Deep learning in clinical natural language processing: A methodical review

Article *in* Journal of the American Medical Informatics Association · December 2019

DOI: 10.1093/jamia/ocz200

CITATIONS
21

READS
1,460

12 authors, including:

Stephen T Wu
University of Texas Health Science Center at Houston
67 PUBLICATIONS **956** CITATIONS
SEE PROFILE

Surabhi Datta
University of Texas Health Science Center at Houston
11 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Jingcheng Du
University of Texas Health Science Center at Houston
61 PUBLICATIONS **360** CITATIONS
SEE PROFILE

Journal of the American Medical Informatics Association, 0(0), 2019, 1–14
doi: 10.1093/jamia/ocz200
Review

Review

Deep learning in clinical natural language processing: a methodical review

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu

School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Corresponding Author: Stephen Wu, PhD, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA (wu.stephen.t@gmail.com)

Received 3 July 2019; Revised 15 October 2019; Editorial Decision 18 October 2019; Accepted 9 November 2019

ABSTRACT

Objective: This article methodically reviews the literature on deep learning (DL) for natural language processing (NLP) in the clinical domain, providing quantitative analysis to answer 3 research questions concerning methods, scope, and context of current research.

Materials and Methods: We searched MEDLINE, EMBASE, Scopus, the Association for Computing Machinery Digital Library, and the Association for Computational Linguistics Anthology for articles using DL-based approaches to NLP problems in electronic health records. After screening 1,737 articles, we collected data on 25 variables across 212 papers.

Results: DL in clinical NLP publications more than doubled each year, through 2018. Recurrent neural networks (60.8%) and word2vec embeddings (74.1%) were the most popular methods; the information extraction tasks of text classification, named entity recognition, and relation extraction were dominant (89.2%). However, there was a "long tail" of other methods and specific tasks. Most contributions were methodological variants or applications, but 20.8% were new methods of some kind. The earliest adopters were in the NLP community, but the medical informatics community was the most prolific.

Discussion: Our analysis shows growing acceptance of deep learning as a baseline for NLP research, and of DL-based NLP in the medical community. A number of common associations were substantiated (eg, the preference of recurrent neural networks for sequence-labeling named entity recognition), while others were surprisingly nuanced (eg, the scarcity of French language clinical NLP with deep learning).

Conclusion: Deep learning has not yet fully penetrated clinical NLP and is growing rapidly. This review highlighted both the popular and unique trends in this active field.

Key words: deep learning, natural language processing, electronic health records, methodical, review, clinical text

INTRODUCTION

Technical research is changing rapidly. Deep learning (DL) techniques have begun to dominate because of their simplicity (no need for handcrafted features), efficient processing (assuming dedicated, massively parallelized hardware), and state-of-the-art results (on a plethora of tasks). Meanwhile, the widespread adoption of electronic health records (EHRs) has produced massive amounts of digital text concerning patients, and the medical informatics community has invested substantial effort to make use of clinical text via natural language processing (NLP). Furthermore, research manuscripts themselves are coming under greater scrutiny for rigor and reproducibility,[1–3] yet they are simultaneously being generated on preprint servers with more momentum and less oversight than ever before.

This work aims to characterize the relationship between DL techniques and the field of clinical NLP, in today's wider landscape of technical research, through a methodical review of the literature. We seek to answer the following research questions:

1. RQ1: Methods. What deep learning methods are being contributed or applied?
2. RQ2: Scope. What types of problems are addressed and solved?
3. RQ3: Context. How do these articles fit into the wider research context?

To answer these questions and draw out other insights, our study methodically considers 25 variables across 212 articles from a variety of venues, mostly published between 2014 and April 2019. Notable data-based observations from this study include:

- Publications on DL in clinical NLP are more than doubling each year, through 2018.
- The majority of this literature uses existing DL models on well-known information extraction tasks in English clinical notes, but there are many exceptions.
- There is growing acceptance of DL as the baseline for NLP research and of DL-based NLP in the medical community.

Working definitions

Deep learning is a modern, popular paradigm for machine learning (ML) heralded for avoiding the extensive manual feature engineering that was common in traditional methods. For the purposes of this study, **deep learning** consists of neural network-based or -inspired methods that utilize modern optimization techniques and training objectives. For example, recurrent neural networks (RNNs) are frequently employed to model sequential data such as language; convolutional neural networks (CNNs) are often used to model signals such as images. We include **embeddings**—dense, data-driven vectorial representations of, for example, a word—under the deep learning umbrella. Embeddings serve as the input layer of most modern neural networks; some embeddings are directly created as part of larger neural networks, while other embedding methods have none of the nonlinearities that are characteristic of neural networks. Finally, though we do not consider older neural networks (eg, multilayered perceptron) to be "deep learning," we nonetheless include them in our study.

We take a broad view of **natural language processing** techniques, namely, any work that computationally represents, transforms, or utilizes text (or speech) and its derivatives. Thus, diverse tasks can be viewed as NLP activities; from producing dependency parses, to text-based event prediction, to image classification via captions. However, our study only considers manuscripts with NLP in the specific clinical setting of **electronic health records**—digital profiles of patients' health, primarily authored by health professionals and administrators. Other health-related settings such as social media, web forums, and messaging platforms differ vastly from EHRs in linguistic profile and data availability and were thus excluded.

Finally, unlike some recent work,[4–6] we have titled this work a methodical review, by which we mean that our work follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines closely, but not completely. Because this work (and with it, most computing literature) does not qualify as a traditional Cochrane-style "systematic review," we have chosen to consider this a "methodical" review.

Related work

While this work emphasizes methods, especially in deep learning, we do not thoroughly discuss the underlying models. Instead, we refer the reader to other treatments thereof, such as Shickel et al's taxonomy of DL methods[7] or classic works on embeddings,[8,9] RNNs,[10,11] CNNs,[12,13] attention,[14,15] and adversarial learning.[16,17]

Deep learning in a clinical NLP is an active and multidisciplinary area of research, and has thus spawned numerous other review articles, as shown in Table 1. Of note, Dreisbach et al had a similar technical focus but overviewed symptom extraction techniques and used patient-authored texts, rather than our focus of all NLP tasks on clinician-authored texts in the EHR. Shickel et al wrote in the context of EHR data, and as such also provided an informative enumeration of important clinical NLP tasks such as representation learning. In contrast with their methodical conceptual survey, our work offers a PRISMA-like review with quantitative analysis and focuses exclusively on NLP. Wang et al performed a thorough methodical review on NLP from EHRs. However, their eligible articles only included work up to September 6, 2016, before DL was really adopted as mainstream in the informatics community (see Figure 5a). Their work is also exclusively focused on applications of NLP, whereas this work also considers primarily methodological contributions. Xiao et al also methodically reviewed DL literature in EHRs up until January 30, 2018, categorizing the tasks involved, deep learning techniques, and the associated challenges. Our work is similar but provides a narrower focus on NLP, more quantitative analysis on a larger number of updated articles through 2019, and consideration of contemporary factors, such as preprints and scientific rigor.

MATERIALS AND METHODS

Our review adheres as closely as possible to the PRISMA guidelines with most analyses considering categorical variables rather than the results of the component articles. The overall workflow is shown in Figure 1 and described below.

Eligibility, sources, and search

Articles eligible for inclusion in our study were characterized by: 1) natural language processing, 2) deep learning or neural networks, and 3) clinical domain tasks using EHR data. These criteria were approximated through librarian-assisted development of queries for each database (the string for Scopus is in Figure 1). Ovid MEDLINE, EMBASE, and Scopus, and the Association for Computing Machinery (ACM) Digital Library were searched on April 10, 2019. We subsequently included articles from "Other Sources." Most notably, we searched the Association for Computational Linguistics (ACL) Anthology with EHR- and DL-related keywords, retrieving 61 articles before deduplication. "Other Sources" also included free suggestions from all authors, who were given a 1-week span to submit relevant works that should be included, preserving the recall-centric style of searching.

Duplicates were removed. The remaining papers were used for screening according to our inclusion/exclusion criteria.

Individual papers' risks of bias were not a concern for our primarily count-based aggregate analyses of categorical variables. Rather, there is the typical risk of selection bias in these articles, since indexes such as MEDLINE or Scopus are not exhaustive, and search strings may have left out relevant articles.

Table 1. Comparison of review articles related to this one, with their methods and scope

Year	Authors	PRISMA Review?	DL?	NLP?	EHR?
2019	*This paper*	✓	✓	✓	✓
2018	Al-Aiad et al[18]	–	✓	Broader	✓
2018	Ching et al[19]	–	–	Broader	Broader: biology
2019	Dreisbach et al[4]	✓	✓	Narrower: IE	Patient-authored; Narrower: symptoms
2019	Esteva et al[20]	–	–	Broader	✓
2017	Gonzalez et al[21]	–	–	✓	Broader: with social media
2016	Liu et al[22]	–	–	Narrower: IE	Broader: biology
2018	Névéol et al[23]	Selection	–	✓	Broader: with social media
2018	Névéol et al[24]	–	–	✓	Non-English; Broader
2019	Sheikhalshahi et al[5]	✓	–	✓	Narrower: chronic diseases
2017	Shickel et al[7]	Methodical	✓	Broader	✓
2018	Velupillai et al[25]	–	–	✓	✓ with health outcomes
2018	Wang et al[26]	✓	–	Narrower: IE	✓
2018	Xiao et al[6]	✓	✓	Broader	✓
2018	Zeng et al[27]	–	–	✓	Narrower: computational phenotyping

Abbreviations: DL, deep learning; EHR, electronic health record; NLP, natural language processing.

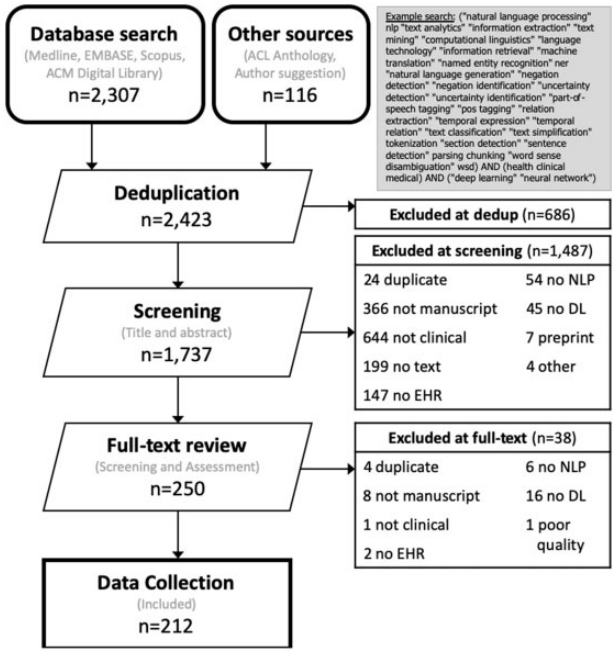

Figure 1. PRISMA flowchart for including articles in our study, with example search string (for Scopus) and primary reasons for exclusion.

Study selection

Title-abstract screening

During the screening stage, 2 randomly assigned co-authors (and 1 adjudicator) screened articles, excluding 1 487 for the ordered list of reasons shown in Figure 1. Thus, the 366 (under "Excluded at screening") that were not typical research manuscripts were already considered to *not* be duplicates.

Note that 7 preprint articles (suggested by authors) were excluded, despite the fact that some such articles (eg, BioBERT[28]) are influential in the research community at the time of this writing. A preprint is a full draft of a research paper that is shared publicly before it has been peer reviewed. Preprints can bring broad and instant visibility to research and have been widely utilized by research communities, though not without controversy. We explore the impact of preprints in RQ3.

Full-text screening

A full-text screening step, performed in conjunction with data collection, further ruled out 38 papers. Here, aside from criteria that were missed in other steps, we also excluded abstract-only publications (8 references), works with insufficient NLP or DL (6 and 16, respectively), or those that were poor in quality (1 reference). Note that we excluded out-of-the-box algorithms, which accounted for 4 of the 6 that were labeled "no NLP," and 12 of the 16 labeled "no DL"—these had no retraining of models, variation in architectures, or systematic evaluation.

Data collection

Nine co-authors extracted 25 variables regarding DL Techniques, Embeddings, Tasks, Data, Experimental Setup, Comparisons, Contributions, and Venue & Timing (see Supplementary Appendix B) from each of 212 included articles (see Supplementary Appendix A), according to a data collection form. These data were normalized, validated, and analyzed by team members.

For all analyses, note that 2019 was a partial year including an imbalanced sampling of underlying data; for example, computer science and NLP conferences tend to occur in the summer, after April 10, of any given year. Also, many of our variables allowed a single paper to have multiple responses; in the text, numbers and percentages represent portions of the 212 included articles unless otherwise indicated, while figures and tables at times utilize relative percentages.

RESULTS

RQ1: What deep learning methods are being contributed or applied?

Methods: Deep learning architectures

Figure 2 illustrates the rise and fall of broad categories of DL architectures (see Supplementary Material Table 1 for more fine-grained categories and mappings). We focus on RNNs, CNNs, feed forward neural networks, and Embeddings-only articles, where Embeddings-only papers used embeddings without neural networks in a similarity function or in a more traditional ML classifier, for example, support vector machine. Note that Figure 2 intentionally double-counts papers that combine multiple methods (n = 41, 19.3%).

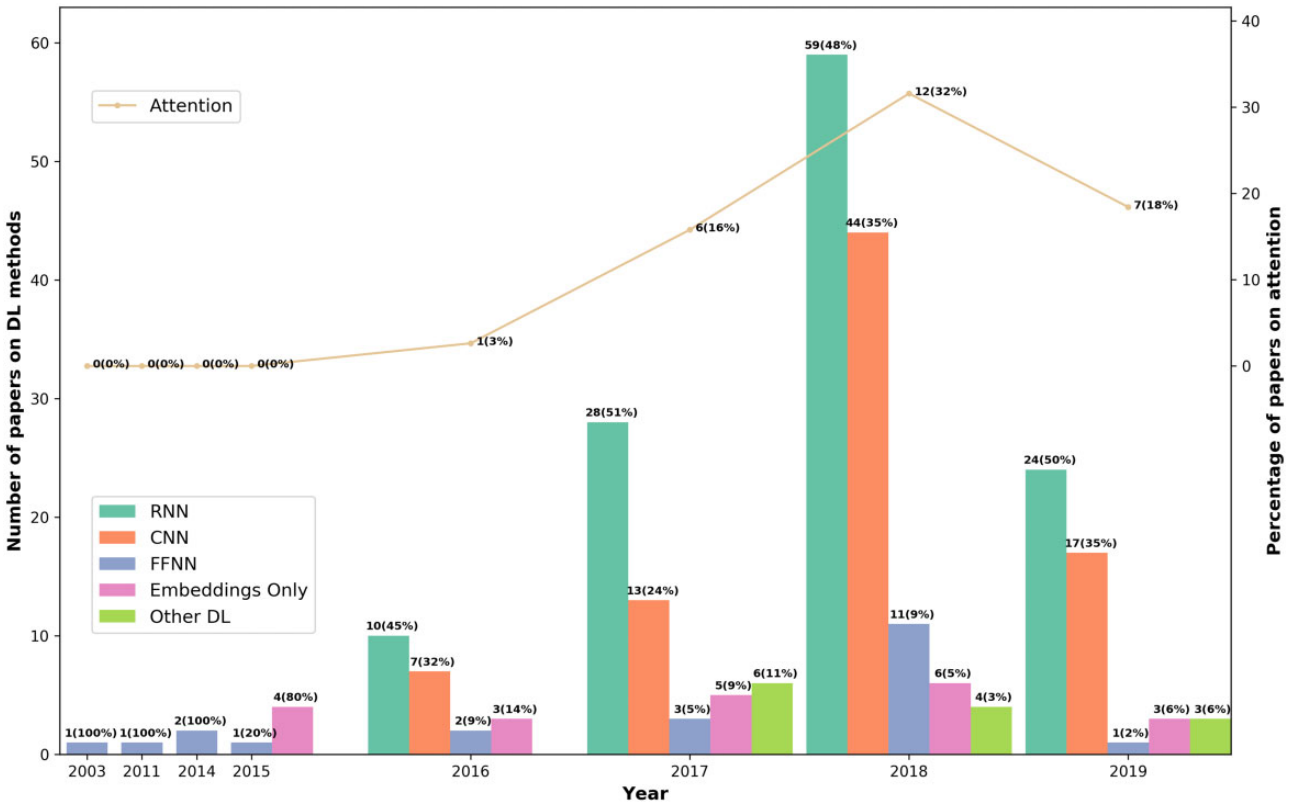

Figure 2. Growth of broad architectures in deep learning over the years. Percentages are relative to the number of studies published in that year. Overall, RNN variants were the most common (n = 129, 60.8%), CNNs were second (n = 83, 39.2%), traditional feed-forward networks were third (n = 22. 10.4%), and embeddings-only were fourth (n = 21, 9.9%).

Abbreviations: CNN, convolutional neural networks; RNN, recurrent neural networks.

The volume of DL publications is currently increasing quickly each year (>200% through 2018), and the same is true of each type of architecture (with the possible exception of "Other DL"). Across years, RNN variants were the most common (60.8%), split among memory units, such as long short-term memory (LSTM, 52.8%), gated recurrent unit (7.5%), and the "vanilla" RNN (2.4%). The second-most common architecture type was CNNs (n = 83, 39.2%). Both show consistent growth in the last few years.

Interestingly, traditional neural networks (largely multi-layer perceptrons) also grew in percentage usage (from 5% in 2017 to 9% in 2018), perhaps as a byproduct of excitement over neural methods. On the other hand, Embeddings-only papers (n = 21, 9.9%) decreased in percentage after initial interest, similar to general domain NLP. Other DL architectures employed in this study included autoencoders,[29–32] residual neural networks,[33–36] deep belief networks,[37] capsule networks,[38] memory networks,[39] seq2seq extensions,[40] and the attention-based Transformer via BERT.[41]

Attention mechanisms (in Figure 2, the line and right-hand axis), which can increase predictive performance and—debatably[42]—improve model explainability, were often combined with other methods and are increasingly popular (overall n = 26, 12.3%). We anticipate a significant future uptick in attention mechanisms due to BERT.[43] Adversarial learning, while typically popular and even motivational in natural language generation tasks,[31] was surprisingly only used by 2 papers.[38,44]

Though the norm was to repurpose or combine existing models for clinical NLP tasks, a few unique architectures were introduced that reflected their clinical task and domain particularly well. For example, Xie et al[44] used a tree-of-sequences LSTM to fit the tree structure of ICD codes, along with adversarial learning, isotonic constraints for ordering, and attention-matching. A second case is the deep averaging network used by Dligach & Miller,[45] whose input is concept-level embeddings that are simply averaged and passed to a downstream layer. This allows the model to robustly focus on sets of conditions for a variety of phenotyping tasks without overemphasizing specific context.

Methods: embeddings

Table 2 details the embedding techniques in our study. Overall, the most prominent embedding model is word2vec (74.1%, combining values in Table 2a and Table 2b), followed by GloVe (9.9%). To avoid out-of-vocabulary problems in representing clinical words, many studies combine character embeddings with word-level embeddings (13.7%), or utilize the fastText subword model (3.8%).

Especially for the tasks of concept and relation extraction, a number of studies explored additional lexical features combined with word embeddings: syntax embeddings, such as parts-of-speech[68,69] and dependency trees;[70–72] semantic embeddings, such as dictionary features,[35,73–76] controlled unclassified information from Unified Medical Language System (UMLS),[77–79] and semantic role labels;[78] and position embeddings, such as word[34,38,70,75,80–84] and section[85] positions. The number of embedding methods combined with different input features is shown in Table 2a. A few studies (Table 2b) compared multiple word embedding algorithms in

Table 2. Embedding techniques among the included articles. Due to the use of multiple approaches in individual papers, percentages overlap and may not add up

(a) Popular embedding techniques

Method	*# papers*
word2vec only	89 (42.0%)
+ character	26 (12.3%)
+ syntax	14 (6.6%)
+ position	8 (3.8%)
+ semantics	5 (2.4%)
+ 2 more features	6 (2.8%)
GloVe	17 (8.0%)
+ syntax	1 (0.5%)
+ character	3 (1.4%)
fastText	8 (3.8%)

(b) Embedding comparisons

Methods	*Reference*
word2vec vs GloVe	[46–50]
word2vec vs fastText	[51–54]

(c) Less common embeddings

Method	*Task*
Collobert[55]	NER,[56] Abbrev. Disambiguation[57]
RNNLM[58]	De-identification[59]
Starspace[60]	Representation learning[61]
VecMap[62]	Cross-lingual concept extraction[63]
ELMo[64]	NER[65–67]
BERT[43]	NER,[65] pretrained resource[41]

Abbreviation: NER, named entity recognition.

terms of extrinsic, downstream tasks, finding no significant differences between the popular methods.

Less common embedding models are reported in Table 2c, including Collobert's ranking-based embeddings;[55] the early RNN-based language model RNNLM;[58] Wu et al's[60] generalization of fastText, StarSpace; the VecMap[62] framework for learning cross-lingual word embeddings; contextual word embeddings ELMo;[64] and the recent language model, BERT.[43]

Large, unlabeled data sources are often used to train effective of word embeddings. Among the 63% of articles that reported using pre-trained resources, clinical word embeddings are built on clinical notes (29%), like MIMIC-III; health-related text from biomedical literature (25%), like PubMed; and healthcare websites (5%), like WebMD. The rest (46%) have randomly initialized the word embeddings and are trained on specific target data. Few works directly compared the effectiveness of pretraining on different resources, yet large pretraining data does not guarantee effective word embeddings for clinical NLP. Notably, in 1 study,[86] embeddings concatenating domain-specific with domain-agnostic embeddings yielded the best results.

Methods: medical knowledge

Clinical NLP has traditionally relied heavily on medical-specific knowledge resources, such as the UMLS.[87] Interestingly, only 38 papers (17.9%) used external (ie, not entirely derived from the training samples) medical knowledge. Of these 38, only 12 papers (5.7%) incorporated that knowledge into the deep learning architecture itself. With only 1 exception,[44] these 12 produced and incorporated knowledge-resource embeddings (eg, concatenating a UMLS controlled unclassified information embedding with a word embedding). Further detail on popular knowledge resources is in the Supplementary Material Table 2.

Methods: implementation

Many papers did not clearly report the use of DL libraries (only 53.4%) or NLP tools (only 36.8%). The most popular DL libraries in our study are TensorFlow and Keras, with increasing trends starting from 2015. The most frequently used NLP tools were Gensim (mostly for word-level or paragraph-level embeddings), cTAKES, NLTK, MetaMap, Stanford CoreNLP, and Jieba (a word segmentation toolkit for Chinese). Further details on the DL and NLP implementation tools can be found in Supplementary Material Table 3 and Figure 2.

We also note that the vast majority had underspecified experimental setups and did not report the type of hardware used (172 of 212, or 81.1%). Of those that did report computational equipment, a few used commercially available resources (3 of 40, 7.5%), some used local CPUs (3 of 40, 7.5%), but most used local GPUs (34 of 40, 85.0%).

RQ2: What types of problems are addressed and solved?

Scope: tasks

Figure 3a illustrates the NLP and clinical tasks addressed among the 212 papers. We categorized the main NLP tasks in each paper into 4 types: Text Classification (40.5%), NER (34.0%), Relation Extraction (13.7%), and Others (10.8%). Note that the Text Classification category is composed primarily of document-level tasks (64 of 86, 74%), but sentence-level (12 of 86, 14%) and Other (10 of 86, 12%) tasks are also present (not pictured). Figure 3a further subdivides clinical domain-specific tasks. The top clinical tasks are clinical concept extraction (ie, the extraction of common clinical concepts, such as problem, lab test, treatment, time expressions, and events); phenotyping (ie, the broad characterization of patients' conditions); and clinical relation extraction (ie, the identification of relations between the common clinical concepts). The figure also reveals a "long tail" of papers on more variegated tasks, include Information Retrieval,[86,88–90] Natural Language Generation,[31,40,91] Abbreviation Disambiguation,[57,68] Section Boundary Detection,[92,93] Shallow Parsing,[94] Machine Translation,[95] Spelling Correction,[96] and others. Overall, many common Clinical NLP tasks are present (eg, negation/attribute identification,[33,72,97,98] event detection,[99–101] and adverse drug event relation extraction[38,81,102,103]) but other tasks are infrequent or absent (sentence detection, part-of-speech tagging, and text simplification) from our study, presumably because DL techniques have yet to be applied on these tasks.

Figure 3b shows the DL algorithm distribution for each task, confirming widely held associations in the community: CNNs are the most common approach to the Text Classification task, and RNNs are the most common approach to NER and Relation Extraction.

Scope: data sources

Table 3a shows the source of corpora for the included studies. Almost half of the studies used private datasets (n = 104; 49.06%), which are rarely shared or replicated due to patient privacy concerns. Though at times used in conjunction with private datasets, publicly available data sources were employed by more than half of the papers in our study. Excluding counts of multiple public corpora in Table 3a, this is 54.7% of 212 papers, whether from research

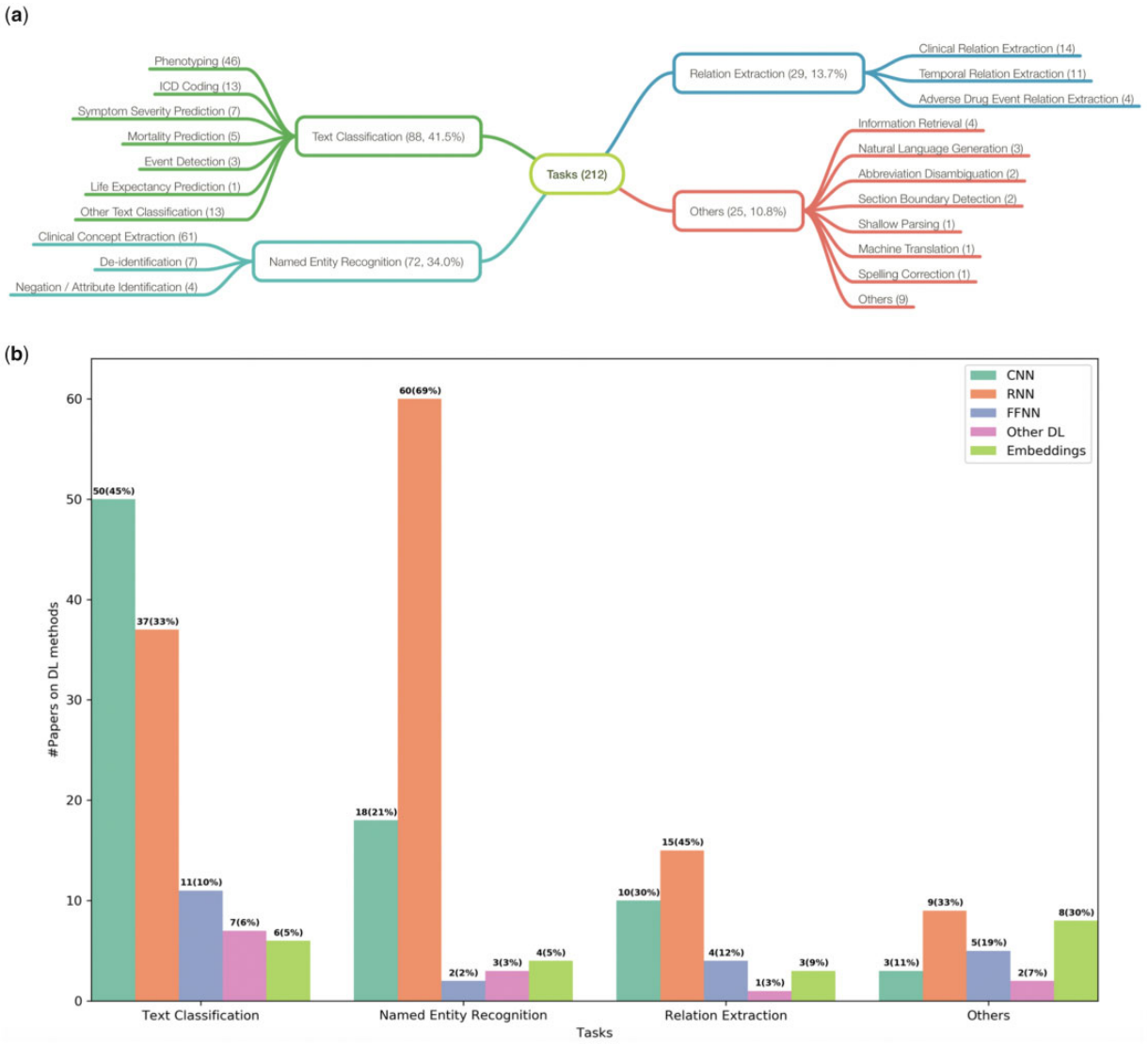

Figure 3. (a) Tasks and their prevalence in our study, from the NLP and the clinical perspective. (b) Deep learning architectures for major task groupings.

Abbreviation: NLP, natural language processing.

challenges (i2b2/n2c2, CCKS, SemEval, etc; 34.0%) or otherwise contributed by the research community (MIMIC, THYME, MEDLINE, etc; 20.8%). Considering the publicly available corpora used for each task, the most popular were: MIMIC data for text classification (15 of 86 papers); i2b2 challenges for NER/concept extraction (19 of 74 papers); i2b2 challenges for Relation Extraction (5 of 17 papers); and SemEval challenges for temporal events and relations (4 of 7 papers).

Table 3b presents some statistics on the languages of these corpora. As expected, most of the studies used datasets in English (n = 151; 71.23%), but a significant proportion utilized Chinese corpora (n = 42; 19.81%). Datasets in all the other languages (such as Spanish, Japanese, and Finnish) were used by 5 (2.36%) or fewer studies. Interestingly, this differed drastically from Névéol et al's[24] recent review of non-English Clinical NLP, especially with the reversed roles of Chinese and French. Of course, Névéol et al cannot be considered head-to-head with this work, because it had an eligibility cutoff date while DL was still nascent in the field (January 2017). However, our data suggest that deep learning-related articles were published sooner in Chinese (2016) than French (2018), and the disparity in volume is growing (see Supplementary Material Table 4).

Further details on datasets available for specific tasks, size of corpora, and subdomain of data are available in the Supplementary Material Table 5 and data-related subsections.

Scope: contributions

RQ1 (Methods) asks what types of contributions are being made. Thus, we judged each of our study articles for its contribution type: Application (to a new dataset, new domain, or setting); Methods (new DL architecture, new embedding method, new NLP task approach, or a variant of existing methods); Resource; or Evaluation; 44 articles made multiple contributions.

Pairing this with RQ3 (Context), we are interested in the novelty of articles, which is typically quite subjective. Thus, we defined a *low bar* of "novelty" as attempting methodological contributions that included new DL architecture, new embedding method, or new NLP Task approach (grouped together in the bottom center of Fig-

Table 3. (a) Types of labeled corpora used among the included articles, and their availability for 3rd party researchers. Each percentage uses 212 papers as its denominator, but, due to the use of multiple corpora in individual papers, percentages do not add up within any grouping. (b) The languages for labeled corpora, in comparison with Névéol et al's[104] reviewed papers. In 4 of our cases of non-English corpora, an English corpus was also used

(a) Source of Labeled Corpus

Availability	Corpus	Count	Percent
Private	Institutional (proprietary) dataset	104	49.1%
Challenge	i2b2 challenges	34	16.0%
	CCKS	12	5.7%
	SemEval challenges	8	3.8%
	CEGS N-GRID challenge	7	3.3%
	MADE challenge	7	3.3%
	Other	11	5.2%
Public	MIMIC data	21	9.9%
	THYME	5	2.4%
	MEDLINE case reports	2	0.9%
	MedlinePlus	2	0.9%
	Other	20	9.4%
Other	~Not reported~	3	1.4%

(b) Language of Labeled Corpus

		Count		Rank	
Language		This study	Névéol et al	This study	Névéol et al
English	71.2%	151	–	–	–
Chinese	19.8%	42	11	1	5
Spanish	2.4%	5	17	2	3
Japanese	1.9%	4	8	3	6
Finnish	1.9%	4	6	3	7
French	0.9%	2	36	5	1
Italian	0.9%	2	3	5	9
Dutch	0.5%	1	5	7	8
Thai	0.5%	1	–	7	–
German	0.5%	1	19	7	2
Swedish	0.5%	1	12	7	4
Not reported	0.5%	2	–	–	–

ure 4), while excluding "variant of existing method" Methods contributions. While there were contributions of each kind for each task, Text Classification papers were overrepresented in Applications (43% vs 36% in Methods); NER papers were overrepresented in Methods variants (40% vs 24% in new Methods and 30% in Applications); and Relation Extraction was overrepresented in new Methods (17% vs 7% in Methods variants and 8% in Applications).

Overall, there were 46 contributions of "novel" methods (excluding duplicates, this is 20.8% of papers), most introducing new DL architectures of some kind, for example, concatenating 2 new types of embeddings. The 5 papers proposing a new approach to an NLP task were related to temporal extraction/relation/inference,[105] spelling/grammar error identification and recovery,[96] computational semantics,[106] information retrieval (IR),[88] and NER.[69] The 2 papers proposing new embedding methods were associated with text classification[107] and word sense disambiguation.[57] One paper was marked for having both DL and NLP innovation: Cai et al's[69] self-attention to label part-of-speech tags and named entities.

Considering the temporal trends of these novel contributions (see Supplementary Material Figure 3), the new embeddings Methods were the earliest (2015 and 2016), whereas Methods contributions proposing a new DL architecture began in 2016; Methods papers did not take new approaches to NLP tasks until 2017. Overall, "novel" Methods contributions are perhaps the slowest growing contribution type in recent history (from 10 to 17, from 2017 to 2018). Methods variants are the fastest-growing type of contribution (an increase from 18 to 52, over the same period), Applications and Resource papers show intermediate rates of increase (from 22 to 51; and from 3 to 5, respectively).

RQ3: How do these articles fit into the wider research context?

Context: research communities

In Figure 5a, we categorized each publication venue as being a Conference or Journal, and arising from 1 of 5 communities:

1. *Computer science* (CS, 22.6%), such as NeurIPS, or AAAI;
2. *Informatics* (48.6%), such as AMIA, JAMIA, or JBI;
3. *Medical* (4.7%), such as Radiology, Drug Safety, or Nature Medicine;
4. *NLP* (18.4%), such as BioNLP, EMNLP, or ACL; or
5. *Other* (5.7%), such as PloS One, or IEEE Transactions on Nano-Bioscience.

The NLP community saw the earliest push (in 2016) in DL-related papers, but the CS and Informatics communities followed quickly (in 2017). Interestingly, medical venues seem to exhibit an increasing acceptance (beginning 2018) of deep learning, despite its reputation as a "black box" which clinical experts might hesitate to use.

Overall, more articles have been published in conferences than journals, but the differences per community demonstrate implicit community preferences; the computer science and NLP communities tend to prefer conferences, which boast quicker turnaround times and higher standards of novelty. In contrast, medical conferences are not considered full publications, so only journals are present for that community. The informatics community is a multidisciplinary blend, and has thus far also produced the largest volume of literature using DL for Clinical NLP.

Figure 5b shows what types of contributions each community has made. This illustrates relatively similar contributions from the Informatics and CS venues, though resources and evaluation are stronger in the Informatics community. As might be expected, new Methods are overrepresented in the NLP venues but underrepresented in Medical venues.

Context: preprint status

We checked the preprint status of each article to see whether it had been posted on the preprint servers arXiv (https://arxiv.org/) or bioRxiv (https://www.biorxiv.org/). Out of 212 papers, 35 were posted on arXiv, and 0 on bioRxiv. Of the 35, 9 were ultimately published in peer-reviewed journals, including 8 informatics-related journals (eg, JAMIA, JBI, BMC Medical Informatics and Decision Making). The number of preprints is also increasing per year.

Context: scientific rigor

A renewed emphasis on scientific rigor has permeated the academic establishment,[108,109] with replicability and reproducibility studies (eg, from ACM's definitions[110,111]) showing how much of the literature has fallen short of its scientific goals. While a direct study on the reproducibility and replicability of our 212 articles is beyond the scope of this work, we labeled each article for methodological implementation details that may contribute to later studies on scientific

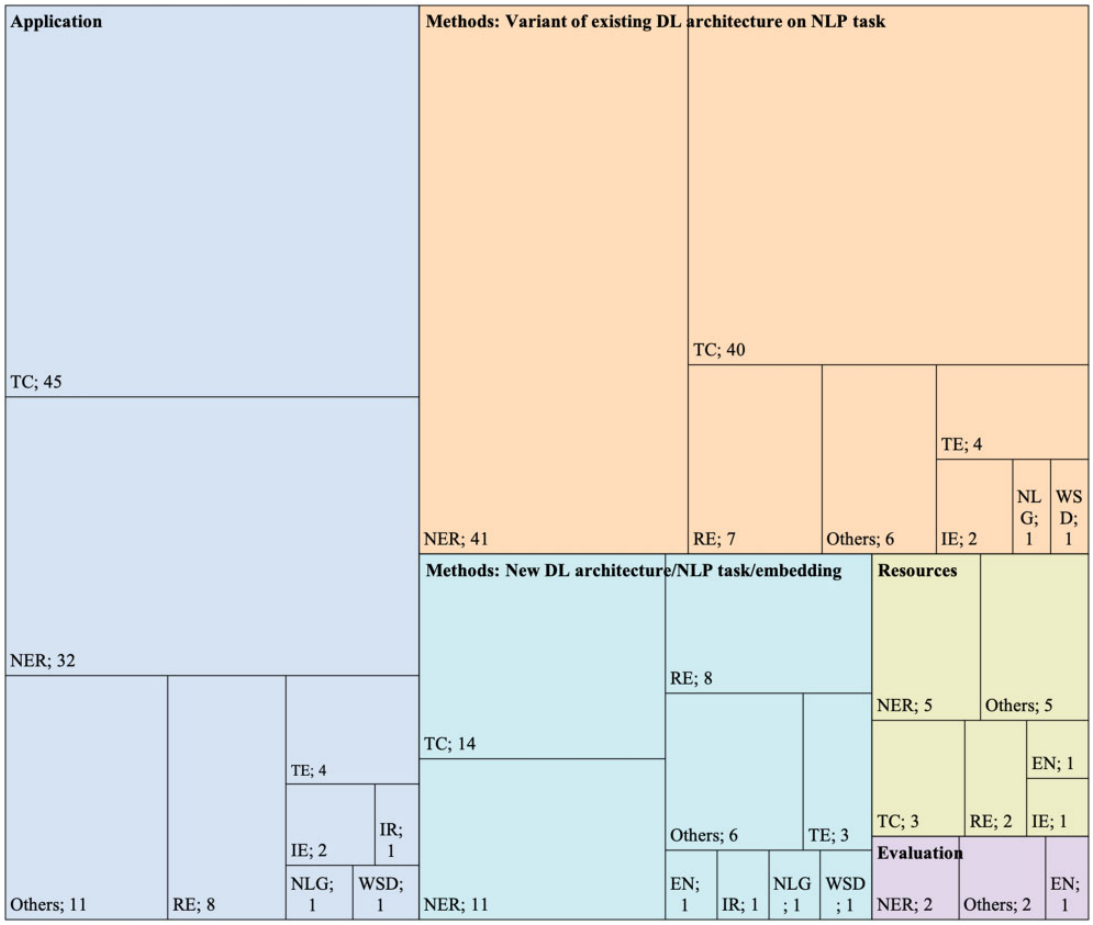

Figure 4. Number of papers on various NLP tasks per contribution category. Articles tagged with multiple contributions and/or Tasks appear more than once.

Abbreviations: EN, Entity Normalization; IE, Other Information Extraction; IR, Information Retrieval; NER, Named Entity Recognition; NLG, Natural Language Generation; NLP, natural language processing; RE, Relation Extraction; TC, Text Classification; TE, Temporal Expressions; WSD, Word Sense Disambiguation.

rigor. First, regarding the type of contributed software, we categorized papers as *Open-source* vs *Restricted* (the software exists but requires, for example, a commercial license) vs *Not provided*. Second, regarding hyperparameters (eg, number of layers in the model, the size of dimensions for each layer, the learning rate, the optimizer, dropout rate), we tagged each paper with *Not provided*, *Partially provided*, and *Present* for the 212 papers.

The results in Table 4 demonstrate that although most papers did not provide software (91.51%), many of these nonetheless offered enough hyperparameters and details (55.19%) so that other researchers can potentially reimplement their methods. An encouraging majority of papers (63.68%, summing the values in bold and/or italic in Table 4) have contributed software or offered hyperparameters.

In addition, we find that the papers whose contributions were tagged as *Method* type tend to provide more implementation details (either contributed software or presented enough hyperparameters) (66.43% = 93/140) than papers with types of *Application and Resources* (58.33% = 42/72).

Context: comparisons with traditional machine learning

Figure 6 shows that the percentage of studies comparing DL against traditional ML decreased from 2016 (70%) to 2017 (63%) to 2018 (55%) and 2019 (47%). This is likely due to the increasing acceptance of other DL algorithms as baseline models. Of the 212 papers, just over half compared their proposed methods with traditional ML methods (n = 108, 50.9%). Within these 108 studies, the majority (72%) proposed DL methods that outperformed the traditional ML methods, though there were also some negative results (11% of DL algorithms were worse than traditional ML).

DISCUSSION

Growing acceptance

There is evidence that in addition to growing in volume, DL for Clinical NLP is becoming more widely accepted. This acceptance is demonstrated in the fact that deep learning approaches are increasingly considered the baseline technique, with no need for comparison with traditional ML. Additionally, despite their genesis in the CS and NLP communities, DL-based NLP approaches have thoroughly permeated the informatics community and penetrated reputable clinical journals.

The implications are that informaticians and clinicians will increasingly be willing to adopt DL technologies in clinical settings as it becomes more familiar and widespread. This is both an opportunity and a hazard, so health professionals need to both accept and discern the associated risks appropriately.

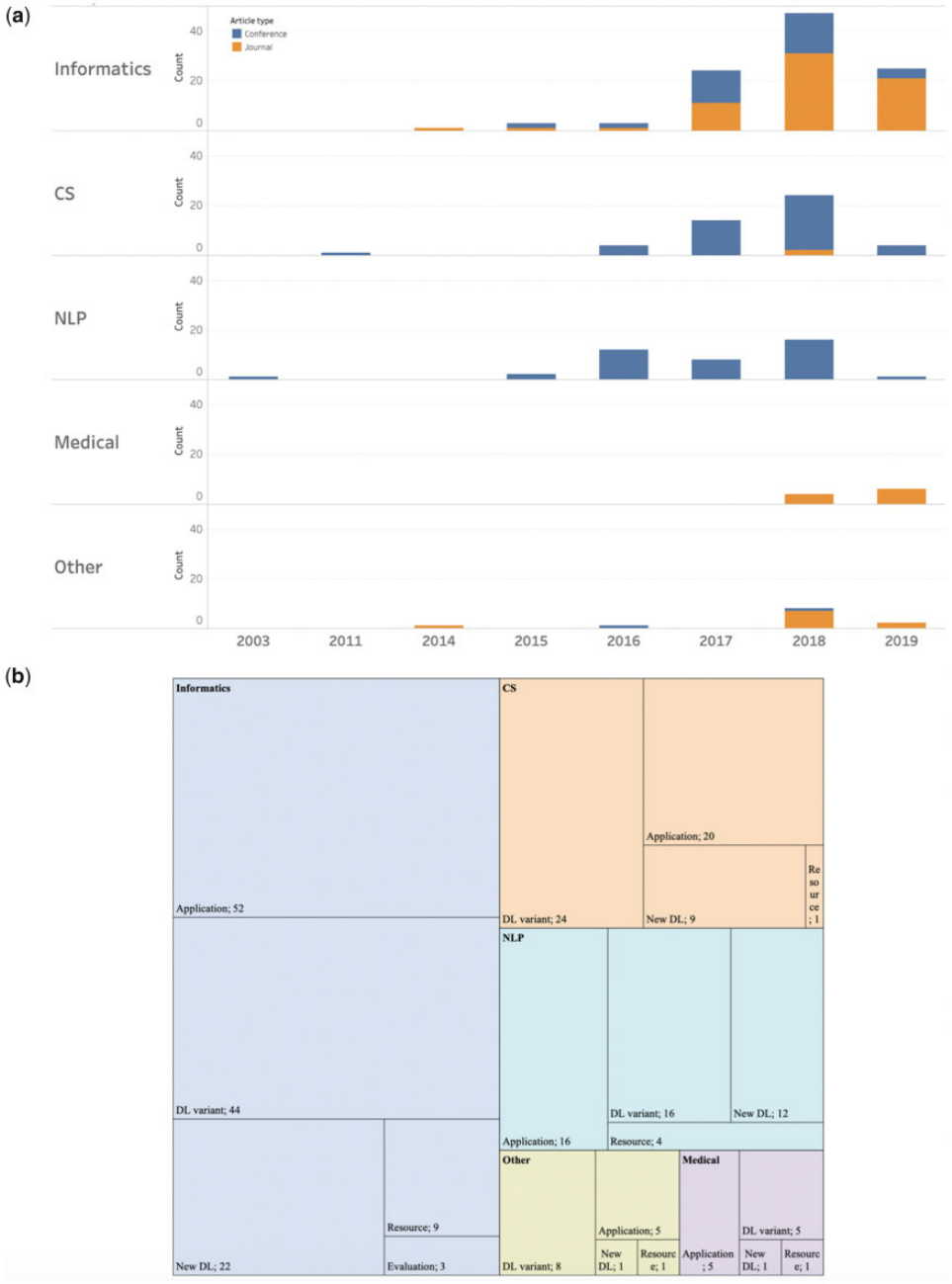

Figure 5. (a) Number of papers from different research communities and their publication types (Journal, Conference) over time; overall, the informatics community produced the highest volume of literature (48.6%), followed by CS (22.6%), followed by NLP (18.4%), followed by Others (5.66%) and Medical (4.7%). (b) Types of contributions for each community.

Abbreviations: CS, computer science; NLP, natural language processing.

Table 4. Availability of rigor-related implementation details (count and proportion)

		Hyperparameters			
		Not provided	Partial	Present	Total (row)
Software	**Not provided**	48 (22.64%)	29 (13.68%)	*117 (55.19%)*	194 (91.51%)
	Restricted	**1 (0.47%)**	**0 (0.00%)**	***3 (1.42%)***	4 (1.89%)
	Open-source	**2 (0.94%)**	**3 (1.42%)**	***9 (4.25%)***	14 (6.60%)
	Total (col)	51 (24.05%)	32 (15.1%)	129 (60.85%)	

Bolded entries indicate that articles identified the software in some way. Italicized entries indicate that articles gave a sufficient set of hyperparameters.

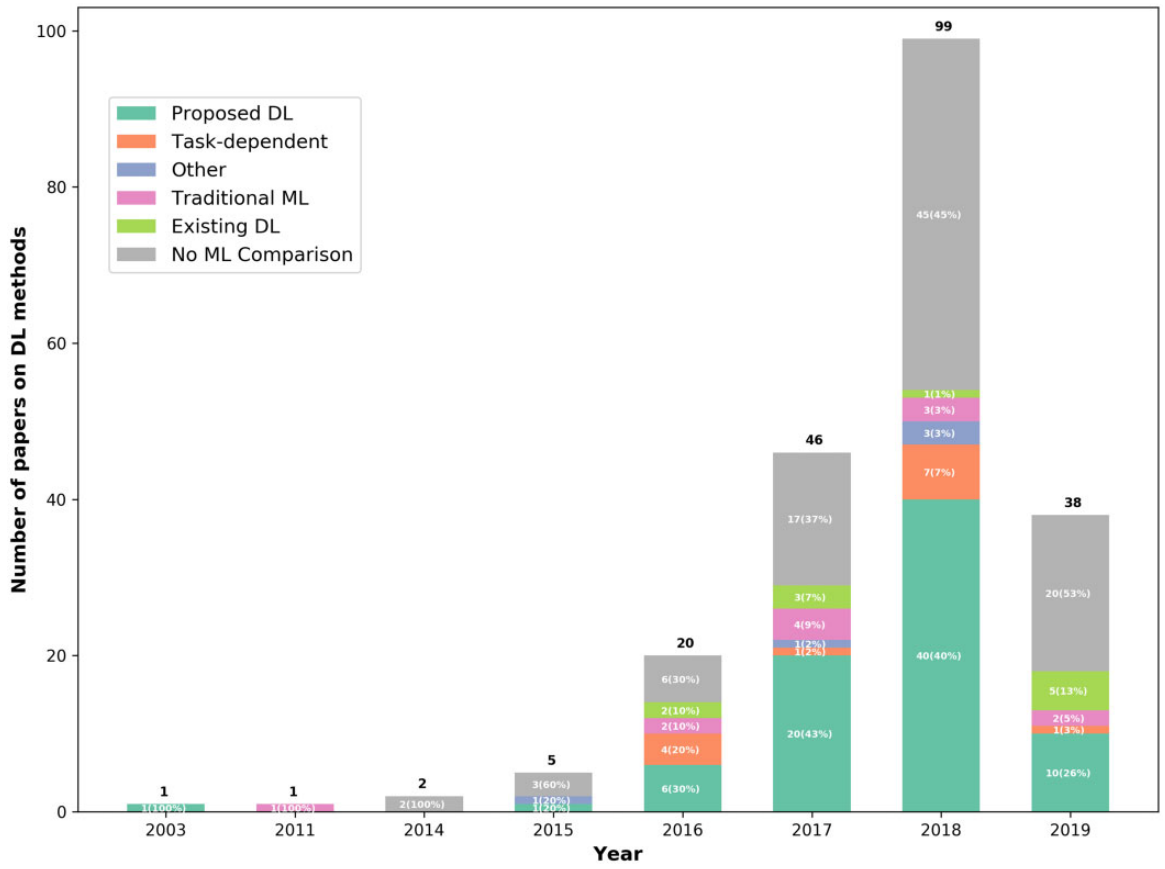

Figure 6. Comparisons between deep learning techniques and traditional machine learning, over time.

Substantiating vs surprising results

A number of commonly held assumptions were substantiated in our data analysis. CNNs have dominated the Text Classification task because of early, successful CNN-based methods.[112] A similar effect was noted for LSTMs with NER (typically cast as a sequence labeling problem). Also, as expected, early adopters of deep learning DL for clinical NLP were publishing in NLP venues rather than informatics or medical venues; there is a lag for adoption of cutting-edge techniques in the informatics community, and an even longer one for the medical community.

However, other analyses were surprising when compared with conventional wisdom. Whereas previous approaches to clinical NLP seemed to extensively utilize knowledge resources, only 17.9% of DL-based approaches did so. Whereas earlier work[24] found French to be the most prolific non-English language in clinical NLP, our study presents a previously unreported fact about non-English clinical NLP: when dealing with deep learning, Chinese has more representation than French. Whereas some have assumed that cutting-edge DL research is entirely contained on preprint servers, we found that only 16.5% of the papers in our study were posted as preprints prior to their publication.

Projections

This review has suggested some potential future trends in DL for clinical NLP. Because of the cost of annotating clinical corpora and the privacy concerns with sharing in-domain training data, *domain adaptation* and *transfer learning* strategies are important. However, there has been little systematic analysis on this issue from a deep learning perspective, perhaps in part due to the lack of convincing results. With the rise of successful pretrained models like BERT, we expect that the use and refinement of transfer learning will rise in popularity quickly.

Despite this, we also believe *medical knowledge resources* have been underutilized. Though the mantra of deep learning has been to "let the weights determine what's important" rather than to handcraft features, DL architectures and inputs still need human input, as evidenced by a recent push to consider inductive bias (eg, gender biases found in word embeddings). Knowledge resources may provide calculable and objective means to guide data-driven DL algorithms, and the medical domain is uniquely equipped with such resources.

Further, it is clear from other subfields involving ML for clinical tasks that deep learning is not always successful.[113,114] In this we mean that oftentimes it fails to outperform basic models, such as logistic regression. In the current set of articles included in this study, however, there were no in-depth analyses of the limitations or failures of DL methods for clinical NLP. In the few cases (n = 12) where DL failed to outperform traditional ML, there was no investigation into the underlying causes of the failure. Missing are investigations into the relative merits and pitfalls of deep learning based on data size, data quality, language, medical specialty, informatics task, and, to our previous point, the inclusion of knowledge resources. Therefore, we assert that there is a critical need for empirical investigations into the limitations of DL methods for clinical NLP tasks. For example, accuracy improvements of DL methods often rely on large computational resources that consume substantial amounts of energy, which can be detrimental to the environment.[104]

Overall, we would argue that at this point it is dangerous to only compare DL-based methods, especially in regard to applications about to be deployed to the clinic. However, this paper also demonstrates that it is scientifically naïve to not compare to some kind of DL baseline as well. Additionally, emerging DL practice suggests some instability surrounding initialization and hyperparameter selection. Few works presented here experimented with re-initialization of random seeds, small adjustments to hyperparameters, etc; 1 example of simple stability characterizations was Tourille et al,[115] who repeated each experiment 30 times and plotted the results. We recommend that as best practices emerge in the NLP community (and these practices still are in the nascent stages), the clinical NLP community quickly adopt these and report results in a consistent manner.

Finally, we suggest that truly novel core methodological contributions in this field may eventually plateau or even decrease, but the application and tweaking of deep learning to many (potentially new) clinical tasks will continue beyond any such plateau.

Limitations

This methodical review has a number of limitations. First, there is the possible selection bias inherent to the search methods used. This includes bias both in the types of searches we performed as well as the underlying limitations of those search engines. For instance, despite the fact that the studies had already been published, our ACL Anthology search missed both emrQA[116] and CliCR[117] whereas a current search with the exact same criteria on the ACL Anthology would have returned those studies. Similarly, we consider that a great variety of relevant papers don't explicitly contain the keywords of our method in their titles, therefore additional papers were manually added based on our authors' research experiences. During the process, we noticed that specific DL models mentioned in the title were probably missed based on the current search and can be considered as keywords in the future—for instance, condensed memory networks,[39] graph-based models, BERT[41], etc. In addition, some conventional clinical tasks relying heavily on NLP are also mentioned in the title, while missed in the previous search, such as de-identification,[59] automatic ICD-9 coding,[44] diagnostic inference,[39] and patient representation learning.[45] If anything, the trend towards universality of DL methods for NLP means that simple keyword searches such as "deep learning" and "neural networks" will increasingly miss relevant papers as more and more of these methods are the assumed default and not an element of novelty.

Second, while our review attempted to define mostly objective criteria for data collection, some data elements (eg, Application vs Evaluation papers) were not precisely defined and still had an inherent element of subjectivity. Related to that, third, many of the data collection elements and the normalization to broader categories were difficult to judge and agree upon, especially as papers are diverse in structure and style, and we allowed multiple tags for many of the data collection elements. All authors who participated in data collection (9) were tested for inter-annotator agreement on 11 of the 212 papers; in brief, this revealed that most data elements' responses could be normalized, but fundamental disagreements between the 9 annotators were still found in virtually every category.

Finally, regarding scope, we found it infeasible to address the very interesting question of comparing clinical NLP vs open-domain NLP. To do so, we would have needed to narrow our focus (eg, only LSTMs or only text classification), which would have been a very different scope than the current study.

CONCLUSION

We have reported quantitative and qualitative analyses on our methodical review of 212 papers regarding DL in clinical NLP, finding an active research area with multiple vibrant communities making diverse contributions. We quantitatively observed some widely held associations in the community, regarding methods (eg, CNNs tend to be used on classification) as well as preferences (eg, NLP community prefers conferences, medical community prefers journals). We expect deep learning to continue and extend its leading role in the wider research context before other technological paradigms supplant it.

FUNDING

This work was supported in part by NIH Grants U01TR002062 and R00LM012104, the UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (CPRIT grant #RP160015), and PCORI grant ME-2018C1-10963.

AUTHOR CONTRIBUTIONS

SW and HX conceived the study, then designed it along with KR. SW oversaw the screening, data collection, analysis, drafting, and editing. SW, KR, SD, JD, ZJ, YS, SS, QW_1, QW_2, and BZ completed initial screening. SW, KR, SD, JD, ZJ, YS, SS, QW_2, and YX performed data collection. Analysis, figures, and drafting was split among algorithms (KR), embeddings (YS), tasks (ZJ), data (SS), contributions (SD), venues (JD), experiments (YX), and comparisons (QW_2). All authors reviewed the manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors would like to thank Dr Cui Tao, Amy Sisson, and Kate Krause.

CONFLICT OF INTEREST STATEMENT

Dr Xu and the University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

REFERENCES

1. Cohen K, Névéol A, Xia J, *et al*. Reproducibility in biomedical natural language processing. In: AMIA Annual Symposium Proceedings LIMSI-Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur; November 4–8, 2017; Washington, DC, USA.
2. Fokkens A, Erp MV, Postma M, *et al*. Offspring from reproduction problems: what replication failure teaches us. In: The 51st Annual Meeting of the Association for Computational Linguistics; 2013: 1691–701.
3. Mieskes M. A quantitative study of data in the NLP community. In: proceedings of the First ACL Workshop on Ethics in Natural Language Processing; 2017: 23–9.
4. Dreisbach C, Koleck TA, Bourne PE, *et al*. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* 2019; 125: 37–46.

5. 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): e12239–18.
6. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc* 2018; 25 (10): 1419–28.
7. Shickel B, Tighe PJ, Bihorac A, *et al*. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2018; 22 (5): 1589–604.
8. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems; 2013: 3111–9; December 5–10, 2013; Lake Tahoe, NV, USA.
9. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014: 1532–43.
10. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005; 18 (5-6): 602–10.
11. Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. *arXiv Preprint arXiv: 160301360*; 2016.
12. LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86 (11): 2278–324.
13. Kim Y. Convolutional neural networks for sentence classification. *arXiv Preprint arXiv: 14085882*; 2014.
14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv Preprint arXiv: 14090473*; 2014.
15. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Advances in Neural Information Processing Systems; 2017: 5998–6008; December 4–9, 2017; Long Beach, CA, USA.
16. Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. In: Advances in Neural Information Processing Systems; 2014: 2672–80; December 8–13, 2014; Montreal, Canada.
17. Yu L, Zhang W, Wang J, *et al*. Seqgan: sequence generative adversarial nets with policy gradient. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017.
18. Al-Aiad A, Duwairi R, Fraihat M. Survey: deep learning concepts and techniques for electronic health record. In: proceedings of the IEEE/ACS International Conference on Computer Systems Applied AICCSA; 2019; 2018-Novem: 1–5.
19. Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018; 15 (141): 20170387.
20. Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med* 2019; 25 (1): 24–9.
21. Gonzalez-Hernandez G, Sarker A, O'Connor K, *et al*. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearb Med Inform* 2017; 26: 214–27.
22. Liu F, Chen J, Jagannatha A, *et al*. Learning for biomedical information extraction: methodological review of recent advances. *arXiv Preprint arXiv:* 1606.07993; 2016.
23. Névéol A, Zweigenbaum P. Expanding the diversity of texts and applications: findings from the section on clinical natural language processing of the international medical informatics association yearbook. *Yearb Med Inform* 2018; 27: 193–8.
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. 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: 11–9.
26. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
27. Zeng Z, Deng Y, Li X, *et al*. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Trans Comput Biol and Bioinf* 2019; 16 (1): 139–53.
28. Lee J, Yoon W, Kim S, *et al*. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv Preprint arXiv: 190108746*; 2019.
29. Li R, Hu B, Liu F, *et al*. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. *JMIR Med Inform* 2019; 7 (1): e10788.
30. Sushil M, Šuster S, Luyckx K, *et al*. Unsupervised patient representations from clinical notes with interpretable classification decisions. In: proceedings of Workshop on Machine Learning for Health (NIPS 2017); December 8, 2017; Long Beach, CA, USA.
31. Lee SH. Natural language generation for electronic health records. *NPJ Digit Med* 2018; 1: 63.
32. Sushil M, Šuster S, Luyckx K, *et al*. Patient representation learning and interpretable evaluation using clinical notes. *J Biomed Inform* 2018; 84: 103–13
33. Rumeng L, Abhyuday NJ, Hong Y. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. *AMIA Annual Symposium Proceedings* 2017; **2017**: 1149–58.
34. Zhang Z, Zhou T, Zhang Y, *et al*. Attention-based deep residual learning network for entity relation extraction in Chinese EMRs. *BMC Med Inform Decis Mak* 2019; 19 (S2): 55.
35. Qiu J, Zhou Y, Wang Q, *et al*. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. *IEEE Trans Nanobiosci* 2019; 18(3): 1–1. doi: 10.1109/TNB.2019.2908678.
36. Qiu J, Wang Q, Zhou Y, *et al*. Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions. In: proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine BIBM 2018; 2019: 935–42. doi: 10.1109/BIBM.2018.8621360.
37. Liang Z, Liu J, Ou A, *et al*. Deep generative learning for automated EHR diagnosis of traditional Chinese medicine. *Comput Methods Programs Biomed* 2018; 0: 1–7. doi: 10.1016/j.cmpb.2018.05.008.
38. Li F, Yu H. An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models. *J Am Med Informatics Assoc* 2019; 26 (7): 646–54. https://doi.org/10.1093/jamia/ocz018
39. Prakash A, Zhao S, Hasan SA, *et al*. Condensed memory networks for clinical diagnostic inferencing. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017.
40. Goodwin TR, Harabagiu SM. Inferring clinical correlations from EEG reports with deep neural learning. *AMIA Annu Symp Proc* 2017; 2017: 770–9. http://www.ncbi.nlm.nih.gov/pubmed/29854143%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi? artid=PMC5977577.
41. Alsentzer E, Murphy JR, Boag W, *et al*. Publicly available clinical BERT embeddings. In: Clinical Natural Language Processing Workshop NAACL; *2019* 2019.
42. Jain S, Wallace BC. Attention is not explanation. In: proceedings of the NAACL-HLT 2019; 2019: 3543–56. http://arxiv.org/abs/1902.10186.
43. Devlin J, Chang M-W, Lee K, *et al*. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv: 181004805*; 2018.
44. Xie P, Shi H, Zhang M, *et al*. A neural architecture for automated ICD coding. In: proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: 1066–76. http://aclweb.org/anthology/P18-1098.
45. Dligach D, Miller T. Learning patient representations from text. In: proceedings 7th Jt Conf Lex Comput Semant; 2018: 119–23. http://arxiv.org/abs/1805.02096.
46. Si Y, Roberts K. A frame-based NLP system for cancer-related information extraction. *AMIA Annu Symp Proc* 2018; 2018: 1524–33.
47. Wunnava S, Qin X, Kakar T, *et al*. Adverse drug event Detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. *Drug Saf* 2019; 42 (1): 113–22.

48. Wu J, Hu X, Zhao R, *et al*. Clinical named entity recognition via bi-directional LSTM-CRF model. *CEUR Workshop Proc* 2017; 1976: 31–6.
49. Gao S, Young MT, Qiu JX, *et al*. Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Inform Assoc* 2017; 16: 16.
50. Rajput K, Chetty G, Davey R, *et al*. Performance analysis of deep neural models for automatic identification of disease status. In: *International Conference on Machine Learning and Data Engineering (iCMLDE2018)*. Institute of Electrical and Electronics Engineers Inc., Piscataway; 2018: 142–8.
51. Miftahutdinov Z, Tutubalina E. Deep learning for ICD coding: Looking for medical concepts in clinical documents in English and in French. In: *9th International Conference on CLEF Association*; 2018: 203–15. doi: 10.1007/978-3-319-98932-7
52. Newman-Griffis D, Zirikly A. Embedding transfer for low-resource medical named entity recognition: a case study on patient mobility. In: proceedings of the BioNLP; 2018: 1–11. http://arxiv.org/abs/1806.02814.
53. Weng WH, Wagholikar KB, McCray AT, *et al*. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017; 17 (1): 1–13.
54. Lin C, Miller TA, Dligach D, *et al*. Self-training for temporal relation extraction with recurrent neural networks. In: proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018); 2018: 165–76.
55. Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch. *J Mach Learn Res* 2011; 12: 2493–537.
56. Wu Y, Xu J, Jiang M, *et al*. A study of neural word embeddings for named entity recognition in clinical text. *AMIA Annu Symp Proc* 2015; 2015: 1326–33.
57. Wu Y, Xu J, Zhang Y, *et al*. Clinical abbreviation disambiguation using neural word embeddings. *BioNLP;* 2015; 15: 171–6.
58. Mikolov T, Kombrink S, Burget L, *et al*. Extensions of recurrent neural network language model. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2011: 5528–31.
59. Ekbal A, Saha S, Bhattacharyya P. Deep learning architecture for patient data de-identification in clinical records. In: proceedings of the Clinical Natural Language Processing Workshop; 2016: 32–41. https://aclweb.org/anthology/W/W16/W16-4206.pdf.
60. Wu LY, Fisch A, Chopra S, *et al*. Starspace: embed all the things! In: Thirty-Second AAAI Conference on Artificial Intelligence; 2018.
61. Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. *Proc Mach Learn Res* 2018; 85: 440–64.
62. Artetxe M, Labaka G, Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv Preprint arXiv: 180506297*; 2018.
63. Weegar R, Pérez A, Casillas A, *et al*. Deep medical entity recognition for Swedish and Spanish. In: proceedings of the 2018 IEEE International Conference on Bioinformatics Biomedical BIBM; 2018: 1595–601. doi: 10.1109/BIBM.2018.8621282.
64. Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. *arXiv Preprint arXiv: 180205365;* 2019: 26 (11), 1297–1304. https://doi.org/10.1093/jamia/ocz096.
65. Si Y, Wang J, Xu H, *et al*. Enhancing clinical concept extraction with contextual embedding. *J Am Med Informatics Assoc* 2019.
66. Zhu H, Paschalidis IC, Tahmasebi A. Clinical concept extraction with contextual word embedding. In: Machine Learning Health Workshop NeurIPS; 2018.
67. Xu G, Wang C, He X, *et al*. Improving clinical named entity recognition with global neural attention. In: *2nd Asia Pacific Web Web-Age Information Management Joint Conference on Web Big Data, APWeb-WAIM*; 2018; 10988: 264–79.
68. Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. *J Biomed Inform* 2018; 86: 71–8.
69. Cai X, Dong S, Hu J. A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records. *BMC Med Inform Decis Mak* 2019; 19 (S2): 65.
70. Li Z, Yang Z, Shen C, *et al*. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. *BMC Med Inform Decis Mak* 2019; 19 (S1): 22.
71. Medina S, Turmo J, Estevez-Velarde S, *et al*. Joint classification of key-phrases and relations in electronic health documents. In: 2018 Taller Workshop on Semantic Analysis at SEPLN, TASS 2018. TALP Research Center, Universitat Politècnica de Catalunya, Spain: CEUR-WS; September 18, 2018; 83–8.
72. Lazib L, Qin B, Zhao Y, *et al*. A syntactic path-based hybrid neural network for negation scope detection. *Front Comput Sci* 2020; 14 (1): 84–94.
73. Ji B, Liu R, Li S, *et al*. A hybrid approach for named entity recognition in Chinese electronic medical record. *BMC Med Inform Decis Mak* 2019; 19 (S2): 64.
74. Liu Z, Tang B, Wang X, *et al*. De-identification of clinical notes via recurrent neural network and conditional random field. *J Biomed Inform* 2017; 75: S34–42.
75. Wang Q, Zhou Y, Ruan T, *et al*. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *J Biomed Inform* 2019; 92: 103133.
76. Banerjee I, Chen MC, Lungren MP, *et al*. Radiology report annotation using intelligent word embeddings: applied to multi-institutional chest CT cohort. *J Biomed Inform* 2018; 77: 11–20.
77. Wu Y, Yang X, Bian J, *et al*. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. In: AMIA Annual Symposium Proceedings; 2018: 1110–7.
78. Li D, Huang M, Li X, *et al*. MfeCNN: mixture feature embedding convolutional neural network for data mapping. *IEEE Trans Nanobiosci* 2018; 17 (3): 165–71.
79. Turner CA, Jacobs AD, Marques CK, *et al*. Word2Vec inversion and traditional text classifiers for phenotyping lupus. *BMC Med Inform Decis Mak* 2017; 17 (1): 1–11.
80. Santiso S, Perez A, Casillas A. Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery. *IEEE J Biomed Heal Inform* 2019; 23 (5): 2148–55.
81. Li F, Liu W, Yu H. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. *JMIR Med Inform* 2018; 6 (4): e12159.
82. Suárez-Paniagua V, Segura-Bedmar I, Martínez P, *et al*. LABDA at TASS-2018 task 3: convolutional neural networks for relation classification in Spanish eHealth documents. In: 2018 Taller Workshop on Semantic Analysis at SEPLN, TASS; September 18, 2018. Computer Science Department, Carlos III University of Madrid, Leganés, Madrid: CEUR-WS 71–6.
83. Dligach D, Miller T, Lin C, *et al*. Neural temporal relation extraction. In: proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017. 746–51.
84. Sahu SK, Anand A. What matters in a transferable neural network model for relation classification in the biomedical domain? *Artif Intell Med* 2018; 87: 60–6.
85. Su J, Hu J, Jiang J, *et al*. Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records. *Comput Methods Programs Biomed* 2019; 172: 1–10. doi: 10.1016/j.cmpb.2019.01.007
86. Soldaini L, Yates A, Goharian N. Denoising clinical notes for medical literature retrieval with convolutional neural model. In: proceedings of the 2017 ACM Conference on Information Knowledge Management; 2017: 2307–10.
87. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. *Yearb Med Inform* 1993; 2: 41–51.
88. Ran Y, He B, Hui K, *et al*. A document-based neural relevance model for effective clinical decision support. In: proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine BIBM; 2017: 798–804. doi: 10.1109/BIBM.2017.8217757

89. Moen H, Ginter F, Marsi E, *et al*. Care episode retrieval: DISTRIBUTIONAL semantic models for information retrieval in the clinical domain. *BMC Med Inform Decis Mak* 2015; 15 (S2): S2.
90. Jimenez-del-Toro OT, Otálora S, Atzori M, *et al*. Deep multimodal case–based retrieval for large histopathology datasets. In: 3rd International Workshop on Patch-Based Technique in Medical Imaging, Patch-MI 2017 held conjunction with 20th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2017; 2017; 10530: 149–57.
91. Wu L, Wan C, Wu Y, *et al*. Generative caption for diabetic retinopathy images. In: 2017 International Conference on Security Pattern Analysis and Cybernetics SPAC 2017; 2018: 515–9. doi: 10.1109/SPAC.2017.8304332.
92. Salloum W, Finley G, Edwards E, *et al*. Automated preamble detection in dictated medical reports. In: proceedings of the BioNLP 2017 Workshop; 2017: 287–95.
93. Sadoughi N, Finley GP, Edwards E, *et al*. Detecting section boundaries in medical dictations: toward real-time conversion of medical dictations to clinical reports. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Cham: Springer; 2018: 563–73.
94. Zhang Y, Tiryaki F, Jiang M, *et al*. Parsing Clinical Text: How Good are the state-of-The-Art Deep Learning Based Parsers? In: 6th IEEE International Conference on Healthcare Informatics Workshops, ICHI-W. Institute of Electrical and Electronics Engineers Inc., Houston; 2018: 80–1.
95. Finley G, Salloum W, Sadoughi N, *et al*. From dictations to clinical reports using machine translation. In: proceedings of the NAACL-HLT; 2018: 121–8. doi: 10.18653/v1/n18-3015.
96. Fivez P, Šuster S, Daelemans W. Unsupervised context-sensitive spelling correction of English and Dutch clinical free-text with word and character N-Gram embeddings. *BioNLP* 2017; 7: 39–52.
97. Rokach L, Romano R, Maimon O. Automatic identification of negated concepts in narrative clinical reports. In: *proceedings of the Eighth International Conference on Enterprise Information Systems-AIDSS*; 2011: 257–62.
98. Taylor SJ, Harabagiu SM. The role of a deep-learning method for negation detection in patient cohort identification from electroencephalography reports. *AMIA Annu Symp Proc* 2018; 2018: 1018–27.
99. Huynh T, He Y, Willis A, *et al*. Adverse drug reaction classification with deep neural networks. In: proceedings of the COLING; 2016: 877–87. http://www.aclweb.org/anthology/C16-1084.
100. Zhou X, Xiong H, Zeng S, *et al*. An approach for medical event detection in Chinese clinical notes of electronic health records. *BMC Med Inform Decis Mak* 2019; 19 (S2): 54.
101. Dev S, Zhang S, Voyles J, *et al*. Automated classification of adverse events in pharmacovigilance. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine BIBM; 2017: 1562–6.
102. Dandala B, Joopudi V, Devarakonda M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. *Drug Saf* 2019; 42 (1): 135–46.
103. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. *J Med Internet Res* 2018; 20: 1–15.
104. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: 57th Annual Meeting of the Association for Computational Linguistics (ACL); 2019. doi: 10.18653/v1/p19-1355.
105. Leeuwenberg A, Moens M-F. Word-level loss extensions for neural temporal relation classification. In: proceedings of the 27th International Conference on Computational Linguistics; 2018: 3436–47. https://www.aclweb.org/anthology/C18-1291.
106. Yerebakan HZ, Shinagawa Y, Bhatia P, *et al*. Document representation learning for patient history visualization. In: proceedings 27th International Conference on Computational Linguistics System Demonstrations; 2018: 30–3. https://aclanthology.coli.uni-saarland.de/papers/C18-2007/c18-2007.
107. Yao L, Zhang Y, Wei B, *et al*. Traditional Chinese medicine clinical records classification using knowledge-powered document embedding In: 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016. Institute of Electrical and Electronics Engineers Inc.: 1926–8.
108. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 2011; 10 (9): 712.
109. Collins FS, Tabak LA. NIH plans to improve reproducability. *Nature* 2014; 505 (7485): 612–3.
110. Association for Computing Machinery. Artifact review and badging; 2018. https://www.acm.org/publications/policies/artifact-review-badging Accessed June 26, 2019.
111. Plesser HE. Reproducibility vs. replicability: a brief history of a confused terminology. *Front Neuroinform* 2018; 11: 1–4.
112. Lai S, Xu L, Liu K, *et al*. Recurrent convolutional neural networks for text classification. In: *AAAI;* 2015: 2267–73.
113. Yang W, Lu K, Yang P, *et al*. Critically examining the" neural hype": weak baselines and the additivity of effectiveness gains from neural ranking models. *arXiv Preprint arXiv: 190409171*; 2019.
114. Jie MA, Collins GS, Steyerberg EW, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019.
115. Tourille J, Doutreligne M, Ferret O, *et al*. Evaluation of a sequence tagging tool for biomedical texts. In: proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis; 2018: 193–203. https://aclanthology.info/papers/W18-5622/w18-5622.
116. Pampari A, Raghavan P, Liang J, *et al*. emrQA: a large corpus for question answering on electronic medical records. In: proceedings of the 2018 Conference on Empirical Methods on Natural Language Processing; 2018: 2357–68. http://arxiv.org/abs/1809.00732.
117. Šuster S, Daelemans W. Clicr: A dataset of clinical case reports for machine reading comprehension. *arXiv Preprint arXiv: 180309720*; 2018.