Drug Safety (2019) 42:99–111
https://doi.org/10.1007/s40264-018-0762-z

ORIGINAL RESEARCH ARTICLE

Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0)

Abhyuday Jagannatha[1] · Feifan Liu[2] · Weisong Liu[3,4] · Hong Yu[1,3,4,5]

Published online: 16 January 2019

Abstract

Introduction This work describes the Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0) corpus and provides an overview of the MADE 1.0 2018 challenge for extracting medication, indication, and adverse drug events (ADEs) from electronic health record (EHR) notes.

Objective The goal of MADE is to provide a set of common evaluation tasks to assess the state of the art for natural language processing (NLP) systems applied to EHRs supporting drug safety surveillance and pharmacovigilance. We also provide benchmarks on the MADE dataset using the system submissions received in the MADE 2018 challenge.

Methods The MADE 1.0 challenge has released an expert-annotated cohort of medication and ADE information comprising 1089 fully de-identified longitudinal EHR notes from 21 randomly selected patients with cancer at the University of Massachusetts Memorial Hospital. Using this cohort as a benchmark, the MADE 1.0 challenge designed three shared NLP tasks. The named entity recognition (NER) task identifies medications and their attributes (dosage, route, duration, and frequency), indications, ADEs, and severity. The relation identification (RI) task identifies relations between the named entities: medication-indication, medication-ADE, and attribute relations. The third shared task (NER-RI) evaluates NLP models that perform the NER and RI tasks jointly. In total, 11 teams from four countries participated in at least one of the three shared tasks, and 41 system submissions were received in total.

Results The best systems F_1 scores for NER, RI, and NER-RI were 0.82, 0.86, and 0.61, respectively. Ensemble classifiers using the team submissions improved the performance further, with an F_1 score of 0.85, 0.87, and 0.66 for the three tasks, respectively.

Conclusion MADE results show that recent progress in NLP has led to remarkable improvements in NER and RI tasks for the clinical domain. However, some room for improvement remains, particularly in the NER-RI task.

Part of a theme issue on "NLP Challenge for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0)" guest edited by Feifan Liu, Abhyuday Jagannatha and Hong Yu.

✉ Hong Yu
hong.yu@umassmed.edu

[1] College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA

[2] Department of Quantitative Health Sciences and Radiology, University of Massachusetts Medical School, Worcester, MA, USA

[3] Department of Computer Science, University of Massachusetts, 220 Pawtucket St., Lowell, MA 01854-2874, USA

[4] Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA

[5] Bedford VAMC, Bedford, MA, USA

Key Points

The MADE (Medication and Adverse Drug Events from Electronic Health Records) 1.0 corpus comprises 1089 electronic health records with detailed named entity and relation annotations.

We provide benchmark results from and analysis of the MADE 1.0 corpus using system submissions in the MADE 1.0 challenge.

MADE 1.0 results suggest that machine learning systems can be useful for automated extraction of adverse drug events and related entities from unstructured texts but that room for improvement remains.

1 Introduction

An adverse drug event (ADE) is "an injury resulting from a medical intervention related to a drug" [1]. ADEs are the single largest contributor to hospital-related complications in inpatient settings [2] and account for approximately one-third of all hospital adverse effects (AE). They affect more than 2 million hospital stays annually [3] and prolong hospital length of stay by 1.7–4.6 days [4, 5]. These events also account for approximately two-thirds of all post-discharge complications, more than one-quarter of which are estimated to be preventable [6]. National estimates suggest that ADEs contribute at least an additional $US30 billion to US healthcare costs [7].

Likely ADEs should ideally be detected in randomized controlled trials (RCTs) before the relevant drug ever enters the market. However, the limited number of participants and inclusion/exclusion criteria reflecting specific subject characteristics (demographic, medical condition and diagnosis, age) [8] means that pre-marketing RCTs frequently miss ADEs. This assertion is supported by the fact that the rate at which the US FDA withdraws previously approved drugs in the first 16 years ranges from 21 to 27% [9]. Drug safety surveillance and post-marketing pharmacovigilance, "the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem" [10], are therefore vitally important tools for monitoring FDA-approved drug safety.

One of the earliest systems for post-marketing pharmacovigilance aimed at improving drug safety is spontaneous reporting systems (SRSs) such as the FDA Adverse Event Reporting System (FAERS), a voluntary SRS. Although SRSs have been highly successful for pharmacovigilance, they have limitations such as under-reporting [11, 12] and missing important patterns of drug exposure [13]. To counter these shortcomings, other resources have been proposed for pharmacovigilance, including biomedical literature [14] and social media [15]. However, biomedical literature has been shown to identify only a limited set of ADEs, mainly rare ADEs [16]. Social media also has challenges, such as incomplete and erroneous drug exposure patterns and duplication [17].

It is well-known that electronic health records (EHRs) contain rich ADE information and have been widely used for drug safety surveillance and pharmacovigilance [2, 6]. Unlike other resources, which are passive in nature, EHRs can be a rich resource for real-time or active pharmacovigilance and patient–drug surveillance. In addition, they can lead to better and more cost-effective patient management [18]. In 2009, the FDA initiated the mini-sentinel program to facilitate the use of routinely collected EHR data for active surveillance of marketed medical product safety [19]. For example, Yih et al. [20] showed an increased risk of intussusception after rotavirus vaccination in US infants.

However, most of the EHR-based pharmacovigilance and patient safety surveillance systems are based on the analyses of the structured data such as *International Statistical Classification of Diseases and Related Health Problems* (ICD) codes [21, 22]. It is well-known that ADEs are often buried in the clinical narrative [23–25] and are not separately recorded in diagnosis codes or other structured data fields. Even information that is expected to be reported in the structured fields, such as bodyweight, frequently appears only in EHR text [26]. In addition, necessary information that can be used to assess the causality of a medication and ADE, including temporal and causal relations, only exists in the narrative. However, extraction of ADE information from EHR narratives remains a challenge because manual data extraction is very costly. It is, therefore, a significant impediment to large-scale pharmacovigilance studies.

Natural language processing (NLP) may be a solution to provide fast, accurate, and automated ADE detection that can yield significant cost and logistical advantages over the aforementioned practices of manual chart review or voluntary reporting [27]. However, despite the advances in NLP, few methods are specifically developed for detecting ADE information from EHR notes. Several NLP systems use resources such as the Unified Medical Language System (UMLS) [28] to extract disease and drug mentions to generate ADE predictions using co-occurrence metrics. Such approaches may miss other important drug information (e.g., indication and dose) and would fail to capture temporal and/or causal associations between a drug and an ADE that may be explicitly expressed in EHR narratives. Moreover, different NLP systems have been evaluated on different gold

standards, making it challenging to identify state-of-the-art NLP technologies.

Therefore, we created the MADE (Medication and Adverse Drug Events from Electronic Health Records) 1.0 corpus, a publicly available, expert-curated benchmark of EHR notes that have been annotated with clinical named entities (i.e., drug name, dosage, route, duration, frequency, indication, ADE, and other signs and symptoms) and relations (ADE–drugname, indication–drugname, drug-name–attributes, etc.). The MADE corpus is the first dataset that provides detailed annotations for medication, indication, ADEs, their attributes and relations relevant to drug safety surveillance and pharmacovigilance studies.

Using this high-quality corpus as a benchmark, we designed three shared tasks (named entity recognition (NER), relation identification (RI), and NER-RI) to assess the state-of-the-art NLP technologies that have the potential to improve downstream pharmacovigilance-related tasks. These shared tasks were organized in the First Natural Language Processing Challenge for Detecting MADE hosted by the University of Massachusetts (Amherst, Lowell, and Worcester, USA) from August 2017 to March 2018. In this paper, we first describe the MADE corpus, then document the shared tasks and provide a comprehensive report of system submissions in the MADE challenge. The main contributions of this paper are as follows.

- Present the first richly annotated and publicly available EHR data for ADE detection and drug surveillance research.
- Describe the carefully designed schema and release the detailed annotation guidelines, which would be a valuable resource to not only drug safety but also any other data-driven clinical informatics research.
- Introduce three shared tasks in the MADE challenge and report the system submissions and results.
- Perform an ensemble-based system aggregation that shows that the top systems are complementary and can be further integrated to push the boundaries of extracting medication, indication, and ADEs from EHRs.

2 Related Work

Natural language processing techniques have been widely applied to biomedicine [28–33]. Much of NLP research in the biomedical domain has centered on NER and normalization tasks. Examples of shared tasks in this domain include BioNLP [34], BioCreAtivE [35], i2b2 shared NLP tasks [36], and ShARe/CLEF evaluation tasks [37].

Existing NLP approaches for EHR ADE detection can be grouped into rule-based, lexicon-based, supervised machine learning, and hybrid approaches. For example, Li et al. [38] built an NLP system with the knowledge of a domain expert. Melton and Hripcsak [39] applied the NLP system MedLEE, a rule-based semantic parser, to detect concepts. Similarly, Humphreys et al. [40] applied MedLEE to map free text to the UMLS concepts and semantic types. UMLS [41] is a resource that combines multiple biomedical and clinical resources into a unified ontology. Rochefort et al. [42] developed supervised machine learning classifiers to classify whether a clinical note contains deep venous thromboembolisms (DVT) and pulmonary embolism (PE) using bag-of-words from EHR narratives. Haerian et al. [43] applied distance supervision to identify terms (e.g., including suicidal, self-harm, and diphenhydramine overdose) associated with an assigned suicide ICD-9 code and then used those terms to recover suicide events. Wang et al. [44] used MetaMap to identify drugs mentioned in the text threads of online health forums. Nikfarjam et al. [45] annotated ADE information on user posts from Daily-Strength and Twitter. They then used word-embedding models and conditional random fields (CRF) for prediction. Li et al. [46] developed NLP methods to extract medication information (e.g., drug name, indication, contraindication) and adverse events from FDA drug labels. Duke and Friedlin [47] applied MetaMap to identify ADEs from structured product labels.

Related works on corpora for clinical NLP research include the GENIA corpus [48] and the TREC Genomics [49]. Shared tasks such as BioNLP [34] and BioCreAtivE [35] have been widely used to train NLP applications. Other annotated corpora include the disease corpus [50], the BioScope corpus [51], and the MEDLINE abstract corpus from Gurulingappa et al. [52]. The corpus closest to ours is the i2b2 2009 corpus by Uzuner et al. [53], which provides annotations for medication and related named entities. However, our work extends the annotation schema used in the i2b2 corpus and provides a common dataset for medication and ADEs. Another similar work is that by Henriksson et al. [54], in which they annotated a dataset focused towards ADE extraction. In contrast to their work, the MADE corpus also

Table 1 The overall statistics for the MADE corpus

Description	Mean ± SD (max, min)
Words/note	948.2 ± 484.9 (3804, 76)
Named entity annotations/note	72.6 ± 52.6 (363, 0)
Relation annotations/note	25.0 ± 24.1 (181, 0)
Notes/patient	51.8 ± 40.3 (166, 1)

MADE Medication and Adverse Drug Events from Electronic Health Records, *SD* standard deviation

Table 2 Annotation counts, and word counts for each named entity type

Named entity type	Number of annotations	Total annotated words
ADE	1940	3255
Indication	3804	8240
Other SSD	39,384	82,956
Severity	3908	5069
Drugname	15,902	19,075
Dosage	5694	11,820
Duration	898	1768
Frequency	4806	11,400
Route	2667	2805

ADE adverse drug event, *SSD* sign, symptom, or disease

provides annotation for medication details such as dosage frequency, etc., which are extremely relevant for pharmacovigilance studies.

3 The MADE Corpus

The MADE corpus comprises 1089 fully de-identified longitudinal EHR notes from 21 randomly selected patients with cancer at the University of Massachusetts Memorial Medical Center. Therefore, the notes include diverse note types such as discharge summaries, consultation reports, and other clinic notes (Table 1).

We used an iterative process throughout the annotation, going back and forth between document annotations and establishing annotation guidelines. In this process, we created a comprehensive annotation guideline[1], which addresses various aspects on how to handle language variations and ambiguities in clinical narratives related to this annotation task. The guideline adapted and substantially extended the 2009 i2b2 shared task of the Medication Challenge annotation guideline [53]. The MADE annotation guideline is designed with a focus on extracting ADEs and other relevant clinical information. It defines nine named entity types and seven relation types. The relation types define relationships between pairs of annotated named entities. A succinct overview of the annotation categories is provided in the following subsections. The entities and relation types are described in detail in the text of the annotation guideline.

[1] The complete annotation guideline and dataset is available at bio-nlp.org/dataset/made1.

Table 3 Annotation counts and relation length for each relation category

Relation type	Occurrences	Relation length
ADE–drugname	2612	82 ± 187 (3662, 1)
SSD–severity	4035	4.7 ± 34.41 (1861, 0)
Drugname–route	3006	18 ± 25 (224, 1)
Drugname–dosage	6043	11 ± 22 (230, 0)
Drugname–duration	1053	20 ± 27 (273, 1)
Drugname–frequency	5149	25 ± 30 (295, 1)
Indication–drugname	5430	96 ± 164 (2742, 1)

A relation is defined as a relation between two named entities. The relation length format is "mean ± SD (max, min)"

ADE adverse drug event, *SSD* sign, symptom, or disease

3.1 Named Entity Types

The named entity types can be broadly defined as either events or attributes. Events are annotations that denote a change in a patient's medical status. This includes the prescription of a medication and identification of a symptom or diagnosis. Events have attributes, including severity and information related to medications (e.g., dosage). The occurrence of each named entity type is provided in Table 2. As evident from the table, there is a large label imbalance in the data, which means developing an NLP system to detect those entity types is challenging.

Based on their context, the named entity annotations can be clustered into those related to sign, symptom, or disease (SSD) mentions and those related to medication (drugname) mentions. The two categories are described in the following subsections.

3.1.1 Sign, Symptom, or Disease (SSD)

Annotations in the SSD group define events and properties relevant to SSD mentions. The relevant named entity types are ADEs, indication, other SSD, and severity. ADEs, indication, and other SSDs are event annotations.

ADE ADEs are a type of SSD. They are adverse events caused by a drugname. An ADE annotation requires a direct linguistic cue that links the adverse effect to a drugname, e.g., "Patient had *anaphylaxis* after getting penicillin."

Indication An indication is annotated if it is explicitly linked to a medication, e.g., "The patient was troubled with *mouth sores* and is being treated with Actiq."

Other SSD Any SSD event that is not annotated as an indication or ADE is categorized as "other SSD". In our EHRs, other SSDs frequently occur in the history section of notes, e.g., "headache in the back of the head."

Severity These annotations are attributes of SSDs that indicate the severity, e.g., acute, mild, and severe, of a particular SSD.

3.1.2 Medication (Drugname)

Medication includes drugname and its attributes.

Drugname The drugname annotation includes descriptions that denote any medication, procedure, or therapy prescription, e.g., warfarin, propofol, chemotherapy, etc.

Duration Duration is the time range for the administration of the drugname, as explicitly described in the notes, e.g., 2 weeks, 15 h.

Dosage Dosage is the amount of drug in a unit dose. It is a numerical value and is an attribute of drugname entity, e.g., two tablets, 4 ml/h.

Frequency Frequency is the rate of administration of the drug and is an attribute of drugname, e.g., every hour, three times daily (t.i.d.), four times daily (q.i.d).

Route Route is the path through which a drug is taken into the body. It is an attribute of drugname entity, e.g., orally, central line.

3.2 Relation Types

Table 3 shows the seven relation types and their frequencies in the MADE corpus. A relation type is defined as a relation between two different named entity types. A brief description of each relation, along with the relevant named entities, is provided in the following.

Drugname Attribute Relations

The MADE corpus contains four different relation types that describe a relation between the drugname entity and its various attributes:

- Drugname–dosage
- Drugname–route
- Drugname–frequency
- Drugname–duration.

The attributes (dosage, route, frequency, duration) are properties of the drugname entity.

SSD–severity Severity is an attribute to an SSD (ADE, indication, other SSD). It is typically a modifier (e.g., mild) for an annotated entity (e.g., fever).

ADE–drugname ADE is an adverse effect of the prescription of the drugname entity.

Indication–drugname The drugname entity has been prescribed as a direct treatment for the indication entity.

In the MADE corpus, the relations between the named entities can occur within a sentence or across multiple sentences in a note. Table 3 provides the relation length in characters. The ADE–drugname and indication–drugname relations have heavy long tails, indicating that, in several instances, they connect named entities that are several sentences apart. We discuss the implications of this trend in Sect. 4.3.

3.3 Annotators

The annotation process involved multiple annotators, including physicians, biologists, linguists, and biomedical database curators. Annotators were used both in document annotation and in the development of annotation guidelines. The following process was used to annotate each file. The first annotator individually labeled the span and type of named entities and relations. A second annotator then reviewed the annotations and modified them to produce the final version. This annotation process was used to reduce the annotation cost of each document while ensuring high annotation quality.

Since the annotations provided by the two annotators in this process were not independent, they could not be used to obtain estimates of inter-annotator agreements (IAAs). To get a fair IAA estimate, we performed a smaller study wherein five annotators independently annotated three documents from our corpus. We used the Fleiss' kappa (κ) [55] measure of IAA. The κ for the named entity annotation and relation annotation agreement were 0.628 and 0.424, respectively. The relation κ measures the agreement in both named entity and relation prediction. The added complexity of combined named entity and relation annotation may explain its comparatively lower annotation agreement value. However, both values fall in the fair-to-significant agreement range [56], suggesting that our annotations are reliable for evaluating information extraction systems.

3.4 De-identification

The EHR data were de-identified using the Safe Harbor methods defined in 45 CFR 164.514b(2) by the US Department of Health and Human Services. First, the EHR data were processed by a publicly available de-identifier [57] so that the 18 types of Safe Harbor identifiers would be automatically annotated. Second, each clinical note was manually reviewed to ensure all identifiers were marked fully and correctly during the annotation process. All the marked identifiers were finally removed before releasing the data to the teams participating in the MADE challenge.

3.5 Evaluation Script

To standardize the evaluation of NER and RI, we developed an evaluation script[2]. Our evaluation script uses bioc[3]

[2] The evaluation script is included with the MADE data release.

[3] http://bioc.sourceforge.net; https://github.com/yfpeng/bioc.

format, a simple format developed by the research community to share text data and annotations.

We used exact phrase-based evaluation, i.e., a named entity is correct only when the predicted span and entity type exactly matches the reference annotation. This is important as partial matching (e.g., infarction) may be semantically different from the exact matching (e.g., myocardial infarction). For relations, a predicted relation between two entity types is regarded as correct only if both the relation type is correct and the prediction of all relevant named entities is correct. We used F_1 score for our system evaluation because it combines precision and recall, both of which are important metrics in the evaluation of information extraction systems. A micro-averaged F_1 score was used to get an aggregate F_1 score over all classes. We strictly followed the micro-average implementation used by scikit-learn[4]. We also report both micro-averaged precision and recall scores for the systems in the interests of interpretability.

Our evaluation script also provides an approximate metric that uses word-based evaluation, i.e., a named entity is correct if one or more words match. However, approximate match was not used for evaluation in the MADE challenge. During the course of the challenge, we found instances of inconsistent inclusion or exclusion of the period for a named entity. Therefore, our evaluation script ignores span errors of one trailing character length to account for such inconsistencies. Details regarding the annotation inconsistencies can be found in Sect. 5.

3.6 Test and Train Data

In total, 213 notes from our MADE corpus were selected for the testing split, and the remaining 876 notes formed the training split of the MADE challenge. To minimize the potential of over-fitting and maximize the evaluation quality, we used two approaches to select the test set. We first selected three patients (of 21) from the MADE cohort and included all their EHR notes (a total of 153) in the test set. We then selected 0–4 notes from the remaining 18 patients to add an additional 60 notes for the test set.

4 The MADE Challenge

The MADE challenge invited participants to submit systems for three shared tasks. The MADE corpus training data were released in November 2017, 4 months before the final test run. Submissions were evaluated using the criterion described in Sect. 3.5. We designed two different runs: standard and extended. Submissions in the standard runs were limited in the type of external tools they could use; as such, it provided a fair evaluation of the NLP models. For this run, teams had no restriction in using open-NLP systems. They could use outputs of open NLP tools such as Stanford NLP, Natural Language Toolkit (NLTK), and the UMLS tools for feature engineering. However, they were not allowed to use custom clinical NLP software, other EHR datasets, or NLP tools trained on other EHR datasets. They were also not allowed to use proprietary or in-house NLP software. Systems not adhering to these constraints were categorized as extended runs. We allowed two submissions for each run from each participating team. However, we only considered standard runs in the final evaluation. Please refer to the relevant articles for an evaluation and analysis of their extended runs.

The three shared tasks were designed to evaluate submissions with an overall goal of identifying ADEs and other relevant entities and relations from EHR notes.

Task 1: Named Entity Recognition (NER)

This task required extraction of both EHR named entities spans and their types from EHR notes. The named entity types are described in Sect. 3.1. Input was an unlabeled raw EHR note. Output was a bioC file containing the entity span and type. The evaluation for this task used F_1

Table 4 Performance metrics for the best runs by teams for the named entity recognition task (shared task 1)

Ranking	Team names	Recall	Precision	F_1 score
1	WPI-Wunnava [58]	0.8247	0.8333	0.8290
2	IBMResearch-dandala [59]	0.8243	0.8327	0.8285
3	UFL-gators [60]	0.8148	0.8318	0.8232
4	UArizonaIschool-Xu [61]	0.8042	0.8272	0.8156
5	UofUtah-Patterson [62]	0.7667	0.8280	0.7962
6	AEHRC-HoaNGO [63]	0.7463	0.8349	0.7881
7	ASU-BMI [64]	0.7074	0.8358	0.7663
8	MITRE-Tresner-Kirsch	0.7581	0.7443	0.7511
9	SUNY-Tao	0.6658	0.8054	0.7290
10	UCA-I3S-SPARK [65]	0.7198	0.6814	0.7001

Table 5 Performance metrics for the best runs by teams for the relation identification task (shared task 2)

Ranking	Team names	Recall	Precision	F_1 score
1	UofUtah-Patterson [62]	0.8806	0.8565	0.8684
2	IBMResearch-dandala [59]	0.8736	0.8093	0.8402
3	UArizonaIschool-Xu [61]	0.8846	0.7849	0.8318
4	ASU-BMI [64]	0.7693	0.8680	0.8157
5	KazanFederalUniversity-Alimova	0.6225	0.3343	0.4350

[4] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

Table 6 Performance metrics for the best runs by teams for the joint relation identification task (shared task 3)

Ranking	Team names	Recall	Precision	F_1 score
1	IBMResearch-dandala [59]	0.6317	0.6029	0.6170
2	UArizonaIschool-Xu [61]	0.6005	0.5965	0.5985
3	UofUtah-Patterson [62]	0.5176	0.6918	0.5921
4	ASU-BMI [64]	0.4350	0.6431	0.5189

score evaluation metrics based on exact phrase matches, described in Sect. 3.5.

Task 2: Relation Identification (RI)

This task required classification of relation and its type between two provided named entities. Since the named entities were provided as input, this task did not require detection of named entity spans or types. The relation types are described in Sect. 3.2. The input was the unlabeled EHR notes and a bioC file containing the list of present named entities. The output was a bioC file containing the relationships between the provided list of named entities, if any. The evaluation for this task used F_1 score evaluation metrics, described in Sect. 3.5.

Task 3: Joint Relation Extraction (NER-RI)

This task required prediction of both the named entities and their relations. Therefore, the systems submitted to this task had to jointly conduct both NER and RI. Input was an unlabeled EHR document. Submissions were expected to correctly extract the named entities, predict the entity type, and predict their relations. Output was a bioC file containing the named entity and relations. The evaluation for this task used F_1 score evaluation metrics based on exact phrase matches, as described in Sect. 3.5.

4.1 Submissions

The workshop submissions included 41 runs from 11 teams. The largest participation was in the first (NER) task and the smallest in the third (NER-RI). We show the exact F_1 score, precision and recall for all teams in Table 4 (NER task),

Table 7 Architecture details shared by teams within the workshop proceedings

Team names	LSTM	CRF	PWE	CE	Features[a]	Relation classifier[b]
UCA-I3S-SPARK [65]	+	−	+	+	POS	−
UFL-gators [60]	+	−	+	−	−	−
UofUtah-Patterson [62]	−	+	+	−	POS, surface	Random Forest
ASU-BMI [64]	+	+	+	+	Surface	Random Forest
IBMResearch-dandala [59]	+	+	+	+	POS	Attention Bi-LSTM
WPI-Wunnava [58]	+	+	+	+	−	−
UArizonaIschool-Xu [61]	+	+	+	+	Prefix, SUffiX EMBEDDING	SVM
AEHRC-HoaNGO [63]	−	+	+	−	Snomed-CT, POS, Dependency	−

CE character embeddings, *CRF* conditional random field, *CT* clinical terms, *LSTM* long short-term memory, *POS* part-of-speech, *PWE* pre-trained word embeddings, *SVM* support vector machine

[a]Documents additional features used by the submission for their named entity recognition system

[b]Classification methodology used by the submission for relation classification

Table 8 Label-wise recall, precision, and F_1 score values for the top submission and ensemble in task 1

NE type	Best submission			Top 3 ensemble		
	Recall	Precision	F_1 score	Recall	Precision	F_1 score
ADE	0.5266	0.7229	0.6093	0.5058	0.7956	0.6184
Indication	0.5959	0.6467	0.6202	0.5833	0.7860	0.6696
Other SSD	0.8490	0.8294	0.8391	0.8547	0.8483	0.8515
Drugname	0.8743	0.9057	0.8897	0.8922	0.9319	0.9116
Duration	0.7443	0.6428	0.6898	0.6766	0.9000	0.7725
Frequency	0.7541	0.7611	0.7576	0.8376	0.9123	0.8734
Dosage	0.8888	0.8757	0.8822	0.8926	0.8915	0.8920
Severity	0.8014	0.8492	0.8246	0.8014	0.8577	0.8286
Route	0.9383	0.9125	0.9252	0.9203	0.9396	0.9298

The ensemble takes the majority vote from the top three systems

ADE adverse drug event, *NE* named entity, *SSD* sign, symptom, or disease

Table 9 Label-wise recall, precision, and F_1 score values for the top submission and ensemble in task 2

Relation type	Best submission			Top 3 ensemble		
	Recall	Precision	F_1 score	Recall	Precision	F_1 score
ADE–drugname	0.7377	0.7032	0.7200	0.7207	0.7047	0.7126
SSD–severity	0.9677	0.9359	0.9516	0.9534	0.9334	0.9433
Drugname–route	0.9120	0.9346	0.9232	0.9274	0.9483	0.9377
Drugname–dosage	0.9676	0.9544	0.9610	0.9665	0.9479	0.9571
Drugname–duration	0.9047	0.7732	0.8338	0.9523	0.6278	0.7567
Drugname–route	0.9260	0.9428	0.9343	0.9369	0.9647	0.9506
Indication–drugname	0.7671	0.7187	0.7421	0.8242	0.7680	0.7951

The ensemble takes the majority vote from the top three systems

ADE adverse drug event, *SSD* sign, symptom, or disease

Table 10 Label-wise recall, precision, and F_1 score values for the top submission and ensemble in task 3

Relation type	Best submission			Top 3 ensemble		
	Recall	Precision	F_1 score	Recall	Precision	F_1 score
ADE–drugname	0.4264	0.4280	0.4272	0.3264	0.6784	0.4407
SSD–severity	0.5080	0.4694	0.4879	0.4740	0.5875	0.5247
Drugname–route	0.7956	0.7801	0.7878	0.7868	0.8689	0.8258
Drugname–dosage	0.7840	0.7324	0.7573	0.7655	0.8371	0.7997
Drugname–duration	0.4557	0.4685	0.4620	0.4965	0.5572	0.5251
Drugname–route	0.7424	0.7970	0.7687	0.7260	0.9059	0.8060
Indication–drugname	0.5365	0.4630	0.4970	0.4223	0.6525	0.5128

The ensemble takes the majority vote from the top three systems

Table 5 (RI task), and Table 6 (NRE-RI task). Table 7 provides a tabular view of the features and methods used. The NER task (task 1) had a best F_1 score of 0.829. RI and NER-RI tasks had best F_1 scores of 0.8684 and 0.6170, respectively. We also evaluated the extended runs, although they were not used for MADE task rankings. No extended run performed better than the top system in any of the three tasks. The label-wise recall, precision, and F_1 scores of the best submission system in all three tasks are provided in Tables 8, 9, and 10. These tables also provide the score for an ensemble prediction system composed of the top three runs in each task. More details about the ensemble system are provided in Sect. 4.3.

A brief overview of the methods used is provided in the following section. Please refer to the relevant team papers for further details about their methodologies.

4.2 Methods

As shown in Table 7, although the teams developed a variety of different sequence labeling and machine learning models, long short-term memory (LSTM) and CRF were the most widely used models for NER. The RI task showed a variety of models, such as support vector machines (SVMs), random forest, etc. A brief overview of the methods used in NER and relation classification are provided in the following.

NER

The task of NER can be posed as a sequence labeling problem, where a sentence can be treated as a sequence of tokens. The task is then reduced to the problem of labeling each token with the named entity tag or an "outside" (no named entity) tag. Commonly used algorithms for sequence labeling are Markov models (hidden Markov models, CRF), neural network models [convolutional neural network, recurrent neural network (RNN)], or a combination of both.

Linear-chain CRF [66] and related models (maximum entropy Markov model [MEMM] [67], hidden Markov model [HMM] [68]) belong to a class of methods in machine learning based on Markov models. CRF, which is a widely used method in sequence labeling, maximizes the joint probability of the label sequence conditioned on the input sentence.

RNN such as LSTM [69] or gated recurrent units (GRU) [70] are neural networks with recurrent connections that are designed to process sequential data. They have been shown to be useful in several NLP tasks such as NER [71], language modeling [72], and part of speech [73]. CRFs and RNNs were the two main methods used in the MADE challenge for the NER task. These methods take bag-of-word and other relevant features as inputs and produce the labels as outputs. Several teams experimented with character embeddings and other sub-word representations such

Table 11 Performance metrics calculated for ensemble of submissions as described in Sect. 4.3

Shared task	Top three teams			All teams		
	Recall	Precision	F_1 score	Recall	Precision	F_1 score
Task 1: NER	0.8334	0.87297	0.8527	0.7847	0.9094	0.8425
Task 2: RI	0.8935	0.8625	0.8777	0.8782	0.7845	0.8287
Task 3: NER-RI	0.5841	0.7616	0.6612	0.4633	0.8580	0.6017

The "top three teams" shows the ensemble generated from submissions of top three teams. The "all teams" ensemble uses submissions from all teams. For each team in both ensembles, only the best run is used. So, all ensembles in "top three teams" are ensembles of three different systems

NER named entity recognition, *RI* relation identification

as suffix and prefix embeddings. Features such as part of speech and surface features (case-based features) were also used with both RNN and CRF-based models. All teams use pre-trained word embedding to either pre-initialize their neural models or as features for CRF training.

Relation Classification

RI and NER-RI tasks require classification of the named entity pairs into several relation classes. The absence of a relation between the named entity pairs can be treated as another class in a multi-class classification scheme. Some submissions divide the classification into two separate sequential classification steps. The first classification task is to predict whether a relationship exists between the two named entity pairs. The second classification task predicts the type of that relation.

The classification methods range from neural network-based methods such as a bidirectional LSTM with attention layer [59] to random forests [64, 65] and SVMs [61]. Neural network-based methods use a final soft-max layer and cross entropy loss for training the relation classifier. SVM [74] is a statistical machine learning technique that uses maximum margin loss to train the classifier. Random forests [75] are a class of ensemble methods. They use the combined score from a collection of decision trees to produce the class prediction.

4.3 Analysis

The micro-average F_1 score for task 3 was significantly lower than that for tasks 1 and 2. This was expected, since prediction in task 3 compounded the errors in both the NER and the RI steps. For the real-world application of extracting drugname and related ADEs, the F_1 score needs to be further improved from its current best of 0.4272 in the NER-RI task (Table 10). A major factor behind the low score of the ADE-drugname relation type is the low NER F_1 score of ADE. However, the prediction of this relation itself is also a challenging prediction problem as evidenced by the F_1 score of 0.72 in task 2. This may be because the text span between two entities in this relation could be large (Table 3). Similar arguments can also be made about the relation indication-drugname, which is another important relation type for downstream applications such as drug-efficacy studies.

We ran paired sample t tests to evaluate the statistical significance of the differences between the top three models in each task. The paired t test evaluates the difference between two related variables. The differences were considered significant if the p value was < 0.05. The samples used in our test were obtained by using file-level micro-average F_1 scores. For the first task, we found no statistically significant difference between the first [58] and second [59] systems. However, the third system [60] was statistically significantly different from that in both Wunnava et al. [58] and Dandala et al. [59]. For task 2, all differences between the top three teams were statistically significant. For task 3, the third system [62] was statistically significantly different from the first [59]. All other differences among the top three teams in this task were statistically insignificant.

We built an ensemble system using the submitted runs. The ensemble output for each task is generated using a simple majority vote scheme. A prediction (named entity or a relation instance) is used by the ensemble if a majority of submissions agree on it. For shared task 1, the entire named entity phrase along with its type is taken as one prediction instance. For tasks 2 and 3, the complete relation prediction (relation type along with its constituent named entity predictions) constitutes one instance. The ensemble F_1 score is shown in Table 11. Each ensemble was created by choosing one best standard run from each team. We also used an ensemble composed of one standard run each from the top three teams for each task. The ensembles show significant performance gains in task 2 and 3 when compared with the best individual system in each shared task category. Even in task 1, the performance increase in F_1 score is around 0.02. This indicates that the top systems in the MADE challenge do not all learn the same pattern from the dataset. Instead, there is variability in the information they learn.

The NER and NER-RI tasks are interesting, not only from a research perspective but also because they have applications as steps in practical information extraction pipelines.

It is non-trivial to accurately estimate an F_1 score threshold for a good real-world performance. Therefore, we cannot calibrate an F_1 score of 0.8 in the context of its real-world performance. However, in our experience, a precision score of 0.83 suggests that the system can extract reasonably accurate and useful data from unstructured text. Therefore, we believe that these models should be good enough for large-scale statistical studies where count-based thresholds can be used to reduce the noise in the extracted data. However, applications that require patient-specific information may need NER systems with higher recall and precision. For instance, systems that use statistical methods to predict the outcome on a patient level may be very sensitive to the noise introduced by MADE NER systems.

As mentioned previously, the NER-RI performance was markedly lower than the NER performance of the submitted systems. The precision of NER-RI systems was significantly improved by building an ensemble, as shown in Table 11. However, a relatively low F_1 score of around 0.6 suggests that the current NER-RI systems need to be further improved to be useful in real-world applications. Future steps in improving these models can focus on (1) improving the machine learning models, (2) annotation efforts to build larger labeled corpora, or (3) designing machine learning techniques that use external knowledge and unlabeled text.

5 Corpus Errors

The annotations in the corpus contain a few inconsistencies and errors. Some of these errors were observed while testing the evaluation script on the data, and several were reported by the participating teams. The errors fall into two categories: inconsistency in annotations and overlapping annotations.

The inconsistency in annotations is due to inconsistent annotations of the period character in named entity annotations. As an example, for the phrase "q.i.d", annotations may sometimes miss the last period character, "q.i.d". This inconsistency is only exhibited with the period character and only when it is a trailing period. To account for this inconsistency, the evaluation script ignores trailing mistakes of one-character length for all tasks.

Errors due to overlapping annotations occur when spans of two named entities overlap. The two overlapping entities can be of the same or different types. A common error in this category is overlapping annotations in the same type. For example, both the phrase "vitamin D" and the overlapping sub-phrase "vitamin" are annotated as separate drugname annotations. Another common error in this category is double annotation on the same named entity. For example, the same phrase span "nausea" is annotated twice, as an ADE and as other SSD. Since it is not trivial to disambiguate the correct annotation for these errors, the evaluation script addresses this issue by treating all reference annotations as correct. This essentially means that the evaluation script slightly underestimates the true scores. However, since these errors are exhibited in only around 130 named entity annotations (from over 70,000 NE annotations in MADE), the evaluation script score still accurately accesses the performance of submitted systems.

6 Conclusion

We created an expert-curated corpus comprising longitudinal EHR notes from patients with cancer. The MADE cohort was annotated with medication- and ADE-related information. We released this cohort to the research community and used it as the benchmark to evaluate state-of-the-art NLP models. MADE results show that recent progress in NLP has led to remarkable improvements in NER and RI tasks for the clinical domain. However, as demonstrated by the joint NER-RI task, room for improvement remains. We invite future research efforts to improve the state of the art on these benchmarks.

Acknowledgements The authors are extremely thankful to the MADE 1.0 annotation team: Elaine Freund, Heather Keating, Nadya Frid, Edgard Granillo, Raelene Goodwin, Brian Corner, Zuofeng Li, Rashmi Prasad, Balaji Ramesh, Victoria Wang, and Steven Belknap for their contributions to the MADE project. They were an essential part of the data curation, annotation, and research process for MADE 1.0. They are also the authors of the annotation guideline used throughout the development of this corpus.

Compliance with Ethical Standards

Funding Research reported in this publication was supported by the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under award number R01HL125089.

Declaration The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu have no conflicts of interest that are directly relevant to the content of this article.

Dataset The data used are from the MADE 1.0 corpus available at http://bio-nlp.org/index.php/projects/39-nlp-challenges.

References

1. Donaldson MS, Corrigan JM, Kohn LT, et al. To err is human: building a safer health system, vol. 6. Washington: National Academies Press; 2000.

2. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, Laffel G, Sweitzer BJ, Shea BF, Hallisey R, et al. Incidence of adverse drug events and potential adverse drug events: implications for prevention. JAMA. 1995;274(1):29–34.
3. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA. 1998;279(15):1200–5.
4. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N, Petersen LA, Small SD, Sweitzer BJ, Leape LL. The costs of adverse drug events in hospitalized patients. JAMA. 1997;277(4):307–11.
5. Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. High rates of adverse drug events in a highly computerized hospital. Arch Intern Med. 2005;165(10):1111–6.
6. Gurwitz JH, Field TS, Harrold LR, Rothschild J, Debellis K, Seger AC, Cadoret C, Fish LS, Garber L, Kelleher M, et al. Incidence and preventability of adverse drug events among older persons in the ambulatory setting. JAMA. 2003;289(9):1107–16.
7. Johnson J, Booman L. Drug-related morbidity and mortality. J Manag Care Pharm. 1996;2(1):39–47.
8. Haas JS, Iyer A, Orav EJ, Schiff GD, Bates DW. Participation in an ambulatory e-pharmacovigilance system. Pharmacoepidemiol Drug Saf. 2010;19(9):961–9.
9. Frank C, Himmelstein DU, Woolhandler S, Bor DH, Wolfe SM, Heymann O, Zallman L, Lasser KE. Era of faster FDA drug approval has also seen increased black-box warnings and market withdrawals. Health Aff. 2014;33(8):1453–9.
10. WHO. WHO | Pharmacovigilance; 2017. http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/. Accessed 10 May 2018.
11. Edlavitch SA. Adverse drug event reporting: improving the low us reporting rates. Arch Intern Med. 1988;148(7):1499–503.
12. Hasford J, Goettler M, Munter K-H, Müller-Oerlinghausen B. Physicians' knowledge and attitudes regarding the spontaneous reporting system for adverse drug reactions. J Clin Epidemiol. 2002;55(9):945–50.
13. Begaud B, Moride Y, Tubert-Bitter P, Chaslerie A, Haramburu F. False-positives in spontaneous reporting: should we worry about them? Br J Clin Pharmacol. 1994;38(5):401–4.
14. Xu R, Wang Q. Comparing a knowledge-driven approach to a super-vised machine learning approach in large-scale extraction of drug-side effect relation-ships from free-text biomedical literature. BMC Bioinform. 2015;16:S6.
15. Butt TF, Cox AR, Oyebode JR, Ferner RE. Internet accounts of serious adverse drug reactions. Drug Saf. 2012;35(12):1159–70.
16. Rossi AC, Knapp DE, Anello C, O'Neill RT, Graham CF, Mendelis PS, Stanley GR. Discovery of adverse drug reactions: a comparison of selected phase IV studies with spontaneous reporting methods. JAMA. 1983;249(16):2226–8.
17. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, Jaulent M-C, Beyens M-N, Burgun A, Bousquet C. Adverse drug reaction identification and extraction in social media: a scoping review. J Med Internet Res. 2015;17(7):e171.
18. Smythe MA, Fanikos J, Gulseth MP, Wittkowsky AK, Spinler SA, Dager WE, Nutescu EA. Rivaroxaban: practical considerations for ensuring safety and efficacy. Pharmacotherapy. 2013;33(11):1223–45.
19. McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. Pharmacoepidemiol Drug Saf. 2012;21(S1):18–22.
20. Yih WK, Lieu TA, Kulldorff M, Martin D, McMahill-Walraven CN, Platt R, Selvam N, Selvan M, Lee GM, Nguyen M. Intussusception risk after rotavirus vaccination in us infants. N Engl J Med. 2014;370(6):503–51.
21. Peissig PL, Costa VS, Caldwell MD, Rottscheit C, Berg RL, Mendonca EA, Page D. Relational machine learning for electronic health record-driven phenotyping. J Biomed Informat. 2014;52:260–70.
22. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010;48:S106–13.
23. Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, Burdick E, Seger DL, Vliet MV, Bates DW. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc. 1998;5(3):305–14.
24. Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository. In: AMIA annual symposium proceedings, vol. 2011. American Medical Informatics Association.
25. Schulman S, Kearon C. Subcommittee on Control of Anticoagulation of the Scientific, Standardization Committee of the International Society on Thrombosis, and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J Thromb Haemost. 2005;3(4):692–4.
26. Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. J Biomed Inform. 2015;54:186–90.
27. Classen DC, Pestotnik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. BMJ Qual Saf. 2005;14(3):221–6.
28. Aronson AR. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA symposium, p. 17. American Medical Informatics Association; 2001.
29. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. Medex: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010;17(1):19–24.
30. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. In: ISMB (supplement of bioinformatics), p. 74–82; 2001.
31. Hahn U, Romacker M, Schulz S. Creating knowledge repositories from biomedical reports: the medsyndikate text mining system. In: Biocomputing 2002, pp. 338–349. World Scientific; 2001.
32. Hong Y, Lee M. Accessing bioscience images from abstract sentences. Bioinformatics. 2006;22(14):e547–56.
33. Yu H. Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. In: AMIA annual symposium proceedings, vol. 2006, p. 834. American Medical Informatics Association; 2006.
34. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the workshop on current trends in biomedical natural language processing: shared task, pp. 1–9. Association for Computational Linguistics; 2009.
35. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of biocreative: critical assessment of information extraction for biology; 2005.
36. Li Z, Cao Y, Antieau L, Agarwal S, Zhang Q, Yu H. Extracting medication information from patient discharge summaries. In: Proceedings of the third i2b2 workshop on challenges in natural language processing for clinical data; 2009.
37. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, Savova G. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc. 2014;22(1):143–54.
38. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, Ni Y, Kaiser M, Stoutenborough L, Solti I. Phenotyping for patient

safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. J Am Med Inform Assoc. 2014;21(5):776–84.
39. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005;12(4):448–57.
40. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The unified medical language system: an informatics research collaboration. J Am Med Inform Assoc. 1998;5(1):1–11.
41. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1):D267–70.
42. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTES) from narrative electronic health record data. J Am Med Inform Assoc. 2014;22(1):155–65.
43. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRS. In: AMIA annual symposium proceedings, vol. 2012, p. 1244. American Medical Informatics Association; 2012.
44. Wang S, Li Y, Ferguson D, Zhai C. Side effect PTM: an unsupervised topic model to mine adverse drug reactions from health forums. In: Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics, p. 321–330. ACM; 2014.
45. Nikfarjam Azadeh, Sarker Abeed, O'Connor Karen, Ginn Rachel, Gon-zalez Graciela. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc. 2015;22(3):671–81.
46. Li Q, Deleger L, Lingren T, Zhai H, Kaiser M, Stoutenborough L, Jegga AG, Cohen KB, Solti I. Mining FDA drug labels for medical conditions. BMC Med Inform Decis Making. 2013;13(1):53.
47. Duke JD, Friedlin J. ADESSA: a real-time decision support service for de-livery of semantically coded adverse drug event data. In: AMIA Annual symposium proceedings, vol. 2010, p. 177. American Medical Informatics Association; 2010.
48. Kim J-D, Ohta T, Tateisi Y, Tsujii J. Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(suppl 1):i180–2.
49. Cohen AM, Hersh WR. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. J Biomed Discov Collab. 2006;1(1):4.
50. Doğan RI, Lu Z. An improved corpus of disease mentions in Pubmed citations. In: Proceedings of the 2012 workshop on biomedical natural language processing, p. 91–99. Association for Computational Linguistics; 2012.
51. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinform. 2008;9(11):S9.
52. 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 Inform. 2012;45(5):885–92.
53. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17(5):514–8.
54. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform. 2015;57:333–49.
55. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378.
56. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
57. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, Wang J, Deng Q, Zhu S. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform. 2015;58:S47–52. https://doi.org/10.1016/j.jbi.2015.06.009.
58. Wunnava S, Qin X, Kakar T, Rundensteiner EA, Kong X. Bidirectional LSTM-CRF for adverse drug event tagging in electronic health records. In: Liu F, Jagannatha A, Yu H, editors. Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection, Proceedings of Machine Learning Research, vol. 90, p. 48–56. PMLR; 2018. http://proceedings.mlr.press/v90/wunnava18a.html. Accessed 10 May 2018.
59. 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. https://doi.org/10.1007/s40264-018-0764-x.
60. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Saf. 2019. https://doi.org/10.1007/s40264-018-0761-0.
61. Xu D, Yadav V, Bethard S. Uarizona at the made 1.0 NLP challenge. In: Liu F, Jagannatha A, Yu H, editors, Proceedings of the 1st international workshop on medication and adverse drug event detection, Proceedings of machine learning research, vol. 90, pp. 57–65. PMLR; 2018. http://proceedings.mlr.press/v90/xu18a.html. Accessed 10 May 2018.
62. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. Drug Saf. 2019. https://doi.org/10.1007/s40264-018-0763-y.
63. Ngo D-H, Metke-Jimenez A, Nguyen A. Knowledge-based feature engineering for detecting medication and adverse drug events from electronic health records. In: Liu F, Jagannatha A, Yu H, editors, Proceedings of the 1st international workshop on medication and adverse drug event detection, Proceedings of machine learning research, vol. 90, pp. 31–38. PMLR; 2018. http://proceedings.mlr.press/v90/ngo18a.html. Accessed 10 May 2018.
64. Magge A, Scotch M, Gonzalez-Hernandez G. Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers. In: Liu F, Jagannatha A, Yu H, editors. Proceedings of the 1st international workshop on medication and adverse drug event detection, Proceedings of machine learning research, vol. 90, p. 25–30. PMLR; 2018. http://proceedings.mlr.press/v90/magge18a.html. Accessed 10 May 2018.
65. Florez E, Precioso F, Riveill M, Pighetti R. Named entity recognition using neural networks for clinical notes. In: Liu F, Jagannatha A, Yu H, editors. Proceedings of the 1st international workshop on medication and adverse drug event detection, Proceedings of machine learning research, vol. 90, p. 7–15. PMLR; 2018. http://proceedings.mlr.press/v90/florez18a.html. Accessed 10 May 2018.
66. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning. San Francisco: Morgan Kaufmann Publishers Inc.; 2001. p. 282–9.
67. McCallum A, Freitag D, Pereira FCN. Maximum entropy markov models for information extraction and segmentation. In: ICML, vol. 17, pp. 591–598; 2000.
68. Zhou GD, Su J. Named entity recognition using an hmm-based chunk tagger. In: Proceedings of the 40th annual meeting on association for computational linguistics, p. 473–480. Association for Computational Linguistics; 2002.
69. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput. 2000;12(10):2451–71.
70. Chung J, Gulcehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. In: International conference on machine learning, p. 2067–2075; 2015.

71. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, vol. 2016, p. 473. NIH Public Access; 2016.
72. Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association; 2012.
73. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint, arXiv:1508.01991; 2015.
74. Cristianini N, Shawe-Taylor J, et al. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
75. Breiman Leo. Random forests. Mach Learn. 2001;45(1):5–32.