Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track

Kirk Roberts
School of Biomedical Informatics
The University of Texas Health Science Center
Houston, TX

Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
Bethesda, MD

Joseph M. Tonning
Center for Drug Evaluation and Research
U.S. Food and Drug Administration
Silver Spring, MD

Abstract

This paper describes the Adverse Reaction Extraction from Drug Labels Track, part of the 2017 Text Analysis Conference (TAC). Participants were provided with an annotated set of drug labels and challenged with: (1) extracting adverse reaction mentions and modifier terms such as negation, severity, and drug class; (2) identifying relations between adverse reaction mentions and those modifiers, including negation, hypothetical, and effect relations; (3) determining the unique set of positive adverse reaction mention strings across all sections of a drug label; and (4) normalizing those adverse reaction strings to a standard terminology, MedDRA. Ten teams submitted at least one valid run, with 20 submissions in total.

Background

The U.S. Food and Drug Administration (FDA) is responsible for protecting the public health by assuring the safety, efficacy, and security of all FDA-regulated products, including human and veterinary drugs, prescription and over-the-counter pharmaceutical drugs, vaccines, biopharmaceuticals, blood transfusions, biological products, medical devices, food safety, tobacco products, dietary supplements, cosmetics, and electromagnetic radiation emitting devices. Within the FDA, the Center for Drug Evaluation and Research (CDER) is responsible for regulating over-the-counter and prescription drugs, biological therapeutics, and generic drugs. CDER is interested in developing automatic tools for drug-adverse reaction signal detection. Current post-marketing safety signal generation in CDER relies on analysis of spontaneous adverse event reports submitted to the FDA Adverse Event Reporting System (FAERS). Some of these events are already known and reported in the Structured Product Labels (SPL) of drugs. To detect novel adverse reactions more efficiently, CDER needs to automate the current manual approach that requires reading the text of an adverse event report to determine if a given event is already noted in the SPL. To do this, adverse events need to be extracted from the unstructured SPLs into a structured list in the form of MedDRA (Medical Dictionary for Regulatory Activities Terminology) Preferred Terms (PT). This will allow the linking of adverse events reported in FAERS (which are already normalized to MedDRA PTs) to Structured Product Labels, thus allowing the FDA to automatically determine whether a reported event is either already known (i.e., because it is in the SPL) or a previously undetected adverse event.

This is a problem where natural language processing (NLP) systems can provide a great benefit to the FDA and medical community in general. The purpose of this TAC track, therefore, is to test various NLP approaches for their information extraction (IE) performance on adverse reactions in SPLs. While the

ultimate goal is for NLP systems to extract MedDRA PTs from the drug labels (the standard structured representation for adverse events), this track also evaluates and provides data for several intermediate tasks, such as extracting mentions, relations, and identifying unique adverse reactions prior to mapping them to MedDRA.

Related Work

The FDA has long been interested in applying data mining methods to further its pharmacovigilance goals (Szarfman et al., 2004; Almenoff et al., 2005; Duggirala et al., 2015). While clinical trials form the primary basis for determining the adverse events of a drug under FDA consideration, many issues with a drug only arise after FDA approval. This could be due to the small percentage of people impacted by an adverse reaction, its use in populations not studied in the trial, or a myriad of other potential reasons. As such, both the FDA and outside researchers collect and analyze a significant amount of data for the purpose of detecting adverse reactions as early as possible.

While the primary source for adverse events, the FDA Adverse Event Reporting System (FAERS), uses structured data, quite a few pharmacovigilance methods involve natural language (Harpaz et al., 2014). The particular source of data varies widely, each carrying its own set of challenges. These sources include

- the Vaccine Adverse Event Reporting System (VAERS) (Botsis et al., 2011)
- biomedical literature (Shetty and Dalal, 2011; Wang et al., 2011; van Mulligen et al., 2012; Xu and Wang, 2014)
- electronic health records (EHRs) (Wang et al., 2009; Gurulingappa et al., 2012; Haerian et al., 2012; Harpaz et al., 2013; LePendu et al., 2013)
- social media (Sarker et al., 2015) ranging from online health websites (Leaman et al., 2010; Chee et al., 2011; Liu et al., 2011; Nikfarjam and Gonzalez, 2011; Yang et al., 2012; Liu and Chen, 2013; Nikfarjam and Gonzalez, 2015), to Twitter (Bian et al., 2012; Jiang and Zheng, 2013; Nikfarjam and Gonzalez, 2015), and search logs (White et al., 2013).

In contrast, relatively little work has been performed on drug labels. Fung et al. (2013) focused on drug indications (why the drug was prescribed), as opposed to the subsequent reactions. Closer to the problem studied in this track, two resources have extracted adverse events from drug labels: SPLICER (Friedlin and Duke, 2010) and SIDER (Kuhn et al., 2010). This current TAC track is a refinement of these previous approaches for five reasons. First, the gold standard was created entirely through manual annotation, as opposed to rule- and dictionary-based approaches with minimal manual validation. Second, concepts of adverse reactions were clearly defined by FDA definitions of adverse reactions and annotators followed clear annotation guidelines when compiling (annotating, creating) the training and test datasets. Third, editors from MedDRA-MSSO reviewed and validated annotations for SPL terms not easily mapped to MedDRA PTs, ensuring an optimal mapping. Fourth, only the "Boxed Warning", "Warnings and Precautions", and "Adverse Reactions" sections of the label were annotated, thus avoiding potential confounding if other sections (such as the Contraindications section) were also included. Fifth, the goal of this track is to develop a community-wide, transparent evaluation of SPLs (with a fully reviewed annotated dataset of 200 labels) for further research purposes.

Data

The dataset consists of over two thousand drug labels: a training set of 101 labels, a test set of 99 labels, and an additional 2,109 unannotated labels. Participants were provided with the 101 annotated training labels and the unannotated version of the 99 test labels mixed in with the 2,109 unannotated labels. Since such a small percentage of these unannotated labels made up the gold standard test set, no special test data release was necessary: participants were given immediate access to the combined 2,208 unannotated labels along with the annotated training set.

The drug labels were provided in an XML format that was greatly simplified compared to the original DailyMed[1] XML format. Figure 1 shows part of an original formatted drug label. This formatting was

[1] https://dailymed.nlm.nih.gov/

6 ADVERSE REACTIONS

The following are discussed in more detail in other sections of the labeling:

- Hypertension, Hypokalemia, and Fluid Retention due to Mineralocorticoid Excess *[see Warnings and Precautions (5.1)]*.
- Adrenocortical Insufficiency *[see Warnings and Precautions (5.2)]*.
- Hepatotoxicity *[see Warnings and Precautions (5.3)]*.

6.1 Clinical Trial Experience

Because clinical trials are conducted under widely varying conditions, adverse reaction rates observed in the clinical trials of a drug cannot be directly compared to rates in the clinical trials of another drug and may not reflect the rates observed in clinical practice.

Two randomized placebo-controlled, multicenter clinical trials enrolled patients who had metastatic castration-resistant prostate cancer who were using a gonadotropin-releasing hormone (GnRH) agonist or were previously treated with orchiectomy. In both Study 1 and Study 2 ZYTIGA was administered at a dose of 1,000 mg daily in combination with prednisone 5 mg twice daily in the active treatment arms. Placebo plus prednisone 5 mg twice daily was given to control patients.

The most common adverse drug reactions (≥10%) reported in the two randomized clinical trials that occurred more commonly (>2%) in the abiraterone acetate arm were fatigue, joint swelling or discomfort, edema, hot flush, diarrhea, vomiting, cough, hypertension, dyspnea, urinary tract infection and contusion.

The most common laboratory abnormalities (>20%) reported in the two randomized clinical trials that occurred more commonly (≥2%) in the abiraterone acetate arm were anemia, elevated alkaline phosphatase, hypertriglyceridemia, lymphopenia, hypercholesterolemia, hyperglycemia, elevated AST, hypophosphatemia, elevated ALT and hypokalemia.

Study 1: Metastatic CRPC Following Chemotherapy

Study 1 enrolled 1195 patients with metastatic CRPC who had received prior docetaxel chemotherapy. Patients were not eligible if AST and/or ALT ≥2.5× ULN in the absence of liver metastases. Patients with liver metastases were excluded if AST and/or ALT >5× ULN.

Table 1 shows adverse reactions on the ZYTIGA arm in Study 1 that occurred with a ≥2% absolute increase in frequency compared to placebo or were events of special interest. The median duration of treatment with ZYTIGA was 8 months.

Table 1: Adverse Reactions due to ZYTIGA in Study 1

	ZYTIGA with Prednisone (N=791)		Placebo with Prednisone (N=394)	
System/Organ Class **Adverse reaction**	**All Grades*** **%**	**Grade 3–4 %**	**All Grades %**	**Grade 3–4 %**
Musculoskeletal and connective tissue disorders				
Joint swelling/discomfort†	29.5	4.2	23.4	4.1
Muscle discomfort‡	26.2	3.0	23.1	2.3
General disorders				
Edema§	26.7	1.9	18.3	0.8
Vascular disorders				
Hot flush	19.0	0.3	16.8	0.3
Hypertension	8.5	1.3	6.9	0.3
Gastrointestinal disorders				
Diarrhea	17.6	0.6	13.5	1.3
Dyspepsia	6.1	0	3.3	0
Infections and infestations				
Urinary tract infection	11.5	2.1	7.1	0.5
Upper respiratory tract infection	5.4	0	2.5	0
Respiratory, thoracic and mediastinal disorders				
Cough	10.6	0	7.6	0

Figure 1: Excerpt from the Adverse Reactions section of the Structured Product Label for drug Zytiga (abiraterone acetate tablet).

removed to generate the simplified XML. This format removal process is unfortunately lossy in nature: removing the complex structure of the original XML necessarily results in some undesirable issues in the simplified XML. It was felt, however, that simplifying the XML format to essentially flat text would lower the barrier to entry for participants and allow the focus to be on NLP methods as opposed to XML structure manipulation. The only formatting preserved from the original labels is the sections. SPLs contain many sections (such as the "Adverse Reactions" section seen in Figure 1). Only the sections relating to adverse reactions were kept: "Adverse Reactions", "Warnings and Precautions", and "Boxed Warning". Not all labels have all sections: most of the studied labels have an "Adverse Reactions" section, around half have a "Warnings and Precautions" section, and only a third have a "Boxed Warning" section.

The gold standard contains the following entity-style annotations:

- AdverseReaction: Defined by the FDA as an undesirable, untoward medical event that can reasonably be associated with the use of a drug in humans. This does not include all adverse events observed during the use of a drug, only those for which there is some basis to believe there is a causal relationship between the drug and the adverse event. Adverse reactions may include signs and symptoms, changes in laboratory parameters, and changes in other measures of critical body function, such as vital signs and ECG.
- Severity: Measurement of the severity of a specific AdverseReaction. This can be qualitative terms (e.g., "*major*", "*critical*", "*serious*", "*life-threatening*") or quantitative grades (e.g., "*grade 1*", "*Grade 3-4*", "*3 times upper limit of normal (ULN)*", "*240 mg/dL*").
- Factor: Any additional aspect of an AdverseReaction that is not covered by one of the other entities listed here. Notably, this includes hedging terms (e.g., *may*, *risk*, *potential*), references to the placebo arm of a clinical trial, or specific sub-populations (e.g., *pregnancy*, *fetus*).
- DrugClass: The class of drug that the specific drug for the label is part of. This is designed to capture drug class effects (e.g., "[*beta blockers*]$_{\text{DrugClass}}$ *may result in...*") that are not necessarily specific to the particular drug.
- Negation: Trigger word for event negation.
- Animal: Non-human animal species utilized during drug testing.

Note that Severity, Factor, DrugClass, Negation, and Animal are only annotated when utilized in one of the following relations with an AdverseReaction:

- Negated: Negation or Factor.
- Hypothetical: Animal, DrugClass, or Factor.
- Effect: Severity.

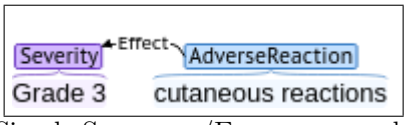
Examples of these annotations can be seen in Figures 2 and 3. See the Annotation Guidelines[2] for more details.

Next, the labels contain the unique, positive adverse reactions: both the de-cased strings of the positive AdverseReactions and the mappings to MedDRA. The mappings contain the MedDRA Lowest Level Terms (LLTs) and Preferred Terms (PTs).[3] This corresponds to the primary goal of the track: to identify the known AdverseReactions in a SPL in the form of MedDRA concepts, as represented by their PTs.

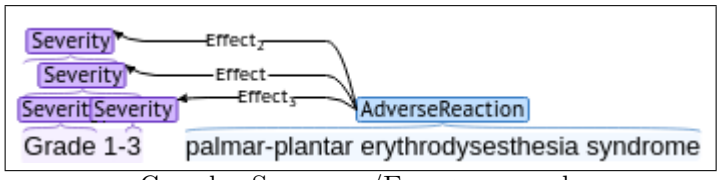
Some basic descriptive statistics about the gold standard corpus are shown in Table 1.

[2]https://bionlp.nlm.nih.gov/tac2017adversereactions/AnnotationGuidelines_TAC2017ADR.pdf

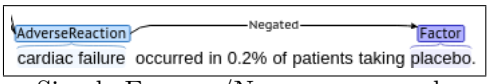
[3]Each LLT has only one associated PT, so there is a deterministic mapping between them. While the PT is the ultimate representation of the adverse event, the main advantage of providing the LLTs is that they are more directly identifiable from natural language. For example, the phrase "*increase in alt*" corresponds to the LLT *ALT increased* and the PT *Alanine aminotransferase increased*.



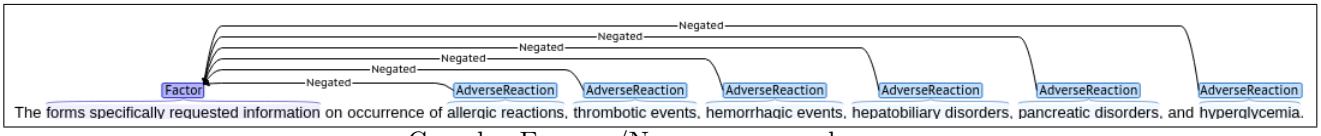
Simple SEVERITY/EFFECT example



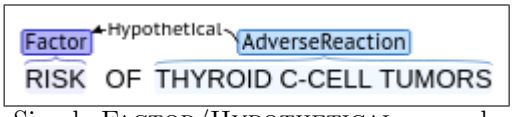
Complex SEVERITY/EFFECT example



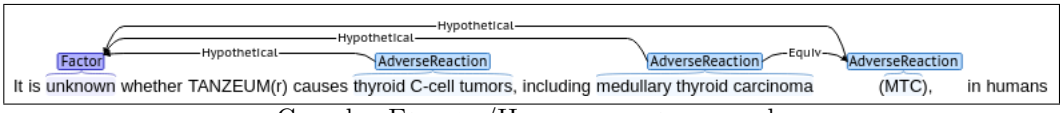
Simple FACTOR/NEGATED example



Complex FACTOR/NEGATED example

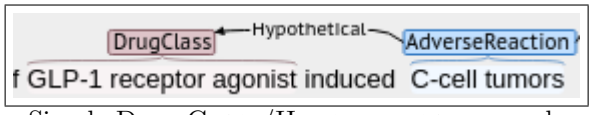


Simple FACTOR/HYPOTHETICAL example

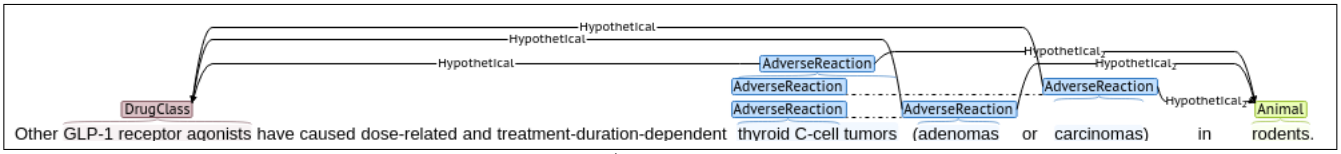


Complex FACTOR/HYPOTHETICAL example

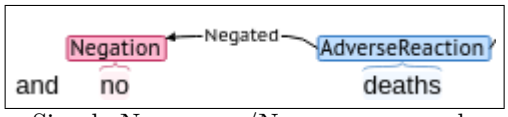
Figure 2: Examples of ADVERSEREACTIONs related to SEVERITY and FACTOR annotations through EFFECT, NEGATED, and HYPOTHETICAL relations.



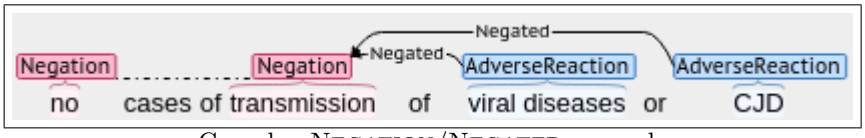
Simple DrugClass/Hypothetical example



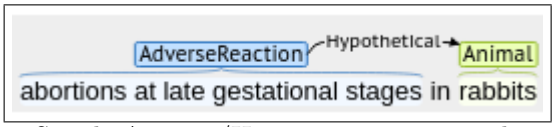
Complex DrugClass/Hypothetical example



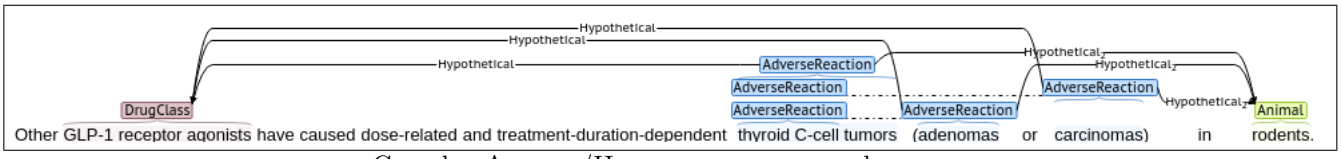
Simple Negation/Negated example



Complex Negation/Negated example



Simple Animal/Hypothetical example



Complex Animal/Hypothetical example

Figure 3: Examples of AdverseReactions related to DrugClass, Negation, and Animal annotations through Hypothetical and Negated relations.

Annotation	Training	Testing	Total
# SPLs	101	99	200
# Sections	239	237	476
# AdverseReaction	13,795	12,693	26,488
# Animal	44	86	130
# DrugClass	249	164	413
# Factor	602	562	1,164
# Negation	98	173	271
# Severity	934	947	1,881
# Effect	1,454	1,181	2,635
# Hypothetical	1,611	1,486	3,097
# Negated	163	288	451
# Reactions	7,038	6,343	13,381
# MedDRA Mappings	5,882	5,185	13,501

Table 1: Basic descriptive statistics of the ADE annotations.

Tasks

The track contained four specific tasks, each one building upon the previous tasks in a "layered" approach:

Task 1 Extract AdverseReactions and related entities (Severity, Factor, DrugClass, Negation, Animal). This is similar to many NLP named entity recognition (NER) tasks.

Task 2 Identify the relations between AdverseReactions and related entities (i.e., Negated, Hypothetical, and Effect. This is similar to many NLP relation identification tasks.

Task 3 Identify the *positive* AdverseReaction entities in the labels. For the purposes of this task, *positive* was defined as all the AdverseReactions that have not been negated (by a Negation or Factor) and are not related by a Hypothetical relation to a DrugClass or Animal. Note that this means Factors related via a Hypothetical relation are considered positive (e.g., "[*unknown risk*]$_{\text{Factor}}$ *of* [*stroke*]$_{\text{AdverseReaction}}$") for the purposes of this task. The result of this task was a list of unique strings corresponding to the positive AdverseReactions as they were written in the label.

Task 4 Normalize positive AdverseReaction entity (strings) to MedDRA PTs. The result of this task was a list of unique MedDRA preferred terms. Note that multiple unique strings may result in the same MedDRA PT (e.g., "*elevated alt*" and "*alt increases*" both normalize to the MedDRA term *ALT increased*), and in some cases a single string may result in multiple MedDRA PTs (e.g., "*infections of the mouth with candida albicans*" is two MedDRA PTs: *Oral candidiasis* and *Candida infection*). Furthermore, there are AdverseReactions in the corpus that human annotators were not able to map to MedDRA terms.

Tasks 1 and 2 correspond to traditional NLP information extraction (IE) tasks, while Tasks 3 and 4 involves more document-level aggregation (similar to phenotyping and problem list extraction). See Tables 2-4 for examples of what participants were expected to extract. Table 2 shows a portion of the label for Abiraterone. Table 3 shows the extracted AdverseReactions for Task 1, represented graphically (no other annotations are shown in this example for simplicity). Table 4 shows the AdverseReaction strings (left column) for Task 2 and the normalized MedDRA Preferred Terms (right column) for Task 3.

While the tasks were designed to build on each other, participation was optional on a per-task basis (e.g., a team could participate in Tasks 1 and 2, or just Task 4).

6 ADVERSE REACTIONS
The following are discussed in more detail in other sections of the labeling:
* Hypertension, Hypokalemia, and Fluid Retention due to Mineralocorticoid Excess [seeWarnings and Precautions (5.1)] .
* Adrenocortical Insufficiency [seeWarnings and Precautions (5.2)] .
* Hepatotoxicity [seeWarnings and Precautions (5.3)] .

EXCERPT: The most common adverse reactions (>= 10%) are fatigue, joint swelling or discomfort, edema, hot flush, diarrhea, vomiting, cough, hypertension, dyspnea, urinary tract infection and contusion.

The most common laboratory abnormalities (> 20%) are anemia, elevated alkaline phosphatase, hypertriglyceridemia, lymphopenia, hypercholesterolemia, hyperglycemia, elevated AST, hypophosphatemia, elevated ALT and hypokalemia.

Table 2: Portion of drug label for Abiraterone.

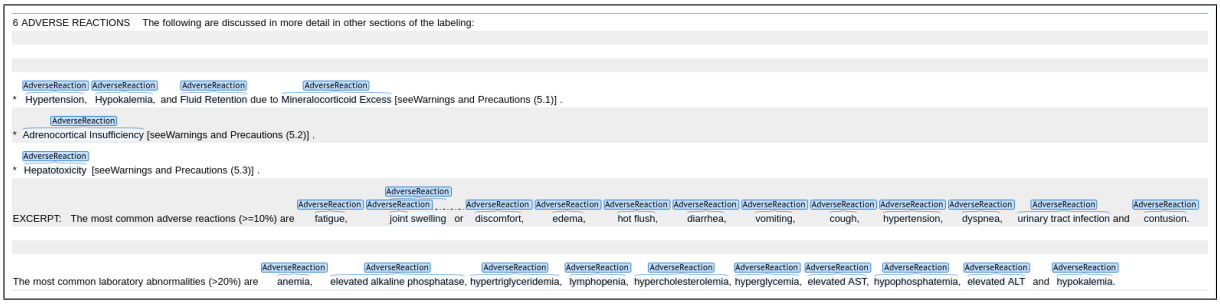


Table 3: Label annotated with AdverseReactions (no other annotations are present for simplicity).

AdverseReaction string	MedDRA PT
hypertension	Hypertension (10020772)
hypokalemia	Hypokalaemia (10021015)
fluid retention	Fluid retention (10016807)
mineralocorticoid excess	Hyperaldosteronism (10020571)
adrenocortical insufficiency	Adrenal insufficiency (10001367)
hepatotoxicity	Hepatotoxicity (10019851)
fatigue	Fatigue (10016256)
joint swelling	Joint swelling (10023232)
joint discomfort	Musculoskeletal discomfort (10053156)
edema	Oedema (10030095)
hot flush	Hot flush (10060800)
diarrhea	Diarrhoea (10012735)
vomiting	Vomiting (10047700)
cough	Cough (10011224)
dyspnea	Dyspnoea (10013968)
urinary tract infection	Urinary tract infection (10046571)
contusion	Contusion (10050584)
anemia	Anaemia (10002034)
elevated alkaline phosphatase	Blood alkaline phosphatase increased (10059570)
hypertriglyceridemia	Hypertriglyceridaemia (10020869)
lymphopenia	Lymphopenia (10025327)
hypercholesterolemia	Hypercholesterolaemia (10020603)
hyperglycemia	Hyperglycaemia (10020635)
elevated ast	Aspartate aminotransferase increased (10003481)
hypophosphatemia	Hypophosphataemia (10021058)
elevated alt	Alanine aminotransferase increased (10001551)

Table 4: Extracted AdverseReaction strings (left) and corresponding MedDRA PT normalizations (right).

Evaluation

Participants submitted system results on **all** unannotated labels (again, while the 99 labels composing the test set were part of the unannotated set, the participants were not aware of specifically which labels these were, so all 2,208 labels were processed by each system).

The evaluation measures were:

Task 1 Precision/Recall/F_1-measure on AdverseReaction, Severity, Factor, DrugClass, Negation, and Animal entities using IE-style measurement (i.e., offset-dependent). Both mentions with type and without type were evaluated. The primary evaluation metric was micro-averaged F_1 across the exact matched entity-level annotations (with type).

Task 2 Precision/Recall/F_1-measure on Negated, Hypothetical, and Effect relations. Both the full relation (all relations connected to an AdverseReaction mention) and binary relations were evaluated, both with and without type. The primary evaluation metric was micro-averaged F_1 across full relations (with type).

Task 3 SPL-level Precision/Recall/F_1-measure on unique *positive* AdverseReaction strings (i.e., unnormalized/un-mapped reactions). Both micro- and macro-averages across labels were evaluated. The primary evaluation metric was F_1 macro-averaged across labels (i.e., so labels with more AdverseReactions are not disproportionately weighted).

Task 4 SPL-level Precision/Recall/F_1-measure on unique MedDRA Preferred Terms. Both micro- and macro-averages across labels were evaluated. The primary evaluation metric was F_1 macro-averaged across labels.

Participants

The ten participants, along with brief descriptions of their approaches, are as follows:

1. **BUPT-PRIS** *Pattern Recognition and Intelligence System Lab*, *Beijing University of Posts and Telecommunications*. For Task 1, utilized a bi-directional LSTM-CRF (long short-term memory conditional random field) model combining word and character embeddings. The word embeddings were static, using a pre-trained `word2vec` model. The character embeddings were learned dynamically. For Task 2, utilized an adversarially-trained piece-wise CNN (convolutional neural network).
2. **CHOP** *The Children's Hospital of Philadelphia*. For Tasks 1 and 3, utilized a bi-directional LSTM using fastText embeddings trained on MEDLINE. A high-precision rule-based system was also used to identify reactions in tables. For Task 4, utilized word embeddings to map reaction strings to the closest MedDRA LLT.
3. **CONDL** *University of North Dakota*. For Task 1, utilized a bi-directional LSTM-CNNs-CRF. For Task 4, utilized a dictionary- and rule-based method where the dictionary was initially created using MedDRA, after which several expansion and exclusion rules were applied.
4. **GN_team** *University of Manchester*. For Task 1, utilized an ensemble of methods, including knowledge-driven rules, CRF, and bi-directional LSTM. Combined classifiers through both voting and stacked generalization.
5. **IBM_Research**. For Task 1, utilized a bi-directional LSTM-CRF to identify contiguous mentions and an attention-based bi-directional LSTM to identify when disjoint words belong to a mention. For Task 2, utilized attention-based bi-directional LSTM.
6. **MC_UC3M** *MeaningCloud*. For Task 1, utilized dictionaries built from MedDRA and SIDER with a support vector machine (SVM) classifier, along with a rule-based method to identify modifiers. For Task 2, utilized an SVM with lexical features. For Tasks 3 and 4, utilized rule-based methods building on the output of Tasks 1 and 2.
7. **Oracle**. For Task 3, utilized a boosted ensemble of CRF models to identify mentions (though did not submit results for Task 1); a rule-based approach to identify disjoint mentions; the ConText algorithm to filter out mentions by negation, animal, or drug class; and a post-processing algorithm to remove common types of error. The boosted ensemble used common NLP features along with features based on the label structure (header, table, etc.) and domain knowledge (MetaMap and UMLS).

System (Run)	Precison	Recall	F_1
UTH_CCB (3)	82.54	82.42	82.48
UTH_CCB (2)	80.22	84.40	82.26
UTH_CCB (1)	83.78	79.74	81.71
IBM_Research	80.90	75.30	78.00
CONDL (1)	76.45	77.49	76.97
GN_team (1)	80.19	72.23	76.00
GN_team (2)	76.84	74.36	75.58
PRNA_SUNY (1)	77.71	63.90	70.13
PRNA_SUNY (3)	77.71	63.90	70.13
CONDL (3)	65.19	69.77	67.41
CONDL (2)	65.47	61.40	63.37
PRNA_SUNY (2)	64.25	61.58	62.89
MC_UC3M (1)	54.79	66.33	60.01
MC_UC3M (2)	54.79	66.33	60.01
trddc_iiith	79.14	43.12	55.83
CHOP	57.95	29.64	39.22
BUPT_PRIS	40.47	11.81	18.29

Table 5: Task 1 Results

8. **PRNA_SUNY** *Philips Research North America / State University of New York at Albany.* For Task 1, utilized CRFs with morphological features, word embeddings, and dictionaries built from VigiAccess.org and UMLS. For Task 2, utilized logistic regression with semantic and syntactic features. For Task 3, utilized rules based on the output of Tasks 1 and 2. For Task 4, utilized MetaMap and the Sub-Term Mapping Tool (STMT) to normalize reactions.
9. **TRDDC_IIITH** *TCS Research / IIT Bombay / IIT Hyderabad.* For Tasks 1 and 2, utilized a joint neural network model, the All Word Pairs neural network (AWP-NN), with pre-trained GloVe word embeddings. An ensemble of AWP-NN models were used, each with a different initialization.
10. **UTH_CCB** *University of Texas Health Science Center at Houston.* For Tasks 1 and 2, utilized a bi-directional LSTM-CRF to identify AdverseReactions, followed by a second LSTM-CRF to identify related modifiers (modifiers not participating in a relation were filtered). Several rules were used to handle disjoint mentions. For Tasks 3 and 4, utilized a learning-to-rank approach. A BM25 model retrieved 10 MedDRA candidates, then RankSVM was used, including features based on the BM25 score, Jaccard similarity, and the translation-based ranking.

Submissions from three additional teams did not pass the validation checks, mostly due to not submitting all labels. Their submission results are not included here.

Results

The results for are shown in Tables 5 through 10. Task 2 was clearly the most challenging (max F_1 of 49.00 compared to F1s in the low 80s for the other tasks). This is likely due to the fact that the modifier terms (Animal, DrugClass, Factor, Negation, and Severity) were only annotated when they existed in a relation with an AdverseReaction, and they were far less common over all. The fairly high results on Task 4, on the other hand, are quite promising, with most systems achieving higher than 75 macro-F_1 and the best system achieving a 85 macro-F_1. Overall this likely demonstrates that the problem with Task 2 (rarity) also limits its downstream impact.

Discussion

We iterate that the knowledge gained from this track can provide great benefit to the FDA and the medical and research community in general. NLP systems which enable linkage of adverse reactions reported to the FDA with adverse reactions noted in labels (SPLs) will allow more rapid determination of whether a particular adverse reaction is already known (i.e., present in the SPL) or unknown.

The results of this track will stimulate development and evaluation of more advanced NLP tools for enhanced (greater precision and recall) extraction of adverse reactions from SPLs. Such enhanced extraction will in turn enable creation of a standardized, searchable dataset containing information about labeled (SPL) adverse reactions. This dataset will not only facilitate post-market surveillance for previously unobserved reactions but also other important activities such as determining whether a drug could

System (Run)	AdverseReaction			Animal			DrugClass			Factor			Negation			Severity		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
UTH_CCB (3)	85.1	85.3	85.2	89.4	68.6	77.6	50.3	49.4	49.8	64.9	71.4	68.0	67.6	57.8	62.3	66.4	61.5	63.8
UTH_CCB (2)	83.4	86.9	85.1	83.4	86.9	85.1	46.2	44.5	45.3	61.3	76.0	67.8	55.4	65.3	59.9	62.5	67.7	65.0
UTH_CCB (1)	86.4	82.4	84.4	89.2	67.4	76.8	50.9	49.4	50.2	65.6	70.6	68.0	69.5	56.6	62.4	67.1	60.4	63.6
IBM_Research	82.3	79.2	80.7	92.6	29.1	44.2	59.1	33.5	42.8	72.4	66.2	69.1	70.6	41.6	52.4	65.6	46.6	54.4
CONDL (1)	78.4	79.6	79.0	85.3	74.4	79.5	44.3	50.0	47.0	65.5	79.2	71.7	69.3	54.9	61.3	62.9	57.2	59.9
GN_team (1)	83.8	75.4	79.4	78.8	90.7	84.3	34.1	37.8	35.8	67.6	54.1	60.1	47.3	50.9	49.0	54.8	48.3	51.3
GN_team (2)	80.5	76.8	78.6	78.8	90.7	84.3	23.7	31.7	27.2	61.7	66.4	63.9	47.2	53.2	50.0	57.7	56.0	56.8
PRNA_SUNY (1)	78.6	67.6	72.7	79.7	68.6	73.8	65.3	28.7	39.8	70.2	44.0	54.0	70.8	19.7	30.8	66.6	40.2	50.2
PRNA_SUNY (3)	78.6	67.6	72.7	79.7	68.6	73.8	65.3	28.7	39.8	70.2	44.0	54.0	70.8	19.7	30.8	66.6	40.2	50.2
CONDL (3)	65.5	70.8	68.0	85.3	74.4	79.5	44.3	50.0	47.0	65.5	79.2	71.7	69.3	54.9	61.3	62.6	56.5	59.4
CONDL (2)	65.5	70.8	68.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0
PRNA_SUNY (2)	67.0	67.2	67.1	80.0	32.6	46.3	41.1	26.8	32.5	62.1	42.9	50.7	44.1	17.3	24.9	19.8	14.7	16.8
MC_UC3M (1)	63.7	70.8	67.1	76.6	57.0	65.3	19.2	39.6	25.9	04.0	07.7	05.3	10.6	53.8	17.7	37.1	49.5	42.4
MC_UC3M (2)	63.7	70.8	67.1	76.6	57.0	65.3	19.2	39.6	25.9	04.0	07.7	05.3	10.6	53.8	17.7	37.1	49.5	42.4
trddc_iiith	80.4	47.7	59.9	100.0	03.5	06.7	30.0	01.8	03.4	65.0	16.2	25.9	00.0	00.0	00.0	54.4	16.4	25.2
CHOP	57.9	34.2	43.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0	00.0
BUPT_PRIS	42.1	13.4	20.3	00.0	00.0	00.0	11.1	02.4	04.0	25.0	03.0	05.4	00.0	00.0	00.0	08.4	01.2	02.0

Table 6: Task 1 Per-Type Results

System (Run)	Precison	Recall	F_1
UTH_CCB (3)	50.24	47.82	49.00
UTH_CCB (1)	51.67	44.45	47.79
UTH_CCB (2)	46.24	48.32	47.26
IBM_Research	48.13	32.54	38.83
PRNA_SUNY (1)	50.48	22.36	30.99
PRNA_SUNY (3)	50.48	22.36	30.99
PRNA_SUNY (2)	31.28	9.34	14.39
MC_UC3M (2)	10.41	10.95	10.67
BUPT_PRIS	0.97	0.38	0.55

Table 7: Task 2 Results

be repurposed (i.e., for a new indication) or finding patterns to predict drug interactions or other toxicity by pharmacologic class or similar chemical moieties.

Conclusion

The goal of the TAC Adverse Reaction Extraction from Drug Labels Track was to evaluate and draw attention to the important problem of identifying the adverse reactions described in SPLs. Having an accurate list of known adverse reactions would be of tremendous value to the FDA in its many activities, including pharmacovigilance. Ten teams submitted a total of twenty submissions across the four tasks (17 submissions for Task 1, 9 for Task 2, 15 for Task 3, 12 for Task 4), with the top submission in the ultimate task (Task 4) achieving a macro-average F_1 of 85.33 at identifying MedDRA PTs in drug labels.

Acknowledgements The organizers would like to thank the corpus annotators: Alan Aronson, Laritza Rodriguez, and Sonya E. Shooshan. The support for this project was primarly through an Interagency Agreement (IAA 224-15-3022S) between the U.S. Food and Drug Administration (FDA) and the U.S. National Library of Medicine (NLM), part of the National Institutes of Health (NIH). This work was also partially supported by NLM/NIH under award number 4R00LM012104-02, as well as the intramural research program at the National Library of Medicine.

References

Almenoff, J., Tonning, J. M., Gould, A. L., Szarfman, A., Hauben, M., Ouellet-Hellstrom, R., Ball, R., Hornbuckle, K., Walsh, L., Yee, C., Sacks, S. T., Yuen, N., Patadia, V., Blum, M., Johnston, M., Gerrits, C., Seifert, H., and LaCroix, K. (2005). Perspectives on the Use of Data Mining in Pharmacovigilance. *Drug Safety*, 28(11):981–1007.

Bian, J., Topaloglu, U., and Yu, F. (2012). Towards Large-scale Twitter Mining for Drug-related Adverse Events. In *Proceedings of the 2010 International Workshop on Smart Health and Wellbeing*, pages 25–32.

System (Run)	Effect			Hypothetical			Negated		
	P	R	F_1	P	R	F_1	P	R	F_1
UTH_CCB (3)	56.8	53.4	55.0	52.5	58.1	55.2	50.4	40.3	44.8
UTH_CCB (1)	58.9	50.2	54.2	53.6	54.0	53.8	52.9	37.5	43.9
UTH_CCB (2)	52.2	57.3	54.6	50.2	58.3	54.0	42.0	46.2	44.0
IBM_Research	55.9	35.1	43.1	52.7	43.2	47.5	54.5	23.3	32.6
PRNA_SUNY (1)	60.5	26.4	36.7	53.8	26.5	35.5	57.8	09.0	15.6
PRNA_SUNY (3)	60.5	26.4	36.7	53.8	26.5	35.5	57.8	09.0	15.6
PRNA_SUNY (2)	11.6	01.7	03.0	43.4	17.9	25.3	25.7	06.6	10.5
MC_UC3M (2)	24.9	25.8	25.3	06.0	09.6	07.3	08.4	04.9	06.2
BUPT_PRIS	01.7	00.5	00.8	01.8	00.9	01.2	00.0	00.0	00.0

Table 8: Task 2 Per-Type Results

System (Run)	Micro-Precison	Micro-Recall	Micro-F_1	Macro-Precision	Macro-Recall	Macro-F_1
UTH_CCB (3)	80.97	84.87	82.87	80.69	85.05	82.19
UTH_CCB (1)	82.83	81.76	82.29	82.61	81.88	81.65
UTH_CCB (2)	79.68	85.57	82.52	78.77	85.62	81.39
Oracle (3)	81.18	79.69	80.43	81.47	79.28	79.67
Oracle (2)	82.71	78.05	80.31	82.64	77.73	79.42
Oracle (1)	81.28	79.32	80.28	81.10	78.81	79.20
CONDL (1)	87.77	67.33	76.21	87.34	67.64	75.15
PRNA_SUNY (1)	73.05	69.90	71.44	73.23	68.91	70.29
PRNA_SUNY (3)	73.05	69.90	71.44	73.23	68.91	70.29
MC_UC3M (1)	70.03	71.42	70.71	69.23	72.93	70.13
MC_UC3M (2)	70.03	71.42	70.71	69.23	72.93	70.13
CONDL (2)	70.86	69.76	70.31	70.16	70.29	69.35
CONDL (3)	70.86	69.76	70.31	70.16	70.29	69.35
PRNA_SUNY (2)	59.57	71.91	65.16	58.16	70.96	63.25
CHOP	64.29	39.57	48.99	62.97	39.95	47.99

Table 9: Task 3 Results

Botsis, T., Nguyen, . D., Woo, E. J., Markatou, M., and Ball, R. (2011). Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association*, 18(5):631–638.

Chee, B. W., Berlin, R., and Schatz, B. (2011). Predicting Adverse Drug Events from Personal Health Messages. In *Proceedings of the AMIA Annual Symposium*, pages 217–226.

Duggirala, H. J., Tonning, J. M., Smith, E., Bright, R. A., Baker, J. D., Ball, R., Bell, C., Bright-Ponte, S. J., Botsis, T., Bouri, K., Boyer, M., Burkhart, K., Condrey, G. S., Chen, J. J., Chirtel, S., Filice, R. W., Francis, H., Jiang, H., Levine, J., Martin, D., Oladipo, T., O'Neill, R., Palmer, L. A. M., Paredes, A., Rochester, G., Sholtes, D., Szarfman, A., Wong, H.-L., Xu, Z., and Kass-Hout, T. (2015). Use of data mining at the Food and Drug Administration. *Journal of the American Medical Informatics Association*, 23(2):428–434.

Friedlin, J. and Duke, J. (2010). Applying natural language processing to extract codify adverse drug reaction in medication labels. Technical report, Observational Medical Outcomes Partnership (OMOP).

Fung, K., Jao, C. S., and Demner-Fushman, D. (2013). Extracting drug indication information from structured product labels using natural language processing. *Journal of the American Medical Informatics Association*, 20(3):482–488.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., and Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H., and Friedman, C. (2012). Detection of Pharmacovigilance-Related adverse Events Using Electronic Health Records and automated Methods. *Clinical Pharmacology and Therapeutics*, 92(2):228–234.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., and Shah, N. H. (2014). Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. *Drug Safety*, 37(10):777–790.

Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Haerian, K., Shah, N. H., Chase, H. S., and Friedman, C. (2013). Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, 20(3):413–419.

Jiang, K. and Zheng, Y. (2013). Mining Twitter Data for Potential Drug Effects. In *Proceedings of the International Conference on Advanced Data Mining and Applications*, pages 434–443.

System (Run)	Micro-Precison	Micro-Recall	Micro-F_1	Macro-Precision	Macro-Recall	Macro-F_1
UTH_CCB (3)	84.17	89.84	86.91	83.02	89.06	85.33
UTH_CCB (1)	85.00	87.75	86.35	84.04	86.67	84.79
UTH_CCB (2)	82.42	90.78	86.40	80.83	89.90	84.53
CONDL (1)	88.81	77.16	82.58	88.20	75.76	80.50
PRNA_SUNY (1)	86.14	74.89	80.12	85.32	72.76	77.97
PRNA_SUNY (2)	81.55	78.24	79.86	79.80	76.03	77.25
PRNA_SUNY (3)	83.60	74.14	78.59	82.22	71.44	75.87
CONDL (2)	74.56	80.96	77.63	73.06	79.92	75.55
CONDL (3)	74.56	80.96	77.63	73.06	79.92	75.55
MC_UC3M (1)	73.40	80.25	76.67	72.10	80.38	75.29
MC_UC3M (2)	73.40	80.25	76.67	72.10	80.38	75.29
CHOP	71.78	50.14	59.04	70.12	49.84	57.27

Table 10: Task 4 Results

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6:343.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., and Gonzalez, G. (2010). Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 117–125.

LePendu, P., Iyer, S., Bauer-Mehren, A., Harpaz, R., Mortensen, J., Podchiyska, T., Ferris, T., and Shah, N. (2013). Pharmacovigilance Using Clinical Notes. *Clinical Pharmacology and Therapeutics*, 93(6):547–555.

Liu, J., Li, A., and Seneff, S. (2011). Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs. In *Proceedings of the First International Conference on Advances in Information Mining and Management*.

Liu, X. and Chen, H. (2013). AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums. In *Proceedings of the International Conference on Smart Health*, pages 134–150.

Nikfarjam, A. and Gonzalez, G. H. (2011). Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments. In *Proceedings of the AMIA Annual Symposium*, pages 1019–1026.

Nikfarjam, A. and Gonzalez, G. H. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Shetty, K. D. and Dalal, S. R. (2011). Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association*, 18(5):668–674.

Szarfman, A., Tonning, J., and Doraiswamy, P. M. (2004). Pharmacovigilance in the 21st Century: New Systematic Tools for an Old Problem. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 24:1099–1104.

van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., and Furlong, L. I. (2012). The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884.

Wang, W., Haerian, K., Salmasian, H., Harpaz, R., Chase, H., and Friedman, C. (2011). A Drug-Adverse Event Extraction Algorithm to Support Pharmacovigilance Knowledge Mining from PubMed Citations. In *Proceedings of the AMIA Annual Symposium*, pages 1464–1470.

Wang, X., Hripcsak, G., Markatou, M., and Friedman, C. (2009). Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. *Journal of the American Medical Informatics Association*, 16(3):328–337.

White, R. W., Tatonetti, N. P., Shah, N. H., Altman, R. B., and Horvitz, E. (2013). Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 20(3):404–408.

Xu, R. and Wang, Q. (2014). Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection. *BMC Bioinformatics*, 15:17.

Yang, C. C., Yang, H., Jiang, L., and Zhang, M. (2012). Social Media Mining for Drug Safety Signal Detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, pages 33–40.