonstra and Broekhuis suggested that the implementation of any technology should be treated as a change project and led by implementers or change managers in medical practices to reduce barriers.[74] Another step to improve the adoption rate would be to adopt a white-box approach, which provides feedback to decision makers and shows them how the tools works.[21,80]

The impact of knowledge discovery on professionals' workload and time is unclear because 77% of the studies included in this review did not use the TM-CDS combination in clinical practice.[81] Future studies should consider integrating the TM-CDS in one system and exploring its effect on work environments. In addition, future research should combine TM with expert opinions from specific domains (ie, oncologists for a cancer study). Most of the articles in this review did not utilize expert opinion in any form, which increases the risk of bias.

In addition, there are limitations to the generalizability of TM-CDS combination. Open sharing of EHR free-text may be impossible due to privacy laws that restrict the sharing of patient health information; however, researchers can continue to develop and use generalized, open-source EHR-related TM systems such as the REX tool and make these TM algorithms available on platforms such as GitHub or they can utilize other frequently used free-text records, like Google or Twitter.[82,83] Making these algorithms available would support the transparency and replicability of study findings and minimize duplicate efforts. This approach is described in a recent systematic review by Koleck et al (2019).[84] Future research should address the issue of replicability, the suitability of technologies, and the usability of these technologies in medical documentation.[40]

Identifying medication errors or adverse drug reactions are important issues in the medical field[85,86]; however, none of the tools in this review were used to identify them. Some studies have used TM to identify adverse drug reactions, but not in combination with a CDS system that could provide active advice to the physician. Future studies should consider combining TM and CDS systems to identify medication errors or adverse drug reactions.[86,87]

CONCLUSION

This review presents a comprehensive collection of representative works from the field of the TM-CDS combinations. All selected publications indicate that the combination may be used to improve diagnostic and therapeutic processes in clinical practice, thus potentially contribute to more efficient, better, and safer healthcare. However, the combination has limitations similar to the respective individual limitations of TM and CDS. Additionally, the adoption rate of these tools among professionals and their use in clinical practice remain low. Furthermore, this review discusses barriers to implement the TM-CDS combination in the medical field. Further research, implementation, and integration of TM into CDS are necessary to understand its impact in daily usage and to ensure that such tools provide relevant information to professionals at the right time.

FUNDING

This work was supported by the e/MTIC research program Medtech solutions for Earlier Detection of CArdiovascular Disease (MEDIC-AID) project.

AUTHOR CONTRIBUTIONS

All authors contributed significantly to this work. BBT, RGS, ABM, AWZ, TES, and EKN conceptualized the study. BBT and EKN searched for and retrieved relevant articles and analyzed data. BBT, BDD, and AWZ interpreted the data. BBT drafted the manuscript, and RGS, ABM, AWZ, TES, BDD, and EKN made substantive revisions to the manuscript. All authors gave final approval of and accept accountability for the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

1. Kohn L, Corrigan J, Donaldson M. *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academies Press; 2000.
2. Radley DC, Wasserman MR, Olsho LEW, Shoemaker SJ, Spranca MD, Bradshaw B. Reduction in medication errors in hospitals due to adoption of computerized provider order entry systems. *J Am Med Inform Assoc* 2013; 20 (3): 470–6.
3. Prgomet M, Li L, Niazkhani Z, Georgiou A, Westbrook JI. Impact of commercial computerized provider order entry (CPOE) and clinical decision support systems (CDSSs) on medication errors, length of stay, and mortality in intensive care units: a systematic review and meta-analysis. *J Am Med Inform Assoc* 2017; 24 (2): 413–22.
4. James B. Making it easy to do it right. *N Engl J Med* 2001; 345 (13): 991–3.
5. Agrawal A. Medication errors: prevention using information technology systems. *Br J Clin Pharmacol* 2009; 67 (6): 681–6.
6. Jia P, Zhang L, Chen J, Zhao P, Zhang M. The effects of clinical decision support systems on medication safety: an overview. *PLoS One* 2016; 11 (12): e0167683.
7. Charles K, Cannon M, Hall R, Coustasse A. Can utilizing a computerized provider order entry (CPOE) system prevent hospital medical errors and adverse drug events? *Perspect Health Inf Manag* 2014; 11 (Fall): 1b.
8. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. *J Healthc Eng* 2018; 2018 (5): 4302425–9.
9. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020; 3 (1): 1–10.
10. Kubben P, Dumontier M, Dekker A, eds. *Fundamentals of Clinical Data Science*. Cham (CH): Springer; 2019: 153–71.
11. Sim I, Gorman P, Greenes R, *et al.* Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc* 2001; 8 (6): 527–34.
12. Eppenga WL, Derijks HJ, Conemans JMH, Hermens WAJJ, Wensing M, De Smet PAGM. Comparison of a basic and an advanced pharmacotherapy-related clinical decision support system in a hospital care setting in the Netherlands. *J Am Med Inform Assoc* 2012; 19 (1): 66–71.
13. Hunt D, Haynes B, Hanna S, *et al.* Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *J Heal Serv Res Policy* 1998; 280 (15): 1339–46.
14. Shortliffe EH. Computer programs to support clinical decision making. *JAMA* 1987; 258 (1): 61–6.
15. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005; 330 (7494): 765–8.
16. Garg AX, Adhikari NKJ, McDonald H, *et al.* Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *J Am Med Assoc* 2005; 293 (10): 1223–38.

17. De Bie AJR, Mestrom E, Compagner W, *et al.* Intelligent checklists improve checklist compliance in the intensive care unit: a prospective before-and-after mixed-method study. *Br J Anaesth* 2021; 126 (2): 404–14.
18. Dexter PR, Perkins S, Overhage JM, Maharry K, Kohler RB, McDonald CJ. A computerized reminder system to increase the use of preventive care for hospitalized patients. *N Engl J Med* 2001; 345 (13): 965–70.
19. Morris AH, Stagg B, Lanspa M, *et al.* Enabling a learning healthcare system with automated computer protocols that produce replicable and personalized clinician actions. *J Am Med Inform Assoc* 2021; 28 (6): 1330–44.
20. Aaron S, McEvoy DS, Ray S, Hickman T-TT, Wright A. Cranky comments: detecting clinical decision support malfunctions through free-text override reasons. *J Am Med Inform Assoc* 2019; 26 (1): 37–43.
21. Pereira L, Rijo R, Silva C, Martinho R. Text mining applied to electronic medical records: a literature review. *Int J E-Health Med Commun* 2015; 6 (3): 1–18.
22. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. *J Heal Inf Manag* 2014; 22 (3): 52–6.
23. Navathe SB, Elmasri R. *Fundamentals of Database Systems*. 6th ed. Boston, MA: Pearson Education; 2000: 205–9.
24. Nisbet R, Elder J, Miner G. Text mining and natural language processing. In: *Handbook of Statistical Analysis and Data Mining Applications*. Burlington: Elsevier Academic Press; 2009: 174–5.
25. Patel FN, Soni NR. Text mining: a brief survey. *Int J Adv Comput Res* 2012; 2 (6): 243–8.
26. Combi C, Zorzi M, Pozzani G, Moretti U, Arzenton E. From narrative descriptions to MedDRA: automagically encoding adverse drug reactions. *J Biomed Inform* 2018; 84: 184–99.
27. Elkin PL, Froehling D, Wahner-Roedler D, *et al.* NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA Annu Symp Proc* 2008; 2008: 172–6.
28. Li Z, Xing X, Lu B, Zhao Y, Li Z. Early prediction of 30-day ICU readmissions using natural language processing and machine learning. *BSI* 2019; 4 (3): 22–6.
29. Knobloch K, Yoon U, Vogt PM. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement and publication bias. *J Craniomaxillofac Surg* 2011; 39 (2): 91–2.
30. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016; 5 (1): 210.
31. Nguyen A, Hassanzadeh H, Zhang Y, O'Dwyer J, Conlan D, Lawley M. A decision support system for pathology test result reviews in an emergency department to support patient safety and increase efficiency. *Stud Health Technol Inform* 2019; 264: 729–33.
32. Yang Z, Huang Y, Jiang Y, *et al.* Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Sci Rep* 2018; 8 (1): 6329.
33. Karwa A, Patell R, Parthasarathy G, Lopez R, McMichael J, Burke CA. Development of an automated algorithm to generate guideline-based recommendations for follow-up colonoscopy. *Clin Gastroenterol Hepatol* 2020; 18 (9): 2038–45.e1.
34. Cruz NP, Canales L, Muñoz JG, Pérez B, Arnott, I. Improving adherence to clinical pathways through natural language processing on electronic medical records. *Stud Health Technol Inform* 2019; 264: 561–5.
35. Wadia R, Shifman M, Levin FL, *et al.* A clinical decision support system for monitoring post-colonoscopy patient follow-up and scheduling. *AMIA Jt Summits Transl Sci Proc* 2017; 2017: 295–301.
36. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. *Int J Med Inform* 2008; 77 (9): 602–12.
37. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform* 2006; 39 (6): 589–99.
38. Sung S-F, Chen K, Wu DP, Hung L-C, Su Y-H, Hu Y-H. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study. *Int J Med Inform* 2018; 112: 149–57.
39. Chen JH, Goldstein MK, Asch SM, Mackey L, Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *J Am Med Inform Assoc* 2017; 24 (3): 472–80.
40. Kivekäs E, Kinnunen UM, Paananen P, Kälviäinen R, Haatainen K, Saranto K. Functionality of triggers for epilepsy patients assessed by text and data mining of medical and nursing records. *Stud Health Technol Inform* 2016; 225: 128–32.
41. Kalra A, Chakraborty A, Fine B, Reicher J. Machine learning for automation of radiology protocols for quality and efficiency improvement. *J Am Coll Radiol* 2020; 17 (9): 1149–58.
42. Raja AS, Pourjabbar S, Ip IK, *et al.* Impact of a health information technology-enabled appropriate use criterion on utilization of emergency department CT for renal colic. *AJR Am J Roentgenol* 2019; 212 (1): 142–5.
43. Shen L, Wright A, Lee LS, Jajoo K, Nayor J. Clinical decision support system, using expert consensus-derived logic and natural language processing, decreased sedation-type order errors for patients undergoing endoscopy. *J Am Med Inform Assoc* 2021; 28 (1): 95–103.
44. Stultz JS, Taylor P, McKenna S. Assessment of different methods for pediatric meningitis dosing clinical decision support. *Ann Pharmacother* 2019; 53 (1): 35–42.
45. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. *AMIA Annu Symp Proc* 2008; 2008: 207–11.
46. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annu Symp Proc* 2006; 2006: 269–73.
47. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. *Proc AMIA Annu Fall Symp* 1996: 542–6.
48. Mendonça EA, Haas J, Shagina L, Larson E, Friedman, C, EM. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform* 2005; 38 (4): 314–21.
49. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. *Nat Lang Eng* 1995; 1 (1): 83–108.
50. Raja AS, Ip IK, Prevedello LM, *et al.* Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. *Radiology* 2012; 262 (2): 468–74.
51. Rosenthal B, Skrbin J, Fromkin J, *et al.* Integration of physical abuse clinical decision support at 2 general emergency departments. *J Am Med Inform Assoc* 2019; 26 (10): 1020–9.
52. Jones BE, Ferraro JP, Haug P, *et al.* Performance of a real-time electronic screening tool for pneumonia. *Am J Respir Crit Care Med* 2012; 185: A5136.
53. Evans RS, Benuzillo J, Horne BD, *et al.* Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. *J Am Med Inform Assoc* 2016; 23 (5): 872–8.
54. Day S, Christensen LM, Dalto J, Haug P. Identification of trauma patients at a level 1 trauma center utilizing natural language processing. *J Trauma Nurs* 2007; 14 (2): 79–83.
55. Smith JC, Spann A, McCoy AB, *et al.* Natural language processing and machine learning to enable clinical decision support for treatment of pediatric pneumonia. *AMIA Annu Symp Proc* 2021; 2020: 1130–9.
56. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: a system for detecting and classifying encounter-based clinical events in any electronic medical record. *J Am Med Inform Assoc* 2005; 12 (5): 517–29.
57. Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. *J Biomed Inform* 2016; 62: 224–31.
58. Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int J Med Inform* 2014; 83 (12): 983–92.
59. Matheny ME, FitzHenry F, Speroff T, *et al.* Detection of infectious symptoms from VA emergency department and primary care clinical documentation. *Int J Med Inform* 2012; 81 (3): 143–56.

60. Garvin JH, Kim Y, Gobbel GT, *et al.* Automating quality measures for heart failure using natural language processing: a descriptive study in the Department of Veterans Affairs. *JMIR Med Inf* 2018; 6 (1): e5.
61. Denny JC, Peterson JF, Choma NN, *et al.* Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010; 17 (4): 383–8.
62. Watson AJ, O'Rourke J, Jethwani K, *et al.* Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. *Psychosomatics* 2011; 52 (4): 319–27.
63. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clin Gastroenterol Hepatol* 2013; 11 (6): 689–94.
64. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol* 2014; 12 (7): 1130–6.
65. Wagholikar KB, MacLaughlin KL, Henry MR, *et al.* Clinical decision support with automated text processing for cervical cancer screening. *J Am Med Inform Assoc* 2012; 19 (5): 833–9.
66. Wagholikar KB, MacLaughlin KL, Kastner TM, *et al.* Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. *J Am Med Inform Assoc* 2013; 20 (4): 749–57.
67. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* 2008; 15 (5): 601–10.
68. Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. *Proc AMIA Symp* 2001: 12–6.
69. Fiszman M, Chapman WW, Aronsky D, Scott Evans R, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc* 2000; 7 (6): 593–604.
70. Kim Y, Garvin J, Goldstein MK, Meystre SM. Classification of contextual use of left ventricular ejection fraction assessments. *Stud Health Technol Inform* 2015; 216: 599–603.
71. Zhou L, Baughman AW, Lei VJ, *et al.* Identifying patients with depression using free-text clinical documents. *Stud Health Technol Inform* 2015; 216: 629–33.
72. Jiang F, Jiang Y, Zhi H, *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230–43.
73. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS. Toward complete structured information extraction from radiology reports using machine learning. *J Digit Imaging* 2019; 32 (4): 554–64.
74. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. *BMC Health Serv Res* 2010; 10: 231.
75. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to electronic health record adoption: a systematic literature review. *J Med Syst* 2016; 40 (12): 252.
76. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. *J Am Med Inform Assoc* 2007; 14 (2): 141–5.
77. Osheroff J, Pifer E, Teich J, Sittig D, Jenders R. *Improving Outcomes with Clinical Decision Support: An Implementer's Guide*. Chicago: Health Information Management and Systems; 2005.
78. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc* 2011; 18 (3): 327–34.
79. Reyes-Ortiz JA, Gonzalez-Beltran BA, Gallardo-Lopez L. Clinical decision support systems: a survey of NLP-based approaches from unstructured data. In: proceedings – 2015 26th international workshop on database and expert systems applications (DEXA). Institute of Electrical and Electronics Engineers Inc; 2015: 163–7.
80. Abuazab AAI, Selamat HB, Yusoff RBCM. Challenge of text mining in clinical decision support system: review. *J Eng Appl Sci* 2017; 12 (20): 5261–73.
81. Islam M, Hasan M, Wang X, Germack H, Noor-E-Alam M. A systematic review on healthcare analytics: application and theoretical perspective of data mining. *Healthcare* 2018; 6 (2): 54.
82. Masino AJ, Forsyth D, Fiks AG. Detecting adverse drug reactions on Twitter with convolutional neural networks and word embedding features. *J Healthc Inform Res* 2018; 2 (1–2): 25–43.
83. MacKinlay A, Aamer H, Yepes AJ. Detection of adverse drug reactions using Medical named entities on Twitter. *AMIA. Annu Symp Proc* 2017; 2017: 1215–24.
84. Koleck TA, Dreisbach C, Bourne PE, Bakken S. 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.
85. Leendertse AJ, Egberts ACG, Stoker LJ, Van Den Bemt PMLA; HARM Study Group. Frequency of and risk factors for preventable medication-related hospital admissions in the Netherlands. *Arch Intern Med* 2008; 168 (17): 1890–6.
86. Wasylewicz A, van de Burgt B, Weterings A, *et al.* Identifying adverse drug reactions from free-text electronic hospital health record notes. *Br J Clin Pharmacol* 2022; 88 (3): 1235–45.
87. Haq HU, Kocaman V, Talby D. Mining adverse drug reactions from unstructured mediums at scale [published online ahead of print, 2022]. https://arxiv.org/abs/2201.01405.