Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Med7: A transferable clinical natural language processing model for electronic health records

Andrey Kormilitzin [a,*], Nemanja Vaci [b], Qiang Liu [a], Alejo Nevado-Holgado [a]

[a] Department of Psychiatry, University of Oxford, Oxford OX3 7JX, United Kingdom
[b] Department of Psychology, University of Sheffield, Sheffield S1 1HD, United Kingdom

ARTICLE INFO

Keywords:
Clinical natural language processing
Neural networks
Self-supervised learning
Noisy labelling
Active learning

ABSTRACT

Electronic health record systems are ubiquitous and the majority of patients' data are now being collected electronically in the form of free text. Deep learning has significantly advanced the field of natural language processing and the self-supervised representation learning and the transfer learning have become the methods of choice in particular when the high quality annotated data are limited. Identification of medical concepts and information extraction is a challenging task, yet important ingredient for parsing unstructured data into structured and tabulated format for downstream analytical tasks. In this work we introduced a named-entity recognition (NER) model for clinical natural language processing. The model is trained to recognise seven categories: drug names, route of administration, frequency, dosage, strength, form, duration. The model was first pre-trained on the task of predicting the next word, using a collection of 2 million free-text patients' records from MIMIC-III corpora followed by fine-tuning on the named-entity recognition task. The model achieved a micro-averaged $F1$ score of 0.957 across all seven categories. Additionally, we evaluated the transferability of the developed model using the data from the Intensive Care Unit in the US to secondary care mental health records (CRIS) in the UK. A direct application of the trained NER model to CRIS data resulted in reduced performance of $F1 = 0.762$, however after fine-tuning on a small sample from CRIS, the model achieved a reasonable performance of $F1 = 0.944$. This demonstrated that despite a close similarity between the data sets and the NER tasks, it is essential to fine-tune the target domain data in order to achieve more accurate results. The resulting model and the pre-trained embeddings are available at https://github.com/kormilitzin/med7.

1. Introduction

Recent years have seen remarkable technological advances in digital platforms for medicine and healthcare. The majority of patients' medical records are now being collected electronically and represent unparalleled opportunities for research, delivering better health care and improving patients' outcomes. Substantial amount of patients' information is collected in a free-text form by clinicians, nurses and caregivers through the interview and assessments. The textual medical records contain rich information about a patient's history as it is expressed in natural language and allows to reflect nuanced details. However, free-texts pose certain challenges in their direct utilisation as opposed to structured and ready-to-use data sources. Manual processing of all patients' free-texts records severely limits the utilisation of unstructured data and makes the process of data mining not scalable. Machine learning algorithms are well poised to process a large amount of data, spot unusual interactions and extract meaningful information. Recent lines of research in the field of natural language processing (NLP), such as deep contextualised word representations [1], transformer-based architectures [2] and large language models [3], showed utility for clinical natural language processing with unstructured medical records [4]. Despite technological advances in NLP models, there is a number of challenges pertinent to the field of clinical NLP which should be addressed in order to develop trustworthy models for information extraction. The foremost challenge is the dearth of high quality annotated examples to robustly train generalisable models. Large amounts of medical data cannot be made publicly available for crowdsourcing annotations, similar to ImageNet [5,6] or by means of Amazon Mechanical Turk, due to ethical consideration of patients' privacy preservation and information security [7]. Therefore, most of the data annotations are made by a limited number of domain experts, such as clinicians or nurses, and cannot be shared. Since 2006, the

[*] Corresponding author.
E-mail address: andrey.kormilitzin@psych.ox.ac.uk (A. Kormilitzin).

https://doi.org/10.1016/j.artmed.2021.102086
Received 18 August 2020; Received in revised form 24 February 2021; Accepted 3 May 2021
Available online 18 May 2021

Informatics for Integrating Biology and the Bedside (i2b2) initiative [8] has been organising regular challenges on clinical natural language processing and the organisers have been providing a sample of carefully selected for each particular task fully anonymised annotated data sourced from the MIMIC-III (Medical Information Mart for Intensive Care-III) electronic health records database [9].

Identification of concepts of interest in free-texts is a sub-task of information extraction (IE), more commonly known as named-entity recognition (NER). The NER task seeks to classify words into predefined categories [10] and to assign labels to them. A robust and accurate NER model for identification of medical concepts, such as drug names, strength, frequency of administration, reported symptoms, diagnoses, health score and many more, is an essential and foundational component of any clinical IE system. However, in order to develop a reliable and generalisable NER model for real-world observational data, one should first address a number of challenges.

Despite the availability of both, annotated i2b2 data and the entire MIMIC-III electronic health records (EHR) database, the models developed using these data sources are not guaranteed to generalise robustly on other, yet similar EHR datasets, even on the same downstream tasks. Many supervised learning algorithms are based on the assumption that the training and test sets are sampled from the same distribution. However, when the target and the source domains are different, it is expected that the model will underperform [11]. One of the potential solutions to transferability of a model trained to recognise concepts from labelled data in a source domain that also performs well on a different but related target domain, regarded as domain adaptation [12,13].

In this study we address the aforementioned problems by implementing three strategies. First, the underlying deep neural network language model was pre-trained in a self-supervised manner on the entire corpus of clinical notes from the MIMIC-III comprising over 2 million free-text documents using the cloze-style approach [14], where some words are masked and must be predicted given the rest of the text. Second, using the weak-supervision method [15], we developed synthetic training data with noisy labels. Lastly, we incorporated all ingredients into an active learning approach, whereby a baseline pattern-matching NER model was used to proactively suggest spans in the unseen text that are highly likely belong to the categories of interest and allowed a human annotator to review suggestions, correct them if needed and to generate gold standard annotations more effectively.

Additionally, we demonstrated that the developed NER model trained on a source domain from the intensive care MIMIC EHR data in the US, failed to generalise well on the target domain sourced from UK Clinical Record Interactive Search (UK-CRIS) platform, which is the largest secondary care mental health database in the United Kingdom. However, using domain adaptation and fine-tuning on UK-CRIS, we showed that the NER model could retain its performance on UK-CRIS.

2. Related work

The topic of clinical natural language processing and information extraction has been actively researched over the past three decades, in particular with the introduction and adoption of electronic health records platforms. The methods have evolved from simple logic and rule-based systems to complex deep learning architectures [16,17]. One of the common approaches to information extraction is by transforming free text data into coded representation via lookup tables, such as universal medical language system (UMLS) [18] or structured clinical vocabulary for use in an electronic health record (SNOMED CT). Some rule-based systems used semantic lexicons to identify concepts in biomedical literature [19] with more complex linguistic features.

With the advances in machine learning algorithms, such methods as hidden Markov models and conditional random fields [20] were used to label entities for the NER task. Since medication and their related information extraction in central to many clinical text mining tasks, there have been many efforts to develop reliable NER models. Among other, MetaMap which was widely used to map biomedical concepts in MEDLINE/PubMed abstract to the unified medical language system (UMLS) Metathesaurus [21]. MedEx was developed using semantic tagger and Chart parser [22] to extract medications and their related eleven semantic categories from discharge electronic medical records [23]. The members of the Open Health Natural Language Processing Consortium has developed a number of models and pipelines, based on Apache UIMA framework, such as MedCoref by Mayo Clinic for coreference resolution to link the markables corresponding to the same entity [24], MedTagger for a collaborative annotation of medical datasets [25], MedTime for temporal information detection, using a conditional random field classifier with lexical features and medical ontology [26] and cTAKES [27] by Mayo Clinic, a comprehensive framework comprising various analysis engines, including event identification, terminology mapping, temporal expressions identification and extraction drug signatures. MedXN uses regular expressions to extract medications, their related information (i.e., dosage, strength, frequency, form, and route) with further normalisation and mapping to RxNorm concept unique identifier [28]. CLAMP (Clinical Language Annotation, Modeling, and Processing) framework [29] comprises several components to facilitate building customised pipelines for diverse clinical applications. General Architecture for Text Engineering (GATE) is an open source framework with customisable tools and engines for diverse text mining and natural language processing tasks [30].

In the last decade, deep learning methods have played an essential role in developing more capable models for natural language processing and in particular, in the biomedical domain. Word embeddings [31,32] were introduced as numerical representation of textual data and were used as input layers to deep neural networks. For a comprehensive review on word embeddings for clinical applications please refer to [33]. More recently, the unsupervised model pre-training on a large collection of unlabelled data with further fine-tuning on a downstream task, has taken off and demonstrated its high potential [34]. Since the introduction of the Transformer-based deep neural network architectures, such as BERT [3], Roberta [35], XLNet [36] and others, the transfer learning approach of reusing pre-trained models became the method of choice for the majority of NLP tasks. Some notable examples of pre-trained deep learning models for biomedical natural language processing are: BioBERT [37] for text-mining, ClinicalBERT [38,39] for contextual word representations fine-tuned on the electronic health records and predicting hospital readmission. Another open source Python library 'scispaCy' [40] was introduced for biomedical natural language processing.

In this work we developed an open source named-entity recognition model ("Med7") for identification of seven categories in free-text electronic patients' records: Dosage, Drug names, Duration, Form, Frequency, Route and Strength.

3. Materials and methods

The Med7 NER model development combined two deeply interconnected processes: the data generation and model training. The data set for training comprised a combination of gold and silver annotated corpora. The model development phase included self-supervised pre-training and hyperparameter optimisation to achieve both high accuracy and robust generalisation. Clinical texts were sourced from both MIMIC-III and UK-CRIS data bases.

3.1. Gold standard corpus

The gold annotated corpus for model development was used from the Track 2 of the 2018 National NLP Clinical Challenges (n2c2) Shared Task on drug related concepts extraction [41]. The data set comprised a collection of discharge letters from the Intensive Care Unit (ICU) and contained very rich and detailed information about medications used for treatment. The data set was annotated by a team of clinicians, randomly

Table 1
The distribution of annotated text spans used for training and evaluation of the Med7 NER model.

	MIMIC-III				OxCRIS	
	Gold (n2c2)	Gold (our)	Silver	Test	Gold (our)	Test
Dosage	4227	3437	2792	2681	298	48
Drug	16257	12687	10551	10575	3253	571
Duration	592	620	462	378	1006	215
Form	6657	5056	4299	4359	410	63
Frequency	6281	5106	4317	4012	1604	305
Route	5460	4554	3761	3513	208	32
Strength	6694	5246	4328	4230	1338	276
Number of documents	303	606	303	202	536	134

split and provided by the organisers into training and test sets with 303 and 202 documents respectively. In this work we primarily focused on seven annotated categories: Dosage, Drug names, Duration, Form, Frequency, Route and Strength with the transferability of the trained model across various settings and clinical specialisations (e.g., primary care, secondary care, ICU, physical and mental health). We aimed to develop a model which will be beneficial to the biomedical community and be robustly used in a variety of downstream natural language processing tasks using free text medical records. Furthermore, our two annotators, trained on the official guide from the n2c2 challenge, marked additional 606 documents randomly sampled from the MIMIC-III data set, ensuring that no samples appeared in the hold-out test set.

In order to evaluate how well a model trained on the MIMIC-III data can be transferred to UK-CRIS clinical notes, we asked one of our two annotators additionally to produce a sample of gold standard corpus using UK-CRIS data. UK-CRIS contains more than 500 million clinical notes from 2.7 million de-identified patient records from 12 National Health Service (NHS) Network Partners across the UK.[1] The manual annotation was performed with Prodigy and following the procedure outlined in [42] by leveraging the active learning annotation approach. The active learning approach in this context means that a rule-based model suggests entities and spans on unseen texts and a human annotator accepts or corrects the model suggestions therefore creating the gold annotated examples.

Furthermore, we aimed to investigate how accurate the developed Med7 model, trained on MIMIC-III electronic health records sourced from the Beth Israel Deaconess Medical Center in Massachusetts (United States), can perform when applied to UK-CRIS secondary care mental health electronic health records in the United Kingdom. We selected a random sample of 670 documents from the Oxford Health NHS Foundation Trust (OHFT) instance of UK-CRIS Network (OxCRIS) and asked one of our clinical annotators to mark them for seven categories following the official guidelines of the n2c2 challenge (see Table 1).

3.2. *Silver standard corpus*

Several recent lines of research have demonstrated a clear benefit in terms of achieving higher accuracy and better generalisation of neural networks trained with corrupted, noisy and synthetically augmented data [43–46]. The silver annotations are automatically generated by using a rule-based approach to match available keywords (e.g., known drug names, standard expressions for duration or frequency) against clinical notes. Training with synthetically produced annotation and data augmentation also alleviates the problem of learning from a limited amount of manually annotated data. Similar to the idea presented in 'Snorkel' [47], we designed a number of labelling functions (LF) by compiling a list of rules, linguistic and keyword patterns for all seven named-entity categories. Additionally, we exploited a 'sense2vec' approach [48] which was fine-tuned on the entire MIMIC-III corpora to bootstrap keywords and patterns. 'Sense2vec' is a more complex version of the 'Word2vec' method [49] for representation of words as vectors. The major improvement over 'word2vec' is that 'sense2vec' also learns from linguistic annotations of words for sense disambiguation in their embeddings. Drug names, both generic and brand names, were additionally sourced from publicly available free-to-access resources for research such as RxNorm [50] and DrugBank [51]. We set aside 30 documents (10%) sampled at random from the n2c2 gold standard training data as a validation set.

In total, the Med7 NER model was trained on 1212 and tested on 202 documents from MIMIC-III as well as further fine-tuned on 536 and tested on 134 from OxCRIS as summarised in Table 1.

3.3. *Text pre-processing*

In order to compare the performance of the developed medication extraction model on both MIMIC-III (n2c2 2018) and OxCRIS data without significantly altering the texts, text cleaning and pre-processing steps were taken to standardise texts. We kept these steps as simple as possible in order to preserve the natural variation between the texts and test how the model generalises across different settings. Some OxCRIS documents, such as scanned letters, were transformed into electronic texts via optical character recognition (OCR), resulted in ASCII artefacts. Specifically, all letters were converted to lowercase, email addresses, non-ASCII characters, website URLs, HTML and XML tags were removed using the Python package 'Beautiful Soup 4'. Standard escape sequences ('\t', '\ n' and '\ r') were removed and the offsets of gold-annotated entities were adjusted accordingly. Extra white spaces were reduced to a single white space. Text was tokenised using the spaCy's native tokeniser.

3.4. *Self-supervised model pre-training*

One of the obstacles to developing an accurate information extraction model is the dearth of a sufficient amount of high-quality annotated data to train the model. In contrast to publicly available large manually annotated data sets for computer vision (e.g., ImageNet [5,6] and for various natural language processing downstream tasks [52–54] manually annotated texts for clinical concepts extraction are quite rare [41]. The shortage of annotated clinical data is mainly due to privacy concerns and potential identification of personal medical information of patients. Several lines of research have addressed the problem of learning from limited annotated data in the clinical domain [55–57] and pre-training of the underlying language model and word representations generally leads to better performance with less data [34].

We used the spaCy's[2] implementation of a cloze-style word reconstruction, similar to the masked language model objectives introduced in BERT [3], but instead of predicting the exact word identifier from the vocabulary, the GloVe [32] word's vector was predicted using a static embedding table with a cosine loss function. The pre-trained language model was then used to initialise the weights of convolutional neural network layers, rather than starting with random weights. We experimented with various combinations of hyperparameters of the language model, such as the number of rows and width of embedding tables and a depth of convolutional layers.

The pre-training task was performed on the entire MIMIC-III data set of over 2 million documents for 350 epochs using a number of configurations of the width and depth of the convolutional (CNN) layers. Each configuration was trained on a single GTX 2080 Ti GPU. CNN dimensions, summary statistics of the pre-training text corpus, the average running time per epoch in minutes and the model size in MB are

[1] https://crisnetwork.co.

[2] https://spacy.io.

Table 2
Model pre-training characteristics for various combinations of convolutional layers dimensions. Time and the resulting model size are reported in minutes and megabytes (MB) respectively.

Configuration	Width	Depth	Time	Size
(default)	96	4	73	3.8
	128	8	90	18.3
	256	8	118	47.6
	256	16	164	66.1
Number of documents		2,083,054		
Number of words		3,129,334,419		

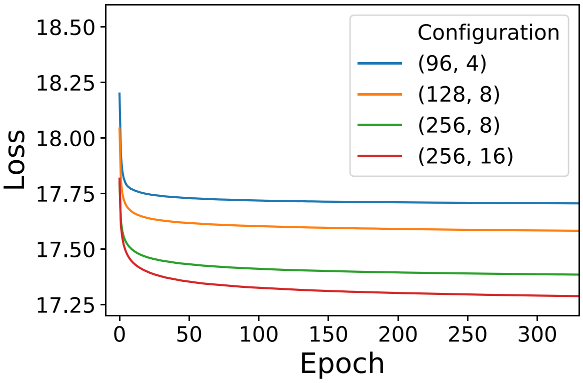

Fig. 1. The decaying loss of pre-trained models.

Table 3
Change in accuracy with more training data. Delta denotes a relative improvement.

Fraction	Accuracy	Delta
0%	0.0	Baseline
25%	90.66	+90.66
50%	91.93	+1.27
75%	92.42	+0.49
100%	92.63	+0.21

summarised in Table 2. The corresponding training losses, logarithmically scaled, are plotted in Fig. 1.

3.5. Named entity recognition model

In this work the named-entity recognition model for extraction of medication-related information was implemented in Python 3.7 using spaCy open source library for NLP tasks [58] which is optimised for speed on CPUs, has an intuitive API and easily integrates with the active learning-based annotation tool Prodigy [59]. The architecture of SpaCy's NER model is based on convolutional neural networks with tokens represented as hashed Bloom embeddings [60] of prefix, suffix and lemmatisation of individual words augmented with a transition-based chunking model [61].

3.6. Rationale for collecting more training samples

Generally, collecting more training data will improve the model accuracy and lead to better generalisation. We simulated an acquisition of more data by training of NER model on fractions (25%, 50%, 75% and 100%) of the training set and evaluating on the validation set, which was not used in the training procedure. We indeed observed (Table 3) a steady upward trend in improvement of accuracy while using more training data, especially in the last fraction of data which indicates the advantages of further collecting more data.

3.7. Model evaluation

In order to estimate the performance of the proposed named-entity recognition model, we used the evaluation schema proposed in SemEval'13 and outlined in Section A. The evaluation schema comprised a number of potential errors categories produced by the model and the model performance metrics, such as precision, recall and $F1$ score were computed using the expressions (1). Under the current evaluation schema, partial match was considered as an exact match between the gold-annotated and the predicted labels while no restriction was imposed on the boundaries of the tokens. The rationale behind this approach was obvious from the ambiguity in gold-annotations examples corresponding to the same concept. For example, both sequences 'for 3 weeks' and '3 weeks' were labelled as 'Duration'. In particular, 492 of 967 (71%) text spans labelled as 'Duration' started with the word 'for'.

We estimated both, strict and lenient metrics. Strict metrics accounts only for the exact match in both, surface strings and the corresponding labels, whereas the lenient metrics allow for partial matches. Specifically, strict and lenient metrics were obtained from equations (1) with $\alpha = 0$ and $\alpha = 1$ correspondingly. Various error types are specified as follow. Correct (COR) represents a complete match of both, the annotation boundary and the entity type. Incorrect (INC) is the case where at least one of the predicted boundaries or the entity type do not match. Partial (PAR) match corresponds to predicted entity boundary which overlaps with ground-truth annotation, but they are not exactly the same. Missing (MIS) the case where the ground-truth annotated boundary is not predicted by the NER, but the ground-truth string is present in the gold-annotated corpus. Spurious (SPU) corresponds to predicted entity boundary which does not exist in the gold-annotated corpus. The examples of each of the error types are shown in Appendix A in Table 7.

$$\begin{aligned} \text{Possible(POS)} &= \text{COR} + \text{INC} + \text{PAR} + \text{MIS} = \text{TP} + \text{FN} \\ \text{Actual(ACT)} &= \text{COR} + \text{INC} + \text{PAR} + \text{SPU} = \text{TP} + \text{FP} \\ \text{Precision} &= (\text{COR} + \alpha\text{PAR})/\text{ACT} \\ \text{Recall} &= (\text{COR} + \alpha\text{PAR})/\text{POS} \\ \text{F1score} &= 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{1}$$

3.8. Transferability across clinical domains

One of the challenges in developing a robust clinical information extraction system is in its generalisability beyond the data distribution it was trained on. Accurate algorithms developed using data from a small number of medical centres, have demonstrated their poor generalisability when applied within a similar context to other medical centres. For example, in a recent study on the algorithmic approach to early detection of sepsis [62], the training data were sourced from electronic health records of two hospitals, while the data from a third hospital were used for testing the developed algorithm. It has been demonstrated and discussed in details [63] that a highly accurate predictive algorithm, validated on a fraction of data from the same two hospitals, failed to achieve the same level of accuracy when tested on the data from the third hospital, not included in the training process. Poor performance using the out-of-distribution (OOD) data poses a significant challenge of wider applications of the developed models and is highly important when algorithms inform real-world decisions [64]. Some other works concerning domain adaption in the context of NER focused on the distributed word representations [65], leveraging rule-based annotators [66] and multi-task learning [67]. We draw inspiration from the ULMFiT approach [34], whereby we first domain adapted the pre-trained spaCy's language model on the entire MIMIC-III data followed by target NER task fine-tuning. For the transferability, we further fine-tuned the MIMIC-III NER model on the annotated data from UK-CRIS. We therefore investigated the feasibility of an application of

Table 4
The evaluation results of the NER model on the hold-out n2c2 Gold test set with 202 documents.

	Strict ($\alpha = 0$)			Lenient ($\alpha = 1$)		
	Precision	Recall	$F1$	Precision	Recall	$F1$
Dosage	0.879	0.831	0.854	0.957	0.904	0.931
Drug	0.954	0.926	0.941	0.984	0.956	0.971
Duration	0.817	0.733	0.773	0.953	0.854	0.901
Form	0.921	0.886	0.903	0.983	0.947	0.965
Frequency	0.801	0.784	0.792	0.989	0.969	0.979
Route	0.961	0.943	0.952	0.973	0.954	0.964
Strength	0.927	0.781	0.848	0.992	0.836	0.907
Average (micro)	0.916	0.871	0.893	0.982	0.933	0.957
Average (macro)	0.897	0.844	0.869	0.977	0.919	0.947

Table 5
The lenient evaluation results of the Med7 model using 134 test documents sourced from OxCRIS – the Oxford Health NHS Foundation Trust from within the OxCRIS electronic health records network.

	Before fine-tuning on OxCRIS			After fine-tuning on OxCRIS		
	Precision	Recall	$F1$	Precision	Recall	$F1$
Dosage	0.826	0.396	0.535	0.656	0.833	0.734
Drug	0.912	0.968	0.939	0.975	0.977	0.976
Duration	0.951	0.107	0.192	0.883	0.934	0.908
Form	0.554	0.611	0.581	0.924	0.968	0.946
Frequency	0.912	0.332	0.487	0.941	0.944	0.942
Route	0.348	0.719	0.469	0.882	0.938	0.909
Strength	0.938	0.877	0.906	0.996	0.917	0.955
Average (micro)	0.864	0.681	0.762	0.941	0.947	0.944
Average (macro)	0.778	0.586	0.609	0.901	0.932	0.914

Med7 NER model across two quite different settings, such as intensive care units in the US and secondary mental health in the UK and provided recommendations how to fine-tune and adapt the model.

4. Results

4.1. Named-entity recognition model

The evaluation results of the developed Med7 NER model are summarised in Table 4 using data shown in Table 1. A detailed token-level confusion matrix of the model predictions is presented in Appendix A in Table 8.

The explicit token-level confusion matrices of both annotators versus the gold standard are presented in Table 10 and Table 11 correspondingly. The resulting summary statistics obtained and the inter-annotator agreement (IAA) in terms of $F1$ score of 0.924 between the gold n2c2 annotations and of our two annotators and $F1$ score of 0.989 between our annotators are shown in Table 12.

4.2. Translation to OxCRIS data

The performance metrics and the token-level confusion matrix of the Med7 model trained on n2c2 data from MIMIC-III and applied to OxCRIS data are presented in Table 5 and in Table 5 correspondingly. Direct comparison to the results presented in Table 4 (*F1=0.762 vs. F1=F*1 0.762 vs. *F* 10.944) clearly shows the problem of direct transferability of NER models trained on different data sources.

4.3. Comparison to MedXN system

We also compared our Med7 system to a widely used alternative MedXN system for medication extraction and concepts normalisation. MedXN was chosen since it has shown better performance then MedEx [28] and cTAKES [68] and close to that of CLAMP [69]. Our Med7 system was compared to MedXN on both, MIMIC and OxCRIS data. For the sake of simplicity we considered only the lenient metric (i.e., $\alpha = 1$). The results are summarised in Table 6.

5. Discussion

The developed named-entity recognition model for clinical concepts in unstructured medical records was trained to recognise seven categories, such as drug names, including both generic and brand names, dosage of the drugs, their strength, the route of administration, prescription duration and the frequency. The data for model development and testing was sourced from the n2c2 challenge, comprising a collection of 303 and 202 documents for training and testing respectively, which represent a sample from the MIMIC-III electronic health records. We demonstrated (Section 3.6) that collecting more annotated examples would improve the model accuracy and therefore implemented two approaches for obtaining more annotations: noisy labelling and active learning with 'human-in-the-loop'. For the noisy labelling, we create a list of unique patterns for each of the seven categories, sourced from the training corpus and from external resources available on the internet, and then used regular expression with string pattern matching to assign labels to tokens. Our two annotators were trained by closely following the official 2018 n2c2 annotation guidelines and demonstrated a high level of inter-annotator agreement among themselves ($F1 = 0.989$) as well as a high-level of concordance ($F1 = 0.924$) with the gold-annotations provided by the organisers of 2018 n2c2 Challenge (cf. Table 12).

The overall (micro-averaged) performance of the NER model across all seven categories was $F1 = 0.957$ (0.893), with Precision = 0.982 (0.916) and Recall = 0.933 (0.871) for lenient (strict) estimates. More detailed breakdown of the performance for each of the categories is presented in Table 4. The performance for 'Duration' and 'Frequency' categories was poorer. There were intrinsically fewer cases of 'Duration' (~1.5%) appeared in texts and these concepts were also ambiguously annotated as mentioned in Section 3.7. A similar situation was also

Table 6
Comparison on the 202 documents from the n2c2 test set and 134 documents from the OxCRIS test set. All metrics are presented as lenient ($\alpha = 1$) to allow for flexible comparison. Abbreviations as follows: P – Precision, R – Recall and F – $F1$ score.

	n2c2						OxCRIS					
	MedXN			Med7			MedXN			Med7		
	P	R	F	P	R	F	P	R	F	P	R	F
Dosage	0.236	0.071	0.111	0.957	0.904	0.931	0.203	0.064	0.097	0.656	0.833	0.734
Drug	0.812	0.692	0.748	0.984	0.956	0.971	0.804	0.741	0.771	0.975	0.977	0.976
Duration	0.514	0.633	0.567	0.953	0.854	0.901	0.933	0.389	0.549	0.883	0.934	0.908
Form	0.801	0.798	0.799	0.983	0.947	0.965	0.601	0.601	0.601	0.924	0.968	0.946
Frequency	0.871	0.845	0.858	0.989	0.969	0.979	0.549	0.684	0.609	0.941	0.944	0.942
Route	0.744	0.759	0.751	0.973	0.954	0.964	0.251	0.503	0.333	0.882	0.938	0.909
Strength	0.721	0.837	0.774	0.992	0.836	0.907	0.933	0.894	0.913	0.996	0.917	0.955
Average (micro)	0.777	0.711	0.742	0.982	0.933	0.957	0.725	0.682	0.703	0.941	0.947	0.944
Average (macro)	0.685	0.668	0.669	0.977	0.919	0.947	0.599	0.561	0.642	0.901	0.932	0.914

Table 7
A list of examples of typical errors produced by the NER model.

Error type			Gold standard		NER prediction	
			Text span	Label	Text span	Label
1	Correct	(COR)	Aspirin	Drug	Aspirin	Drug
2	Incorrect	(INC)	25	Strength	25	Dosage
3	Partial	(PAR)	Augmentin	Drug	Augmentin XR	Drug
4	Partial	(PAR)	For 3 weeks	Duration	3 weeks	Duration
5	Partial	(PAR)	p.r.n.	Frequency	prn	Frequency
6	Missing	(MIS)	Tablet	Form	–	–
7	Spurious	(SPU)	–	–	Codeine	Drug

Table 8
Token-level confusion matrix of the predicted entities versus the ground truth labels. Spurious examples correspond to predicted entity boundary and type which do not exist in ground-truth annotations and partial matches correspond to predicted entity boundary overlap with golden annotation, but they are not the same. Missing entities correspond to ground-truth annotation boundaries which were not identified.

	Predicted categories								
	Dosage	Drug	Duration	Form	Frequency	Route	Strength	Missed	Partial
True categories									
Dosage	2225	0	6	10	24	1	16	200	199
Drug	2	9796	0	7	0	4	1	449	316
Duration	6	0	277	0	8	0	2	39	46
Form	38	31	0	3864	1	65	6	90	264
Frequency	1	3	4	5	3144	2	0	108	745
Route	3	4	0	43	1	3312	1	108	41
Strength	38	3	0	1	2	0	3304	650	232
Spurious	20	120	6	4	7	22	3		

Table 9
The number of the gold and manually annotated entities for the inter-annotator agreement evaluation corpus, comprising 10 randomly sampled texts from the test set of 202 documents.

Types of annotated entities	Gold (n2c2)	Annotator 1	Annotator 2
Dosage	128	139	139
Drug	519	530	526
Duration	28	31	32
Form	234	246	238
Frequency	193	196	201
Route	179	167	167
Strength	200	212	205
Number of documents	10	10	10

observed for the ’Frequency’ category, where in spite of a good number of the annotated examples (~14%), the ambiguity in the presentation of text spans was higher, which resulted in a large number of partial matches (cf. Table 8). Another reason for poor performance for both ’Duration’ and ’Frequency’ was due to inconsistent annotations, where the same text string appeared in both categories.

Self-supervised pre-training of deep learning models has shown its efficiency in many NLP tasks. We experimented with a number of architectural variations of the width and depth of convolutional layers as well as the size of the embedding rows. Empirically, and as confirmed by other studies [70], larger models, with more parameters, tend to achieve better results. Interestingly, the larger model (Width = 256, Depth = 16, Embeddings = 10,000) outperformed the default one (Width = 96, Depth = 4, Embeddings = 2000) by a small margin ($F1_{256} = 0.893$ vs $F1_{96} = 0.884$) however, the differences were more visible for ’Duration’ ($F1_{256} = 0.773$ vs $F1_{96} = 0.729$) and ’Strength’ ($F1_{256} = 0.848$ vs $F1_{96} = 0.801$). The better performance resulted at the expense of the training time, its size on a disk and the memory consumption. We publicly released the pre-trained neural network weights for various architectures through the dedicated GitHub repository.[3]

Another objective of this work was to estimate the degree of transferability of the developed information extraction model to another clinical domain. We evaluated how the Med7 model, trained on a collection of discharge letters from the intensive care unit in the US (MIMIC), performed on the secondary care mental health medical

Table 10
Token-level confusion matrix of the annotator 1 versus the gold-standard annotations provided by 2018 n2c2 challenge.

	Annotator 1								
	Dosage	Drug	Duration	Form	Frequency	Route	Strength	Missed	Partial
Gold (n2c2)									
Dosage	104	0	1	3	0	0	2	17	4
Drug	0	473	0	3	0	1	0	27	21
Duration	0	0	19	0	0	0	0	2	7
Form	1	4	0	201	0	2	0	7	21
Frequency	1	0	0	0	172	0	1	2	17
Route	2	2	0	2	0	156	0	15	2
Strength	2	1	0	0	0	0	171	4	28
Spurious	25	29	4	16	7	6	10		

[3] https://github.com/kormilitzin/med7.

Table 11
Token-level confusion matrix of the annotator 2 versus the gold-standard annotations provided by 2018 n2c2 challenge.

	Annotator 2								
	Dosage	Drug	Duration	Form	Frequency	Route	Strength	Missed	Partial
Gold (n2c2)									
Dosage	104	0	1	3	0	0	2	17	4
Drug	0	472	0	3	0	1	0	30	20
Duration	0	0	19	0	0	0	0	2	7
Form	0	3	0	201	0	2	0	9	21
Frequency	0	0	1	0	172	0	0	2	18
Route	2	2	0	2	0	156	0	15	2
Strength	3	1	0	0	4	0	171	3	21
Spurious	26	28	4	8	7	6	10		

Table 12
The evaluation results of the inter-annotator agreement on a random selection of 10 documents from the 202 test texts. A pair-wise comparison between each of the annotators and the gold-annotated documents as well as the direct comparison between the both annotators.

	Annot. 1 vs. gold			Annot. 2 vs. gold			Annot. 1 vs. annot. 2		
	Pr	Re	*F*1	Pr	Re	*F*1	Pr	Re	*F*1
Dosage	0.777	0.824	0.801	0.816	0.783	0.799	0.986	0.986	0.986
Drug	0.935	0.935	0.935	0.971	0.941	0.962	0.998	0.991	0.994
Duration	0.812	0.923	0.867	0.861	0.9233	0.911	0.969	1.000	0.984
Form	0.933	0.941	0.937	0.933	0.922	0.927	1.000	0.967	0.983
Frequency	0.945	0.984	0.964	0.897	0.984	0.938	0.975	1.000	0.987
Route	0.946	0.883	0.913	0.946	0.883	0.913	1.000	1.000	1.000
Strength	0.941	0.946	0.944	0.941	0.912	0.926	1.000	0.962	0.981
Average (micro)	0.921	0.928	0.924	0.931	0.917	0.924	0.994	0.985	0.989
Average (macro)	0.901	0.921	0.911	0.889	0.909	0.905	0.991	0.986	0.988

Table 13
Token-level confusion matrix of the Med7 model trained on MIMIC-III and applied to 134 manually annotated documents from the Oxford instance (OxCRIS) of the UK-CRIS electronic medical records network.

	Med7-predicted categories: before fine-tuning on OxCRIS								
	Dosage	Drug	Duration	Form	Frequency	Route	Strength	Missed	Partial
Gold annotated									
Dosage	18	0	0	0	0	0	12	17	1
Drug	0	535	0	0	0	0	0	18	15
Duration	0	0	18	0	1	0	0	158	1
Form	0	2	0	34	0	1	0	20	2
Frequency	0	7	0	25	86	40	1	114	7
Route	0	0	0	3	3	23	0	6	0
Strength	3	0	0	0	0	0	238	31	4
Spurious	1	44	1	1	8	2	3		

Table 14
Token-level confusion matrix of the Med7 model trained on MIMIC-III and applied to 134 manually annotated documents from the Oxford instance (OxCRIS) of the UK-CRIS electronic medical records Network.

	Med7-predicted categories: after fine-tuning on OxCRIS								
	Dosage	Drug	Duration	Form	Frequency	Route	Strength	Missed	Partial
Gold annotated									
Dosage	39	0	0	0	0	0	1	7	1
Drug	0	553	0	2	0	0	0	11	4
Duration	0	0	177	0	1	0	0	13	20
Form	0	0	0	61	1	1	0	0	0
Frequency	1	1	0	2	279	1	0	12	6
Route	0	0	0	0	0	30	0	2	0
Strength	16	1	0	0	0	0	242	6	11
Spurious	4	12	26	1	16	2	0		

records in the UK (CRIS). The Med7 model was purposely designed to recognise non-context related medical concepts, such as drug names, strength, dosage, duration, route, form and frequency of administration and we expected to see a comparable level of the model performance across the both EHR systems. To consistently validate the transferability of the Med7 model, a random sample of 670 gold-annotated examples from OxCRIS were split into training (536) and test (134) data sets (cf. Table 5). We compared the performance of the Med7 model without and with fine-tuning on OxCRIS. The direct application of Med7 on the testing set of 134 documents, resulted in a quite poor performance ($F1 = 0.762$). We investigated the cases where the model was predicted incorrectly and in the majority of them, the main reason for poor performance was due to differences in the language presentation of the concepts. For examples, the model largely missed concepts labelled as 'Frequency' in OxCRIS, such as "ON" ("every night"), "OD" ("every day"), "BD" ("twice daily"), "OM" ("every morning"), "mane" and "nocte". Then, we fine-tuned the Med7 model on the training set of OxCRIS (536 documents) and evaluated it on the same testing set as before of 134 documents. Despite the small number of training examples in OxCRIS, leveraging the transfer learning approach, whereby re-using the pre-trained Med7 model on MIMIC and further fine-tuned on OxCRIS data, resulted in higher accuracy ($F1 = 0.944$) comparable with training and testing on the same domain (cf. Table 5).

One particular strength of this project is in the interoperability of the developed model with other generic deep learning NLP libraries, such as HuggingFace and Thinc as well as straightforward integration with pipelines developed under the spaCy framework. This allows to customise the Med7 model and include other pipeline components, such as negation detection, entity relations extraction and to map the extracted concepts onto the universal medical language system (UMLS). Normalisation of concepts to UMLS categories will allow us to systematically parse electronic medical records into structured and consistent tabular form which will be ready for downstream epidemiological analyses. Additionally, the developed model naturally integrates into the Prodigy annotation tool, which allows to efficiently collect more gold-annotated examples. It is also worth mentioning that the Med7 model is designed to run on standard CPUs, rather than expensive GPUs. This fact will allow researchers without access to expensive and complex infrastructure to develop fast and robust pipelines for clinical natural language processing.

However, two limitations should be noted. First, is that some of the categories are naturally underrepresented that impacts the accuracy of the NER model. It was observed empirically that the number of annotated 'Duration' entities was intrinsically skewed in the medical records, in contrast to drug names and strength, making it more challenging to train a robust model to accurately identify these entities. Interestingly, the same pattern of the number of reported mentions of the 'Duration' category persists in both, MIMIC and OxCRIS data, which might be indicative of a general clinical reporting pattern. A second limitation of this study is related to a low number of the manually-annotated examples in OxCRIS, in order to run more rigorous evaluations of the transferability of the Med7 model across all seven categories.

Future research into the robust clinical information extraction system will need to address the feasibility of deploying the model in the UK-CRIS Network Trust members and evaluate its transferability. The aim is to furnish clinical researchers with an open source and a robust tool for structuring free-text patients' data for downstream analytical tasks.

6. Conclusion

In this work we developed and validated a clinical named-entity recognition model for free-text electronic health records. The model was developed using the MIMIC-III free-text data and trained on a combination of the manually annotated data from the 2018 n2c2 challenge, on a random sample from MIMIC-III with noisy labels and manually annotated data using an active learning tool. To maximise the utilisation of a large amount of unstructured free-text data and alleviate the problem of training from limited and imbalanced data, we used self-supervised learning to pre-train the weights of the NER neural network model. We demonstrated that transfer learning plays an essential role in developing a robust model applicable across different clinical domains and the developed Med7 model does not require an expensive infrastructure and can be used on standard machines with CPU. Further research is needed to improve recognition of naturally underrepresented concepts and we are planning to address this problem, as well as extracted concepts normalisation and UMLS linkage in our future releases of the Med7 model.

Data availability

The data that support the findings of this study are available by application through the UK-CRIS network. Owing to the sensitivity of secondary care clinical information, access is dependent on receiving research approvals from NHS trust oversight bodies at Oxford Health NHS Foundation Trust.

Acknowledgements

The study was funded by the National Institute for Health Research's (NIHR) Oxford Health Biomedical Research Centre (BRC-1215-20005). This work was supported by the UK Clinical Records Interactive Search (UK-CRIS) system funded and developed by the NIHR Oxford Health BRC at Oxford Health NHS Foundation Trust and the Department of Psychiatry, University of Oxford. AK, NV, QL, ANH are funded by the MRC Pathfinder Grant (MC_PC_17215). We are thankful to the organisers of the n2c2 2018 Challenge for providing annotated corpus and the annotation guidelines. A.N.-H. reports funding support from Johnson & Johnson, as well as salary support from Akrivia Health.

The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health.

We would also like to acknowledge the work and support of the Oxford CRIS Team: Tanya Smith, Head of Research Informatics and Biomedical Research Centre (BRC) CRIS Theme Lead and Lulu Kane, Adam Pill and Suzanne Fisher, CRIS Academic Support and Information Analysts.

Appendix A. The evaluation schema for extracted concepts

In order to evaluate the output of the NER system, we adopted the notations developed for different categories of errors [71] and the evaluation schema introduced in SemEval'13 (cf. Eq. (1)). The following types of evaluation errors were considered (Table 7):

Appendix B. Inter-annotator agreement analysis

We estimated the level of concordance between the gold-annotated corpus from the n2c2 2018 challenge and two trained annotators. The annotators closely followed the same annotation guidelines as used in the challenge. Ten documents were sampled at random from 202 documents comprising the test set. The distribution of gold-annotated tokens and by two annotators is presented in Table 9.

We examined the cases where our two annotators labelled the concepts of interests differently than those found in the gold-annotated data set provided by the n2c2 team.

Appendix C. Fine-tuning on OxCRIS

Tables 13 and 14.

Appendix D. Reproducibility

Tokens have been represented using the standard GloVe embedding. The neural network architecture was based on a convolutional neural network encoder with `width=256` and `depth=16` of convolutional layers with the fully connected hidden layer of size 128 before the softmax final layer. The cross-entropy loss function was optimised using Adam optimiser with momentum 0.92, the batch size of 64, learning rate of 0.0015 and gradient clipping at 1.0. All experiment and pre-trained models are available at https://github.com/kormilitzin/med7 https://github.com/kormilitzin/med7https://github.com/kormilitzin/med7.

Appendix E. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.artmed.2021.102086.

References

[1] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. 2018. arXiv preprint arXiv:1802.05365.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems 2017: 5998–6008.

[3] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint arXiv:1810.04805.

[4] Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inf 2018;88:11–9.

[5] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. CVPR09 2009.

[6] Lin T, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, et al. COCO: common objects in context. 2014. CoRR abs/1405.0312. http://arxiv.org/abs/1405.0312.

[7] Entzeridou E, Markopoulou E, Mollaki V. Public and physician's expectations and ethical concerns about electronic health record: benefits outweigh risks except for information security. Int J Med Inf 2018;110:98–107.

[8] Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inf Assoc 2008;15(1):14–24.

[9] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. Mimic-iii, a freely accessible critical care database. Sci Data 2016;3:160035.

[10] Schütze H, Manning CD, Raghavan P. Introduction to information retrieval. Proceedings of the international communication of association for computing machinery conference, Vol. 4 2008.

[11] Patel VM, Gopalan R, Li R, Chellappa R. Visual domain adaptation: a survey of recent advances. IEEE Signal Process Mag 2015;32(3):53–69.

[12] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Mach Learn Res 2016;17(1): 2030–96.

[13] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[14] Baevski A, Edunov S, Liu Y, Zettlemoyer L, Auli M. Cloze-driven pretraining of self-attention networks. 2019. arXiv preprint arXiv:1903.07785.

[15] Ratner AJ, Hancock B, Ré C. The role of massively multi-task and weak supervision in software 2.0. CIDR 2019.

[16] Dalianis H. Clinical text mining: secondary use of electronic patient records. Springer; 2018.

[17] Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inf Assoc 2020;27(3):457–70.

[18] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(suppl_1):D267–70.

[19] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inf 2012;45(5):885–92.

[20] Zhou G, Su J. Named entity recognition using an HMM-based chunk tagger. Proceedings of the 40th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics 2002:473–80.

[21] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA symposium, American Medical Informatics Association 2001:17.

[22] Kay M. Algorithm schemata and data structures in syntactic processing, Technical Report. 1980.

[23] Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. Medex: a medication information extraction system for clinical narratives. J Am Med Inf Assoc 2010;17(1):19–24.

[24] Jonnalagadda SR, Li D, Sohn S, Wu ST-I, Wagholikar K, Torii M, et al. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. J Am Med Inf Assoc 2012;19(5):867–74.

[25] Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inf Assoc 2011;18(5): 580–7.

[26] Sohn S, Wagholikar KB, Li D, Jonnalagadda SR, Tao C, Komandur Elayavilli R, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. J Am Med Inf Assoc 2013;20(5):836–42.

[27] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inf Assoc 2010;17(5):507–13.

[28] Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. Medxn: an open source medication extraction and normalization tool for clinical text. J Am Med Inf Assoc 2014;21(5):858–65.

[29] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inf Assoc 2018;25(3):331–6.

[30] Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. PLoS Comput Biol 2013;9(2):e1002854.

[31] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 2013:3111–9.

[32] Pennington J, Socher R, Manning C. Glove: global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014:1532–43.

[33] Kalyan KS, Sangeetha S. Secnlp: a survey of embeddings in clinical natural language processing. 2019. arXiv preprint arXiv:1903.01039.

[34] Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018. arXiv preprint arXiv:1801.06146.

[35] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized bert pretraining approach. 2019. arXiv preprint arXiv:1907.11692.

[36] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems 2019:5754–64.

[37] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36 (4):1234–40.

[38] Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. 2019. arXiv preprint arXiv:1904.05342.

[39] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical bert embeddings. 2019. arXiv preprint arXiv:1904.03323.

[40] Neumann M, King D, Beltagy I, Ammar W. Scispacy: fast and robust models for biomedical natural language processing. 2019. arXiv preprint arXiv:1902.07669.

[41] Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inf Assoc 2019;27(1):3–12.

[42] Vaci N, Liu Q, Kormilitzin A, De Crescenzo F, Kurtulmus A, Harvey J, et al. Natural language processing for structuring clinical text data on depression using UK-CRIS. Evid Based Ment Health 2020;23(1):21–6.

[43] Xie Q, Hovy E, Luong M-T, Le QV. Self-training with noisy student improves imagenet classification. 2019. arXiv preprint arXiv:1911.04252.

[44] Natarajan N, Dhillon IS, Ravikumar PK, Tewari A. Learning with noisy labels. Advances in neural information processing systems 2013:1196–204.

[45] Provilkov I, Emelianenko D, Voita E. Bpe-dropout: simple and effective subword regularization. 2019. arXiv preprint arXiv:1910.13267.

[46] Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, et al. Not enough data? Deep learning to the rescue!. 2019. arXiv preprint arXiv:1911.03118.

[47] Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J 2019:1–22.

[48] Trask A, Michalak P, Liu J. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. 2015. arXiv preprint arXiv: 1511.06388.

[49] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv preprint arXiv:1301.3781.

[50] Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: Rxnorm at 6 years. J Am Med Inf Assoc 2011;18(4):441–8.

[51] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic acids research 2018;46 (D1):D1074–82.

[52] Solution, in: Proceedings of the human language technology conference of the NAACL, companion volume: short papers, NAACL-Short '06, Association for Computational Linguistics, USA, 2006, pp. 57–60.

[53] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. 2016. arXiv e-prints arXiv:1606.05250.

[54] Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA 2011. pp. 142–150. http://www.aclweb.org/anthology/P11-1015.

[55] Hofer M, Kormilitzin A, Goldberg P, Nevado-Holgado A. Few-shot learning for named entity recognition in medical text. 2018. arXiv preprint arXiv:1811.05468.

[56] Gligic L, Kormilitzin A, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. Neural Networks 2020;121:132–9.

[57] Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC medical informatics and decision making 2019;19(1):1.

[58] Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. to appear.

[59] Montani I, Honnibal M. Prodigy: a new annotation tool for radically efficient machine teaching. Artif Intell 2018. arXiv: to appear.

[60] Serrà J, Karatzoglou A. Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks. Proceedings of the Eleventh ACM Conference on Recommender Systems 2017:279–87.

[61] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. 2016. arXiv preprint arXiv: 1603.01360.

[62] Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Crit Care Med 2020;48(2):210.

[63] Morrill J, Kormilitzin A, Nevado-Holgado A, Swaminathan S, Howison S, Lyons T. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. In: 2019 Computing in cardiology conference (CinC); 2019.

[64] Ren J, Liu PJ, Fertig E, Snoek J, Poplin R, Depristo M, et al. Likelihood ratios for out-of-distribution detection. Advances in neural information processing systems 2019:14680–91.

[65] Kulkarni V, Mehdad Y, Chevalier T. Domain adaptation for named entity recognition in online media with word embeddings. 2016. arXiv preprint arXiv: 1612.00148.

[66] Chiticariu L, Krishnamurthy R, Li Y, Reiss F, Vaithyanathan S. Domain adaptation of rule-based annotators for named-entity recognition tasks. Proceedings of the 2010 conference on empirical methods in natural language processing 2010: 1002–12.

[67] Peng N, Dredze M. Multi-task domain adaptation for sequence tagging. 2016. arXiv preprint arXiv:1608.02689.

[68] Fan Y, Wen A, Shen F, Sohn S, Liu H, Wang L. Evaluating the impact of dictionary updates on automatic annotations based on clinical NLP systems. AMIA Summits Transl Sci Proc 2019;2019:714.

[69] Weeks HL, Beck C, McNeer E, Bejan CA, Denny JC, Choi L. medextractr: a medication extraction algorithm for electronic health records using the r programming language. MedRxiv 2019. 19007286.

[70] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019. arXiv preprint arXiv:1910.10683.

[71] Chinchor N, Sundheim B. MUC-5 evaluation metrics. Fifth message understanding conference (MUC-5): proceedings of a conference held in Baltimore, Maryland, August 25–27, 1993 1993. https://www.aclweb.org/anthology/M93-1007.