Deciphering *Clinical Narratives* – key to automated Decision Making in Health Care

Lipika Dey

TCS Research

Outline of the talk

- Clinical Texts – brief introduction
- Clinical Text Processing challenges
- Clinical NLP pipeline
- Clinical Text Processing Use-cases
  - Handling Clinical Trials
  - Predicting ICU stay for patients
  - Detecting Adverse effect of drugs

Clinical narratives - Main form of communication within health care

- **Clinical Data**
  - Electronic Medical Records (EMR)/Electronic Health Records (EHR) (classic)
  - Physician and Care-given notes - Account of patient history and assessments - offering rich insights about clinical decision making
  - Clinical trials management – trial description for recruiting patients, monitor trial progress
- **Social Media (tweets, Facebook comments, message boards, etc.)**
  - personal accounts of patients – signals for mental health – adverse effects of drugs
  - Health care system feedback
- **Medical Literature**
  - News feeds, Medical journals
- **Insurance Providers (claims from private and government payers)**
  - Underwriter notes

Volume of text estimated – 2000+ exabytes at 2020

Analytics applications

- **Success / Failure of Clinical Trials**
- **Hospital Logistics Management**
- **Drug side effects**
- **Prediction of disease spread**
- **Understanding Diseases – cause, spread, pathway**
- **Genomics**

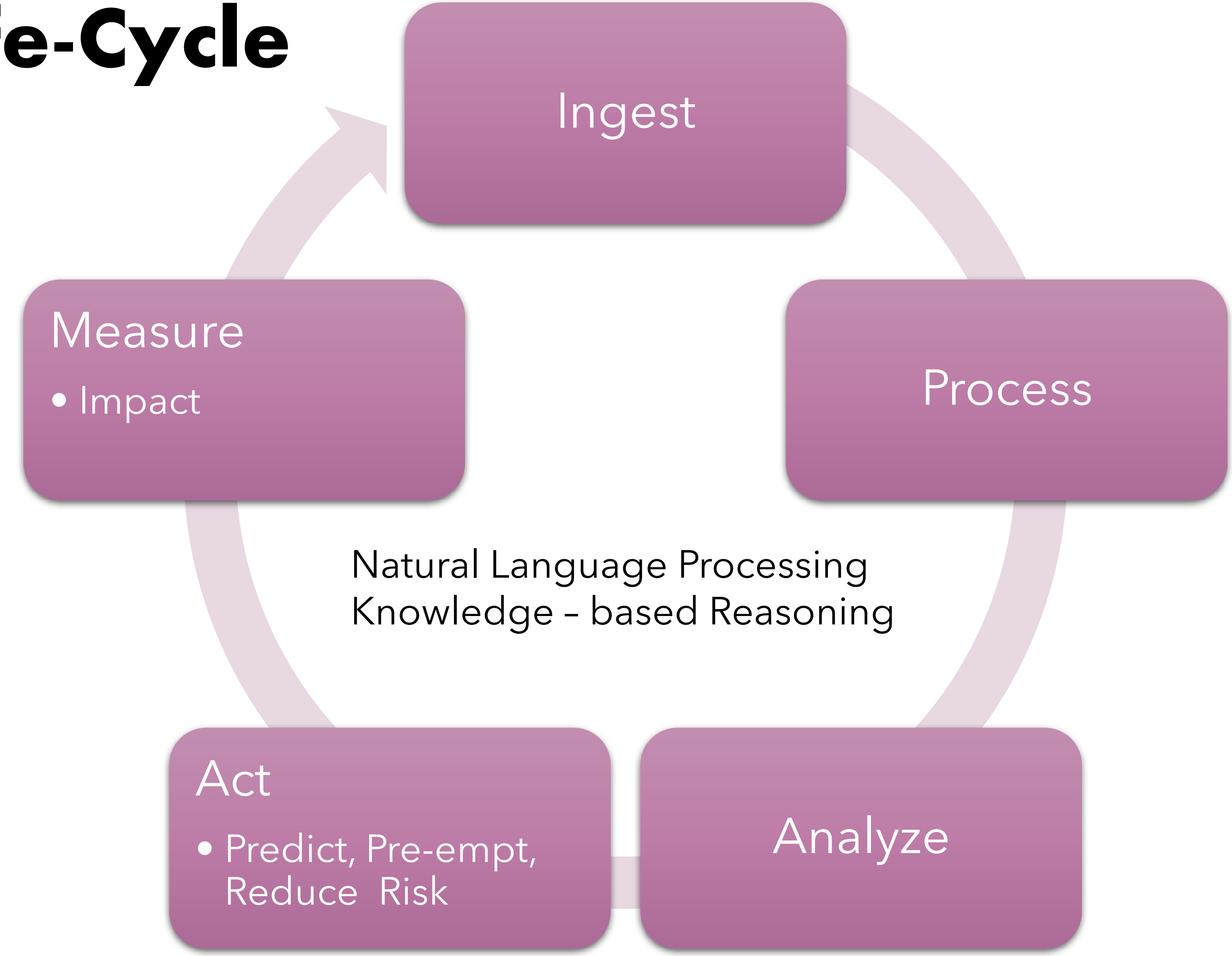
Clinical Trial description

Electronic Health Records

Social Media

Literature

Analytics Life-Cycle



Clinical NLP – ingest and process

- Natural language processing (NLP) plays an important role in unlocking insights embedded in clinical narratives
- Machine learning is the back-bone that aids development of NLP tools by leveraging large amounts of text data

- ✓ Information Extraction
- ✓ Text Classification
- ✓ Paraphrase detection
- ✓ Content Matching
- ✓ Predictive Modeling
- ✓ Summarization

Challenge - Data is totally unstructured

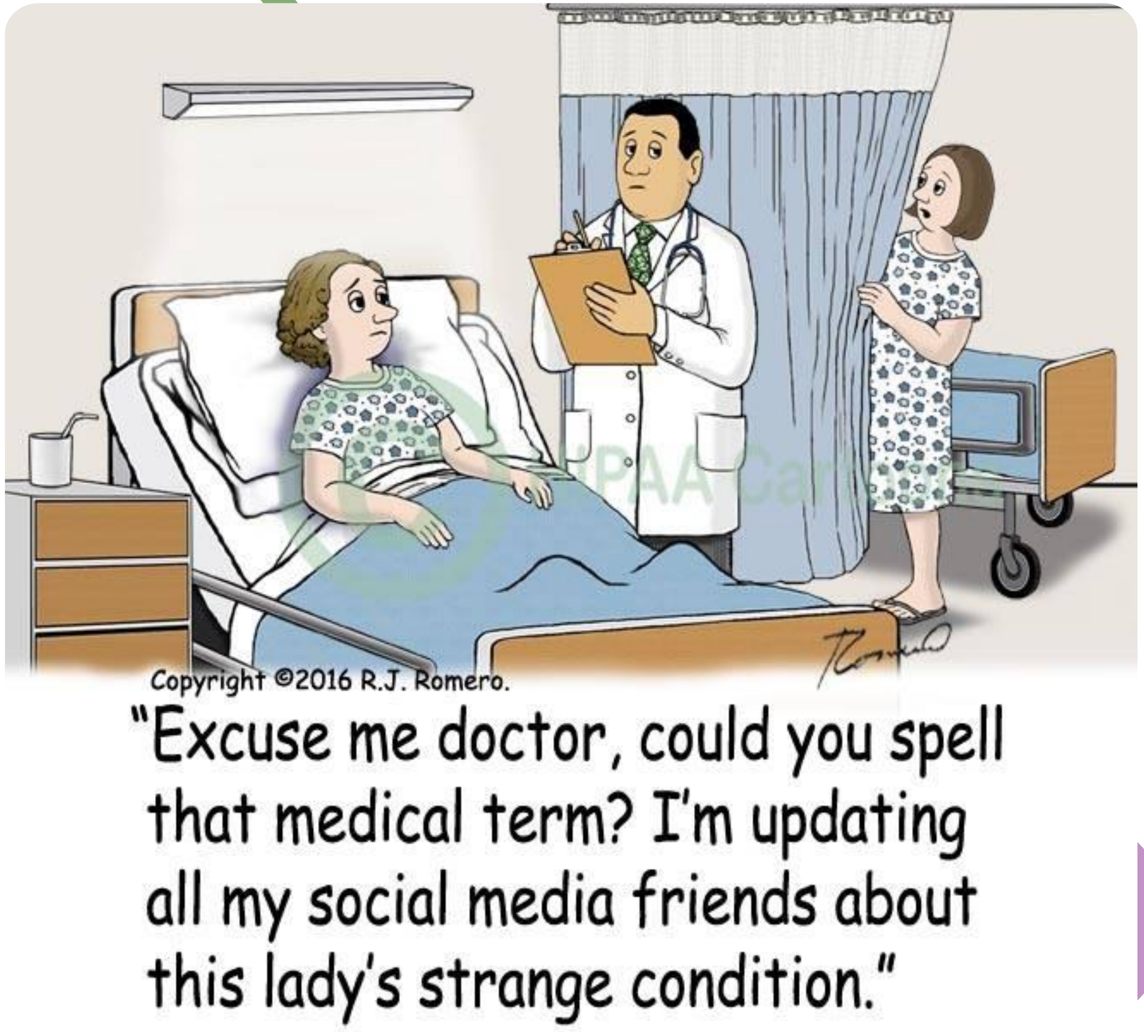
- Bulk of this data does not exist in discretely-labeled fields - but rather available as completely unstructured free text clinical notes
- Traditional healthcare analytics depends predominantly on discrete data fields

Challenge - Unavailability of annotated data

- Adverse-drug effect?
- Bankruptcy?
- Humor in Uniform?

Privacy and Security concerns

- It's more difficult to mask text data



Clinical text is very different

- Domain specific features abound
- Shared vocabularies are needed
  - When we say breathless – the doctor hears “dyspnea”
- Contextual interpretation
  - History
  - Ruling out
  - Grammar is not the mainstay
- Paraphrases
  - Obese with hyperinsulimenia
  - Type II Diabetes

“The Doctor will see you now. Here's your medical jargon dictionary.”

Clinical NLP - pillars

- Entities

Named Entity Detection and Classification

- Relations

Extraction Contextualization

- Concept mapping

Paraphrase detection

- Knowledge Graph

Reasoning

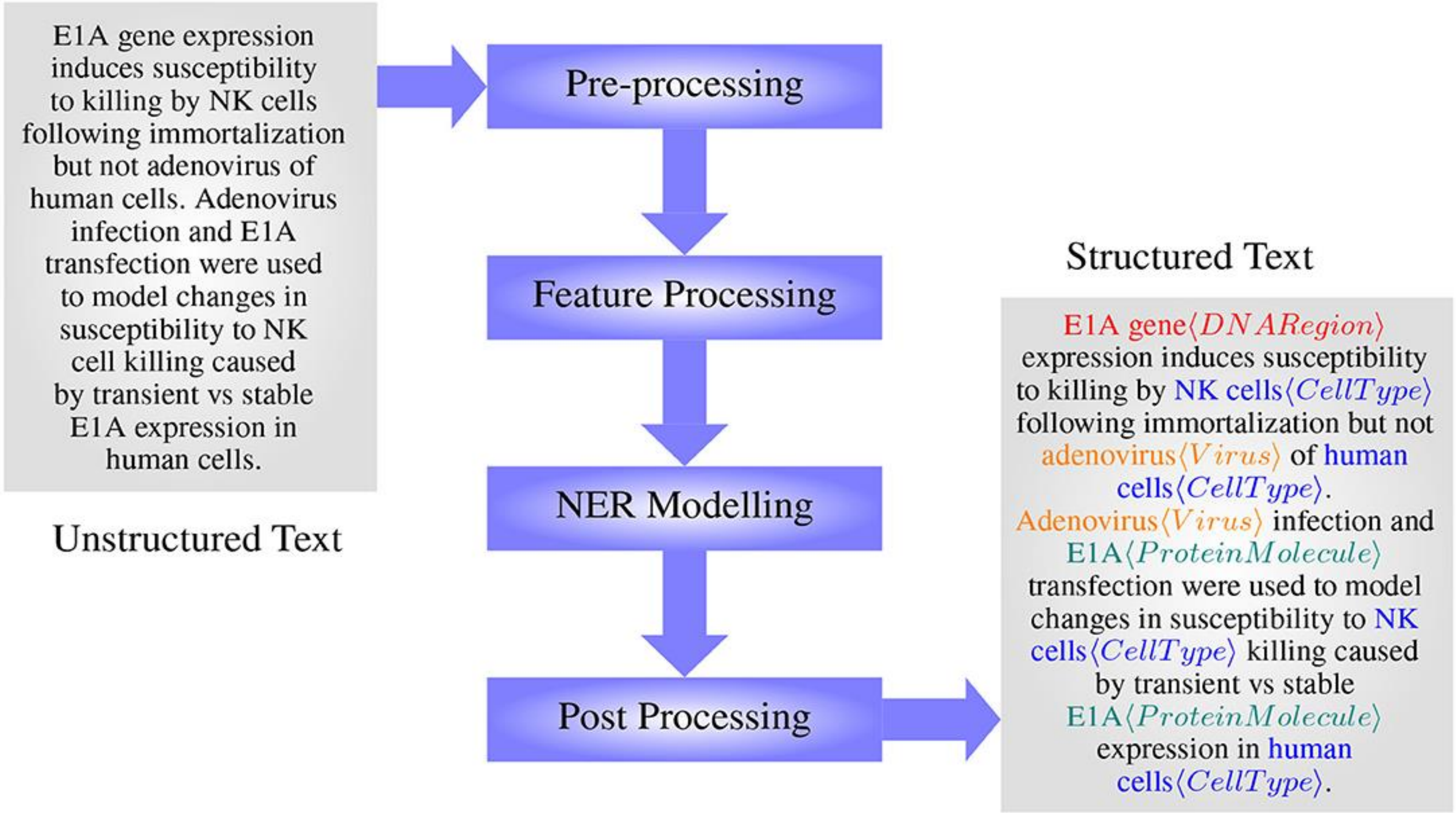
Named Entity Recognition

- A subtask of Information Extraction – that aims to locate and classify Named Elements within unstructured text
- Named Entity Recognition and Classification into *medical concepts*
  - Diseases
  - Symptoms
  - Drugs
  - Genes
  - Chemicals

BioNER Challenges

- Boundary detection
  - *2,4,4,6-Tetramethylcyclohexa-2,5-dien-1-one," "epidemic transient diaphragmatic spasm*
- Phrases
  - Alice in wonderland syndrome - Neuropsychological condition
- *Synonyms*
  - *Lymphocytic Leukemia"* and "*Lymphoblastic Leukemia"*
- Sharing common Head noun in an article
  - *"91 and 84 kDa proteins"*
- Polysemy
  - *GLP1R* may refer to either the gene or protein – needs resolution from context
- Non-standard abbreviations
  - CLD - *Cholesterol-lowering Drug," "Chronic Liver Disease," "Congenital Lung Disease," or "Chronic Lung Disease*

Machine Learning Pipeline – detecting BioNER



BioNER Detection and Classification

Rule Based

- Syntactic features
- Statistically significant features

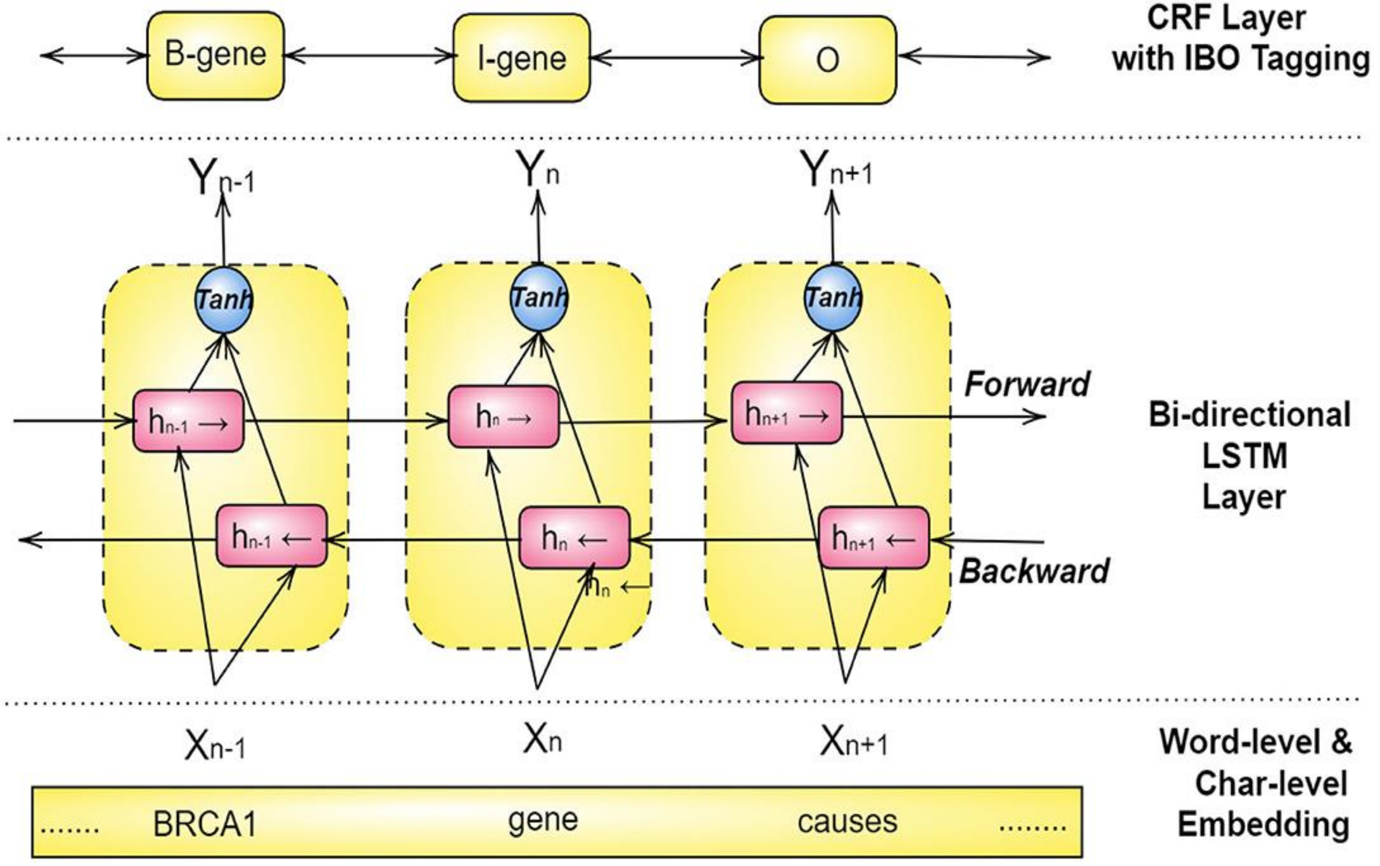
Support Vector Machines, Decision Trees, Conditional Random Fields

- Semantic features – Words, POS tags, dependencies

Deep Neural Networks

- LSTM with Word Embeddings - sequence probabilities
- Transformers – Contextual Text Embedding

BIO tagging for NER



Sequence Labeling Architecture

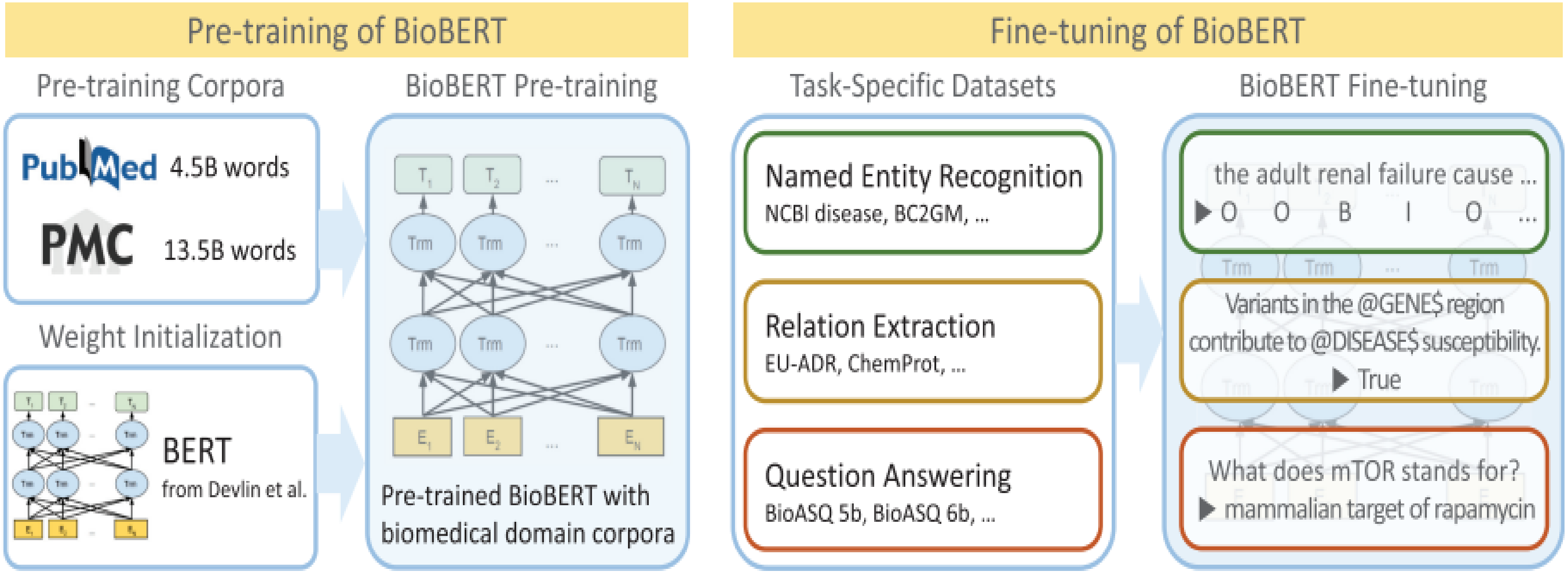
A few words on building models

- Training is expensive
- Annotated resources are not available
  - Spent resources for building annotated data sets
    - Expert availability is a problem
- Transfer Learning
  - Trained on areas for which we have experts
  - Fine-tuned with small sets
  - Distant supervision

Annotated Entity Dataset

Dataset	Entity type	Number of annotations
NCBI Disease (Doğan *et al.*, 2014)	Disease	6881
2010 i2b2/VA (Uzuner et al., 2011)	Disease	19 665
BC5CDR (Li *et al.*, 2016)	Disease	12 694
BC5CDR (Li *et al.*, 2016)	Drug/Chem.	15 411
BC4CHEMD (Krallinger *et al.*, 2015)	Drug/Chem.	79 842
BC2GM (Smith *et al.*, 2008)	Gene/Protein	20 703
JNLPBA (Kim *et al.*, 2004)	Gene/Protein	35 460
LINNAEUS (Gerner *et al.*, 2010)	Species	4077
Species-800 (Pafilis *et al.*, 2013)	Species	3708

BioBERT – a pre-trained model for BioNER



Bio-medical Named Entity Recognizers

Tools	Source	Description
CLiNER	https://github.com/text-machine-lab/CliNER	Clinical Named Entity Recognition system (CliNER) is an open-source natural language processing system for named entity recognition in clinical text of **electronic health records.**
C-NER	Clinical Named Entity Recognition (NER)	
SCiSpacy	https://allenai.github.io/scispacy/	scispaCy is a Python package containing spaCy models for processing **biomedical, scientific or clinical text.**
PICO_Parser	https://github.com/Tian312/PICO_Parser	Recognize PICO Elements from **Clinical Trial Literature** – (Patient, Intervention, Condition, Outcome)
BIO_POS_DEP	https://github.com/datquocnguyen/BioPosDep	Biomedical POS tagging and dependency parsing models

Example 2

Sample Text

The patient was given some hydrocodone for control of her pain. The patient suffers from bulimia and eating disorder, bipolar disorder, and severe hypokalemia.

Label	Id	Type	Context
Hydrocodone	lxg-079584b7217f	Drug	Medication (Ingredients), Medication (Generic)
Control brand of phenylpropanolamine	lxg-b3f1a567f46b	Drug	Medication (Brand), Medication (Ingredients)
Pain	lxg-28ab8b90a066	Disease	Current issue
Bulimia	lxg-793bc328e6b8	Disease	Chronic ailments
Eating Disorders	lxg-aa5be6fbee41	MentalHealth	
Bipolar Disorder	lxg-ca57a7a5d0c9	MentalHealth	
Hypokalemia	lxg-335c798f8bcc	Disease	Modifier

Analytics Tasks

Document Classification

No fever or chills. No nausea or vomiting. He had a mild fever and fainted. He does not recall falling. There is no apparent fracture.

Diagnosis

The patient was given some hydrocodone for control of her pain. The patient suffers from bulimia and eating disorder, bipolar disorder, and severe hypokalemia.

Prescription

Different types of Contexts

- ✓ **Negation and Hypothetical**
- ✓ **Vitals and Lab Values**
- ✓ **Medication - dosage**
- ✓ **Anatomy**
- ✓ **History – past / present**
- ✓ **Allergies**
- ✓ **Modifiers**

Deriving more context

Negation

Deriving more context

Medication

Drug
Dose

Entities are not enough

Entities by themselves are not enough – relationship between entities or concepts around entities are more important

- The **AIDS pandemic** was caused by the spread of **HIV infection.**
- **Infectious diseases** or **communicable diseases** are caused by **bacteria, viruses,** and **parasites.**
- **Serious adverse effect** was observed in patients with heart disease due to high dosage of Spironolactone

Cause – effect relations

Example 1

Sample Text

No fever or chills. No nausea or vomiting. He had a mild fever and fainted. He does not recall falling. There is no apparent fracture.

Label	Id	Type	Reference text
Fever	lxg-45c0d6e146f4	Disease	... No [**fever**] or chills ...
Chills	lxg-fc8fdc162ff2	Disease	... fever or [**chills**] . No ...
Nausea	lxg-9bdf71250b2d	Disease	... . No [**nausea**] or vomiting ...
Vomiting	lxg-5f5bf2c446ed	Disease	... nausea or [**vomiting**] . He ...
Fever	lxg-45c0d6e146f4	Disease	... a mild [**fever**] and fainted ...
Syncope	lxg-7bf138779152	Disease	... fever and [**fainted**] . He ...
Accidental Falls	lxg-91f1adb6e169	Other	... not recall [**falling**] . There ...
Fractures, Bone	lxg-2d8902924530	Disease	... no apparent [**fracture**] .

Presenting Contexts

No fever or chills. No nausea or vomiting. He had a mild fever and fainted. He does not recall falling. There is no apparent fracture.


```
"contexts": [
    {
        "type": "negative",
        "subtype": "definite",
        "explanation": {
            "begin": 0,
            "end": 14,
            "matchedTokens": [
                {
                    "token": "no",
                    "position": 0
                },
                {
                    "token": "evidence",
                    "position": 2
                },
                {
                    "token": "of",
                    "position": 4
                }
            ],
            "fuzzyValues": [],
            "triggerId": "N0000057",
            "triggerLabel": "no evidence"
        }
```

Relation extraction – an NLP task

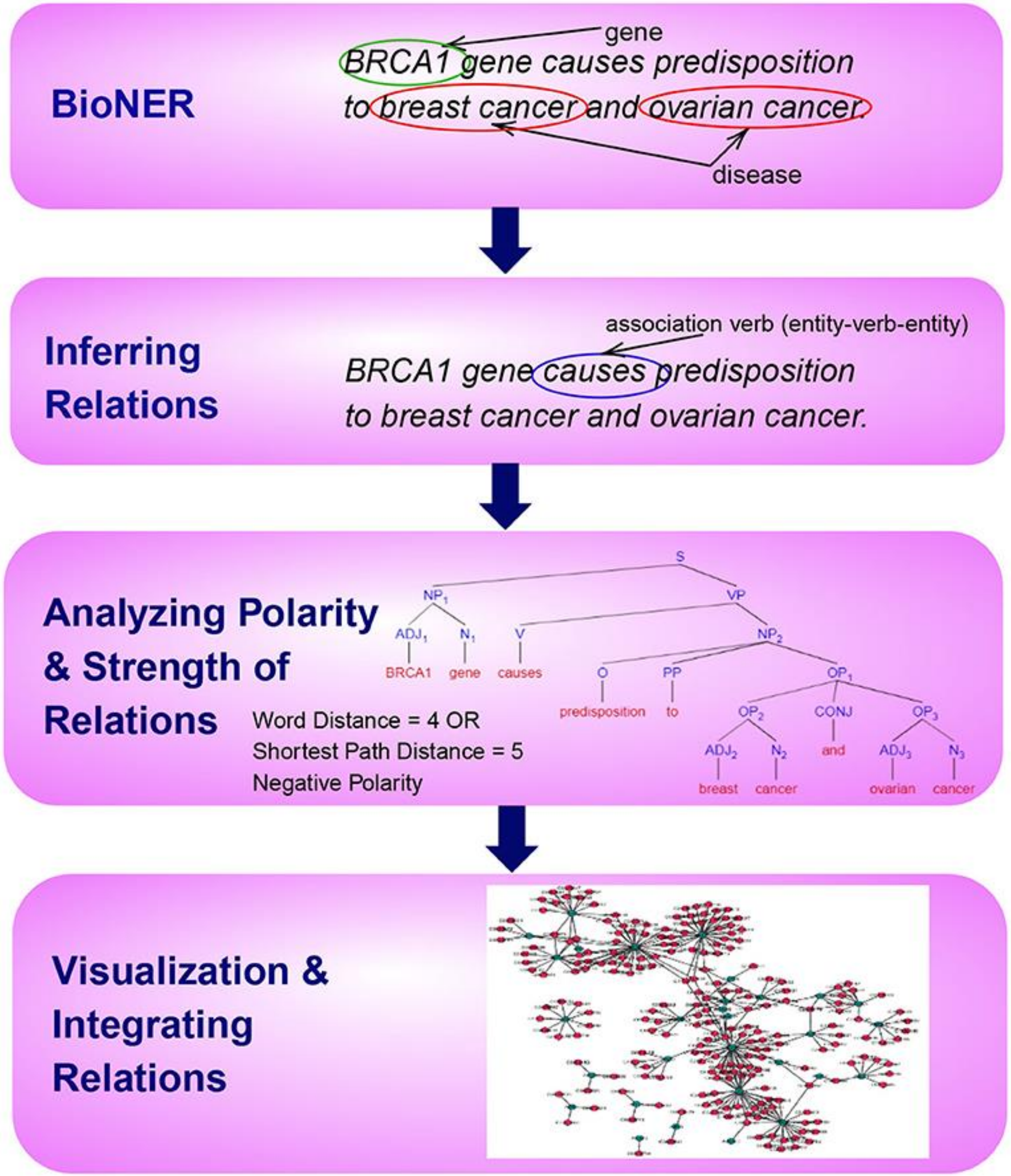
- Relations are defined between entities
  - Defines the context for entities

- Drugs *cure* disease
- Diseases *are caused by* organisms
- Auxins *influence* plant growth
- Genes are *expressed* through the process of protein synthesis.

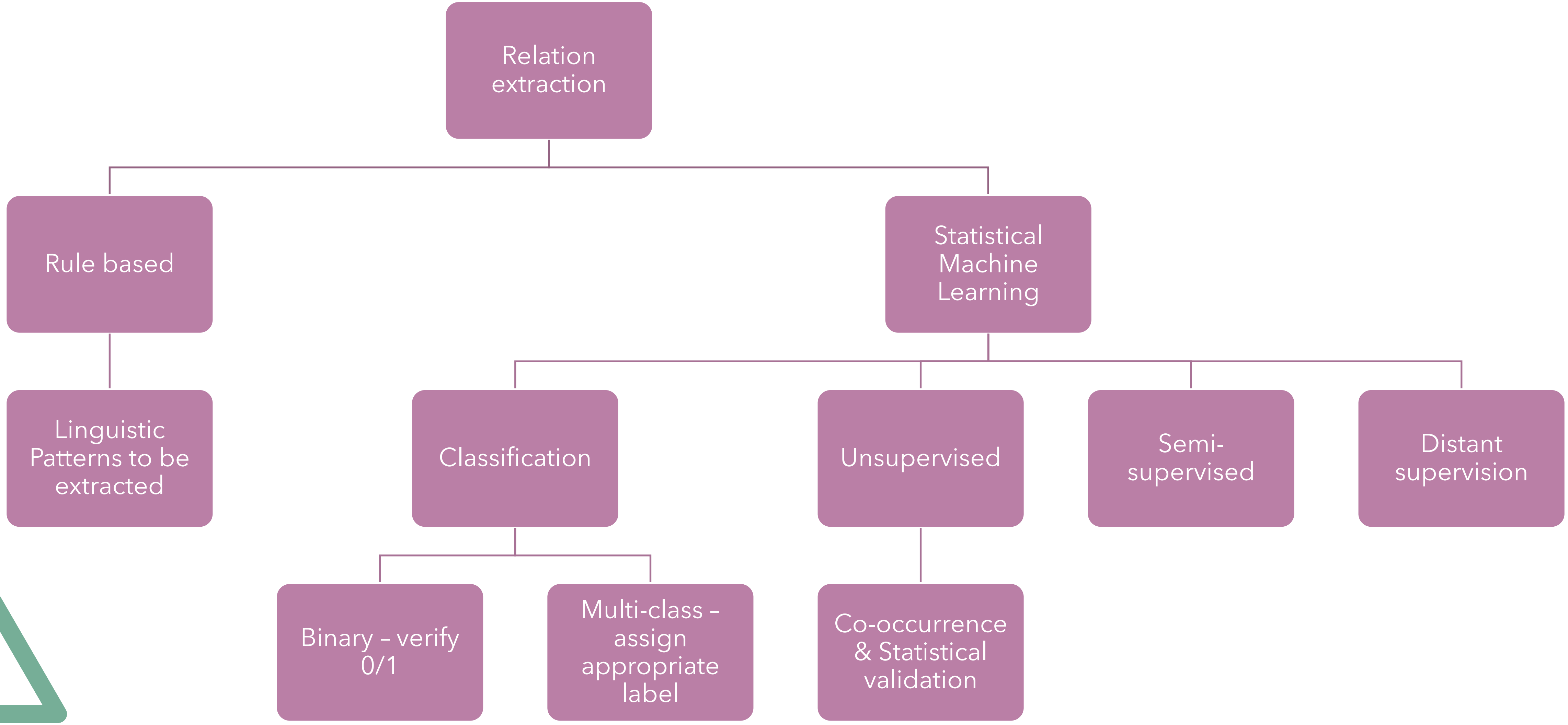
Relation extraction – *Find relations embedded in text*

- Information about possible new drugs from Literature
- Side-effects from Social Media

Inferring Relationships through Big Data Analytics



Bio-medical Relation Extraction



Distant Supervision for Relation Learning

- Applicable when a large pool of relevant but unlabeled data exists
  - Label a small sample using rules / heuristic / noisy operator / existing imperfect classifier
- Small pool of labeled data **may** also exist – but not necessary
- Create a model utilizing the original labeled training data if it existed and the new noisily labeled data to create final output

Obtaining Annotations for Distant Supervision

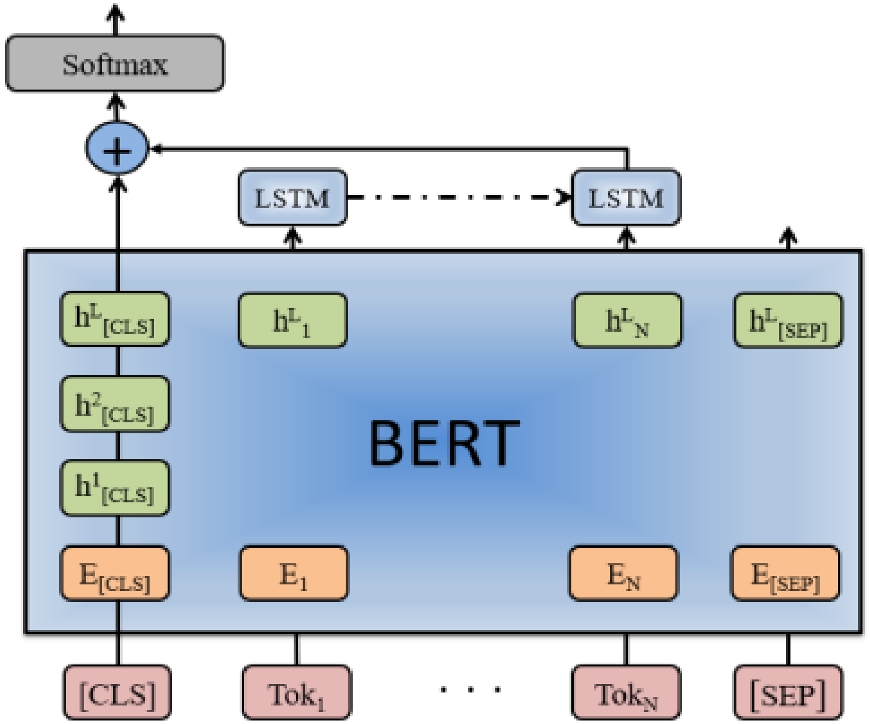
- Weak Labels: Non-expert labels from crowdsourcing / heuristic rules
- Constraints: Specified as constraints - entity-based constraints
- Distributions: Probability distribution
- Invariances: extend the coverage of the labeled distribution to all transformations e.g. – if it is applicable for one entity, it is applicable to conjunction of similar entities

https://www.snorkel.org/blog/weak-supervision

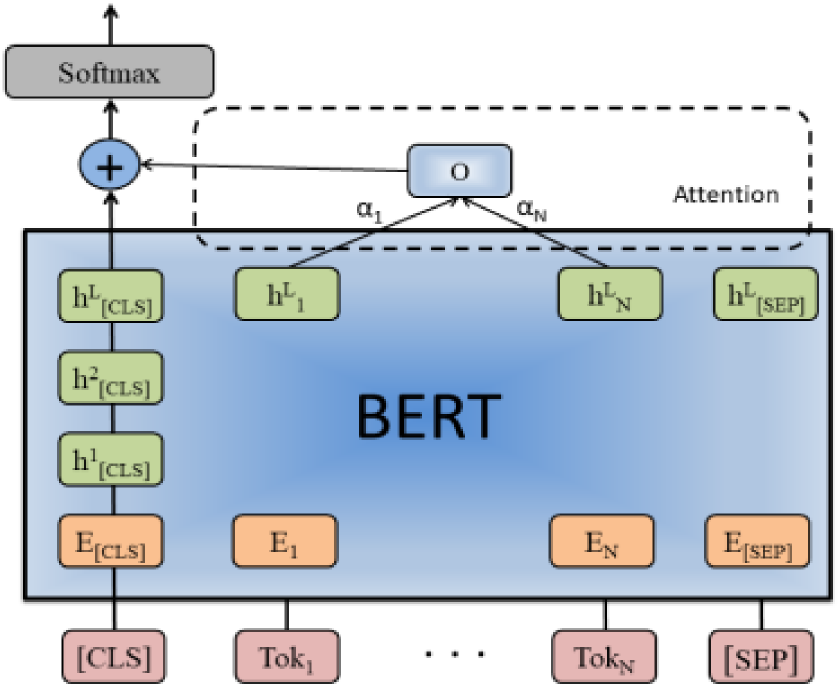
Relation extraction using BERT based classification

- Assume a set of relations
  - Example label set L = {ADVICE, EFFECT, INT, MECHANISM}
- Relation extraction as a classification problem
  - Does this text contain a Drug-Drug Interaction relation?
- Example –
  - S = Tok1 Tok2 . . . Tokn
  - Sentence has to contain drug occurrences, otherwise it will be seen as negative
  - A few tokens should be labeled by two entities Et1, Et2
  - Model needs to predict the probability of each label P(L|Tok1, Tok2, . . . , Tokn, Et1, Et2) based on the sentence text in the sentence

BERT based Architectures



(a) LSTM on the last layer



(b) Attention mechanism on the last layer

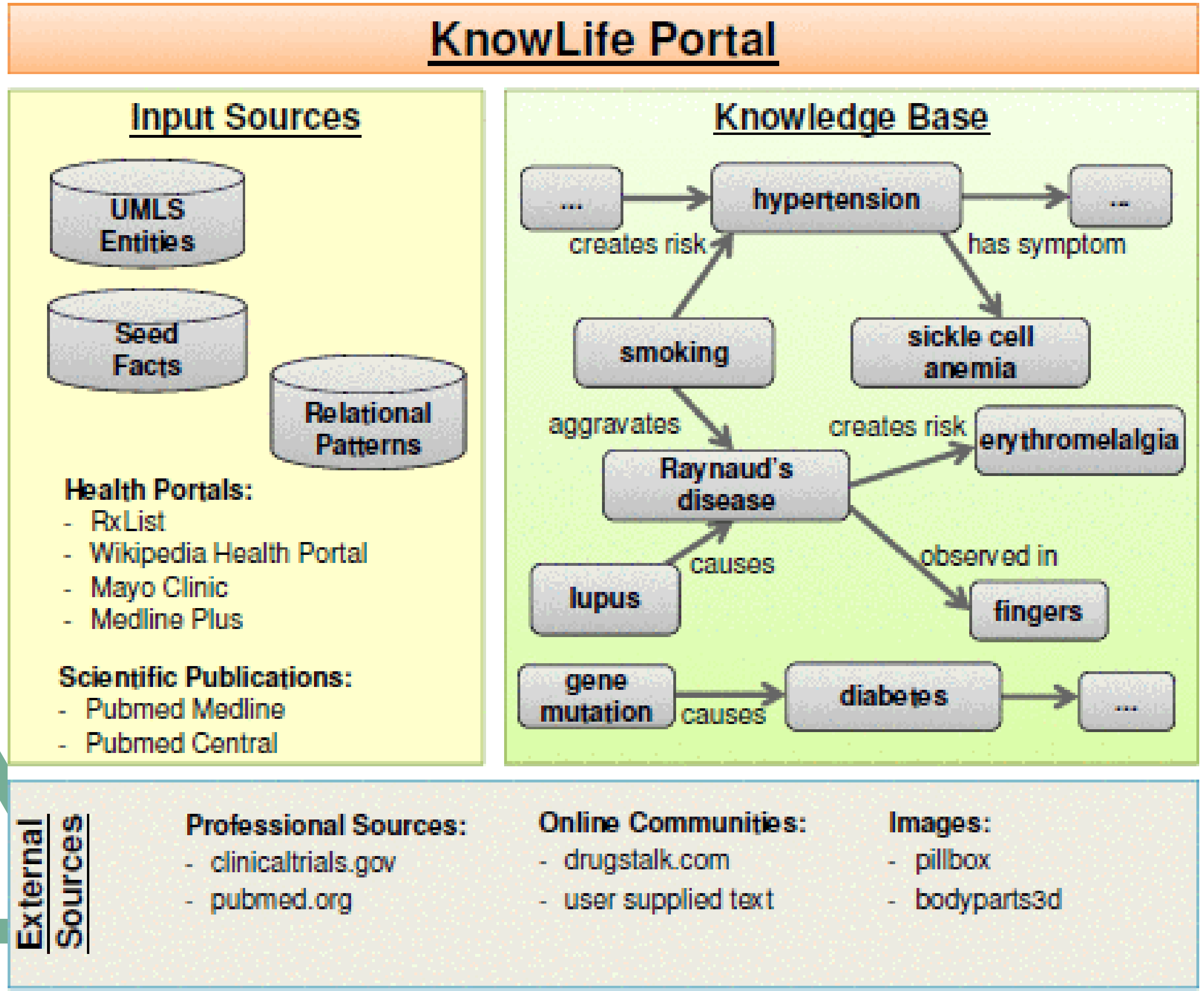
Focus shift needed

Dataset	Entity type	Number of relations
GAD (Bravo et al., 2015)	Gene–disease	5330
EU-ADR (Van Mulligen *et al.*, 2012)	Gene–disease	355
CHEMPROT (Krallinger *et al.*, 2017)	Protein–chemical	10 031

Relation between diseases and other risk factors

- nutritional habits
- life style
- side effects of combinations of drugs rather than single drugs alone
- analyzing the experience of patients on mass diseases such as asthma or diabetes

KnowLife: A knowledge graph for health and life sciences



- Using advanced IE methods for constructing a large KB on a wide range of health-centric relations with entity linking to the Unified Medical Language System (UMLS, uts.nlm.nih.gov): a total of 214k canonical entities and 78k facts for 14 relations
- Tapping into health-related online forums for evidence for relational facts and populate KG
- Automatically annotating new documents from scientific literature or from social media with relevant entities and relationships

https://ieeexplore.ieee.org/document/6816754

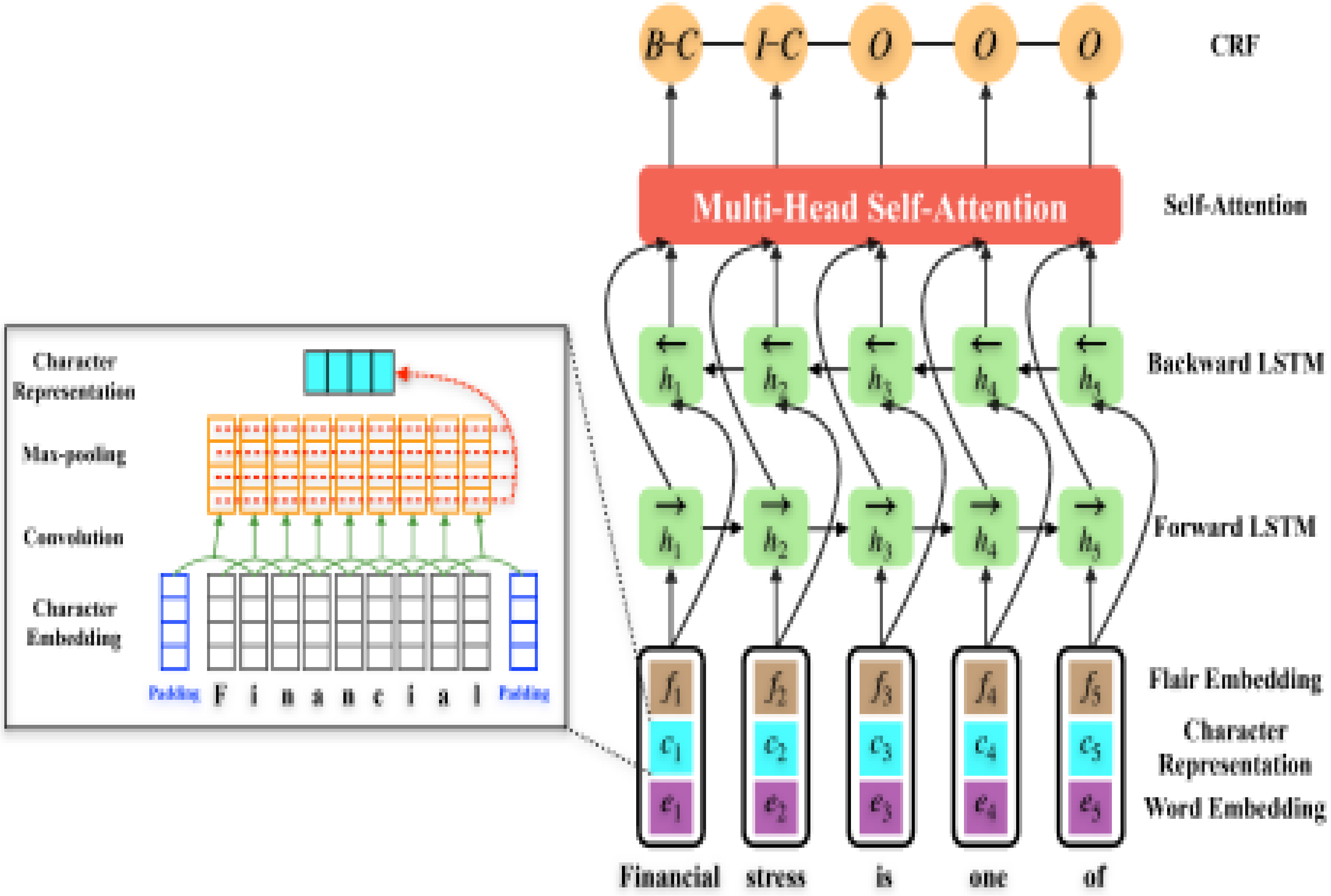
Use Case 1 – Extract Causal relations from text

Identify drug side-effects

Deep Learning Mechanisms for Relation Extraction

- ✓ Use Deep recurrent Networks
- ✓ Train relation classifiers

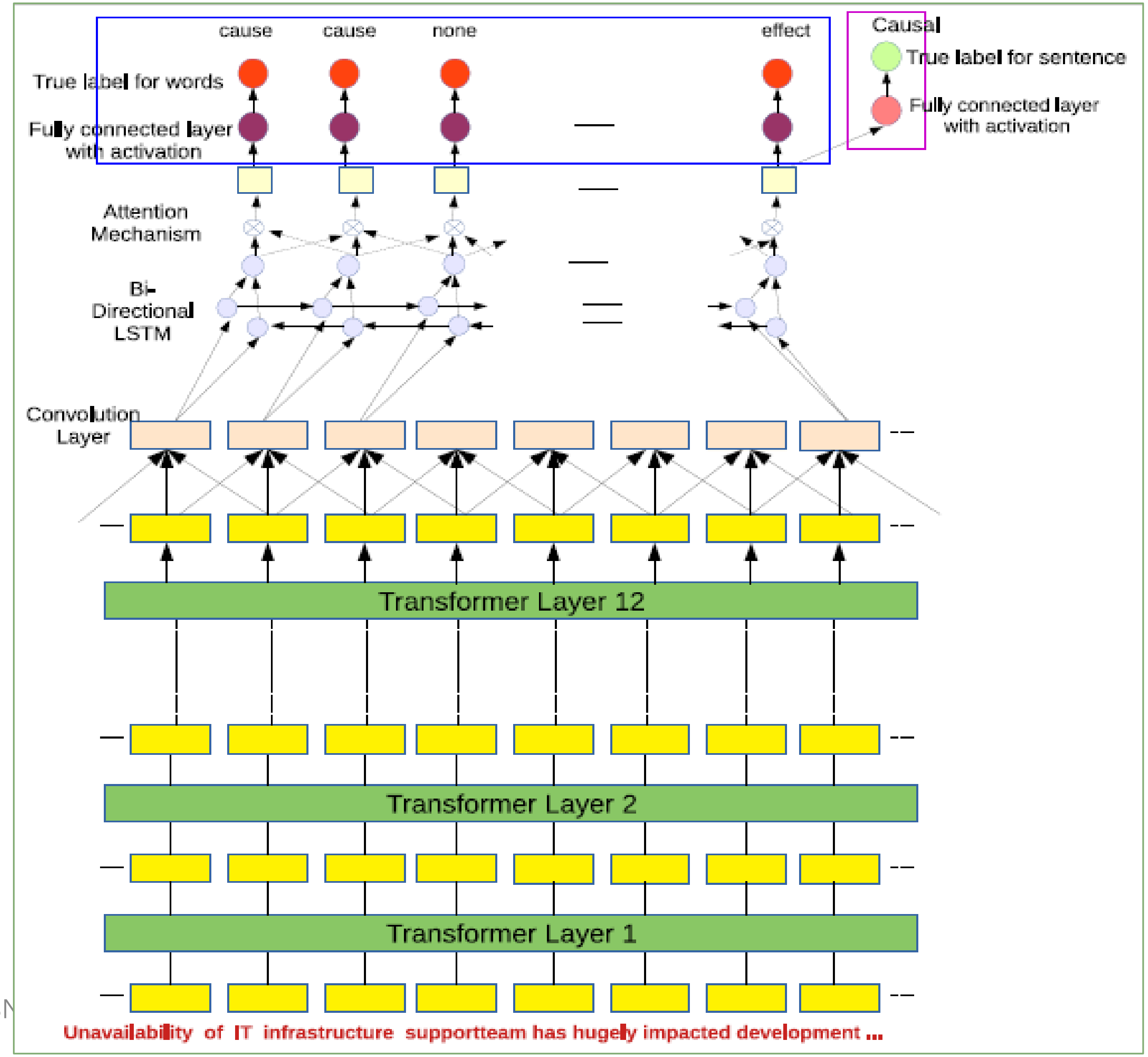
- ✓ Attention mechanism to learn stronger correlations between words, motifs observed for a particular kind of relation



Self-Attentive BiLSTM-CRF with Transferred Embeddings (Li et al., 2020)

Causal Relation Extraction

- Proposed Joint Model for Causal Sentence Classification and Relation Extraction
- Uses a multi-layer bidirectional Transformer encoder architecture based on the transformer model
- 12 layers of transformer blocks, 768 hidden units, and 12 self-attention heads.
- Pretrained with two strategies on large-scale unlabeled text - masked language model and next sentence prediction.
- The pre-trained BERT model provides a powerful context-dependent sentence representation and can be used for various target tasks through the fine-tuning procedure.



Learning The Joint Model for Sentence Classification and Sequence Labeling Tasks

- Loss function

$$L_1(\theta) = -\sum_{t=1}^{M}\sum_{k=1}^{K} \bar{y}_t^k log(y_t)$$

$$L_2(\theta) = -\sum_{t=1}^{N}\sum_{j=1}^{J} \bar{q}_t^{i,j} log(q_t^i)$$

$$L_{joint}(\theta) = \lambda * L_1(\theta) + (1-\lambda) * I_{[y_{sentence}==1]} * L_2(\theta)$$

Resource Creation – Causal fragments

Source	Sentence count	Average sentence length
Analyst Report (AR)	4500	23.7
SEMEVAL (SEM)	1331	18.7
BBC News (BBC)	503	22.5
ADE	3000	20.5
Recall News (RN)	1052	23.1

Results

Features	Dataset	Precision	Recall	F1-Score
BiLSTM	Analyst Reports(AR)	0.71 ± 0.11	0.87± 0.09	0.78 ± 0.11
	BBC News (BBC)	0.71± 0.14	0.82± 0.03	0.75± 0.08
	SEMEVAL(SEM)	0.79± 0.11	0.87± 0.14	0.81± 0.01
	Adverse Drug(ADE)	0.73± 0.07	0.86± 0.03	0.75± 0.06
	Recall News (R)	0.87± 0.02	0.93± 0.14	0.87± 0.10
CNN+BiLSTM	Analyst Reports(AR)	0.84± 0.11	0.87± 0.04	0.82± 0.03
	BBC News (BBC)	0.81± 0.01	0.82± 0.03	0.79± 0.04
	SEMEVAL(SEM)	0.89± 0.03	0.87± 0.01	0.86± 0.05
	Adverse Drug(ADE)	0.83± 0.07	0.96± 0.02	0.90± 0.05
	Recall News (R)	0.87± 0.08	0.93± 0.05	0.89± 0.06
$BERT_{base}$	Analyst Report	0.87± 0.14	0.92± 0.09	0.85± 0.07
	BBC News (BBC)	0.81± 0.01	0.92± 0.05	0.84± 0.06
	SEMEVAL(SEM)	0.91± 0.05	0.97± 0.03	0.95± 0.02
	Adverse Drug(ADE)	0.91± 0.03	0.96± 0.06	0.93± 0.05
	Recall News (R)	0.87± 0.04	0.93± 0.05	0.90± 0.07
BERT+CNNBiLSTM	Analyst Report	0.91± 0.05	0.97± 0.07	0.90± 0.02
	BBC	0.94± 0.13	0.97± 0.16	0.94± 0.09
	SEMEVAL	0.91± 0.15	0.97± 0.07	0.94± 0.05
	Adverse Drug	0.89± 0.05	0.97± 0.01	0.95± 0.05
	Recall News (R)	0.87± 0.02	0.93± 0.01	0.90± 0.02

Use Case 2 - Enabling Automatic Patient Recruitment For Clinical Trials

Clinical Trials

Research studies that are aimed at evaluating a medical, surgical, or behavioral intervention.

Whether a new treatment, like a new drug or diet or medical device is more effective than the existing treatments for a particular ailment.

Clinical Trial description – a sample

TITLE: Randomized Trial of Acetazolamide for Uveitis-Associated Cystoid Macular Edema
CONDITION: Macular Edema, Cystoid
INTERVENTION: Acetazolamide
SUMMARY: To test the efficacy of acetazolamide for the treatment of uveitis-associated cystoid macular edema.
DETAILED DESCRIPTION: Uveitis, an intraocular inflammatory disease, is the cause of about 10 percent of visual impairment in the United States. Uveitis may lead to many sight-threatening conditions including cataract, vitreal opacities, glaucoma, and, most commonly, cystoid macular edema. Reduction of swelling or edema within the retina depends on the movement of fluid from the retina through the choroid. A number of studies indicate that this process requires active transport of fluid ions by the retinal pigment epithelium and may involve the carbonicm anhydrase system.
ELIGIBILITY Gender: All Age: 8 Years to N/A

Males and females 8 years of age or older and weighing at least 35 kg (77 lb) were eligible for the study. Patients had to have a best corrected visual acuity of 20/40 or worse in at least one eye with cystoid macular edema demonstrable on fluorescein angiography. Patients were allowed to receive systemic therapy for their uveitis. Exclusion criteria included current use of acetazolamide as part of a therapeutic regimen; a history of hypersensitivity reactions to acetazolamide, sulfonamides, or angiography dye; unclear ocular media that would obscure fluorescein angiography. Also, patients with macular subretinal for medical reasons may be excluded.

What is the analytics use-case here?

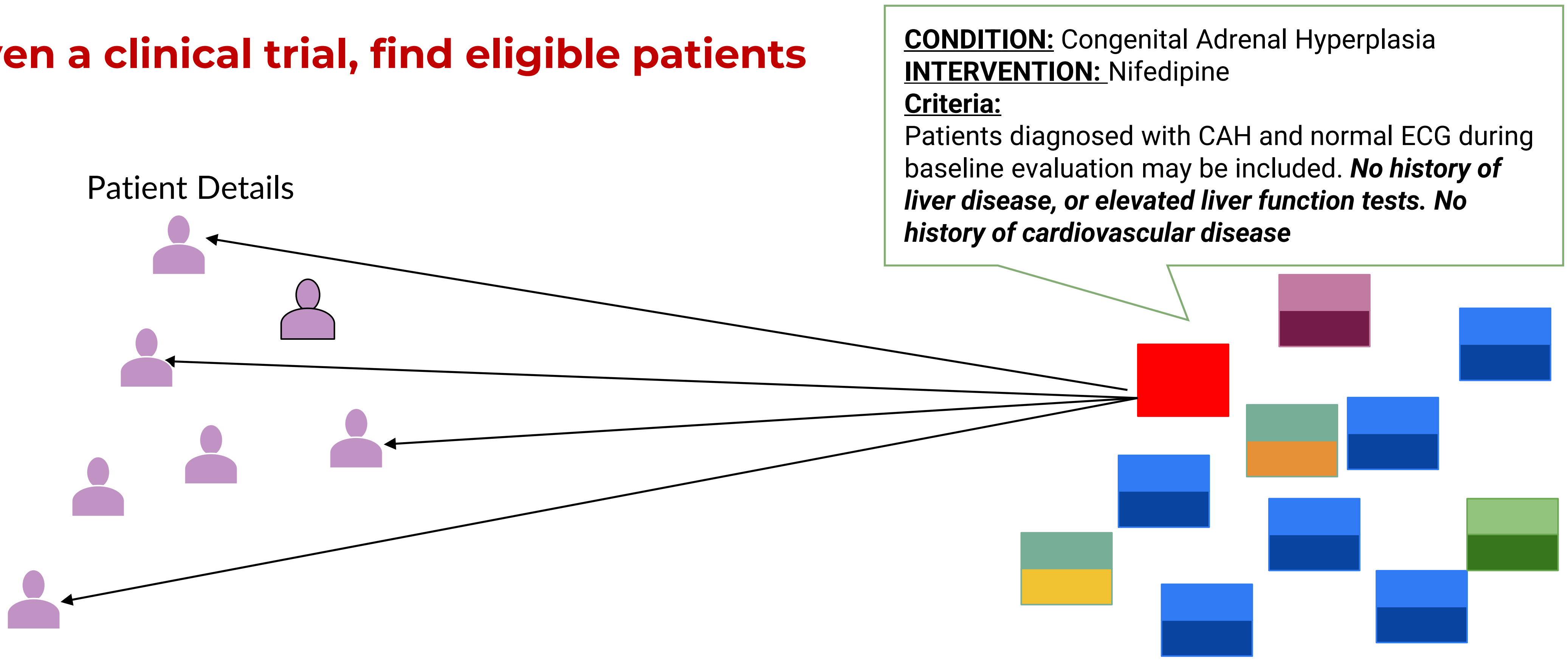
- ❑ A successful completion of a trial is dependent on achieving a significant sample size of patients enrolled into the trial within a limited time period.
- ❑ Identification of participants for a given trial is not trivial.
  - ❑ Involves repeated readings of the patient's electronic health record (EHR) for matching against the inclusion/exclusion conditions.

The Inclusion/Exclusion do not follow standard specification format – makes the problem more complex

- Limits the number of patients that can be evaluated.
- May overlook adverse drug reactions

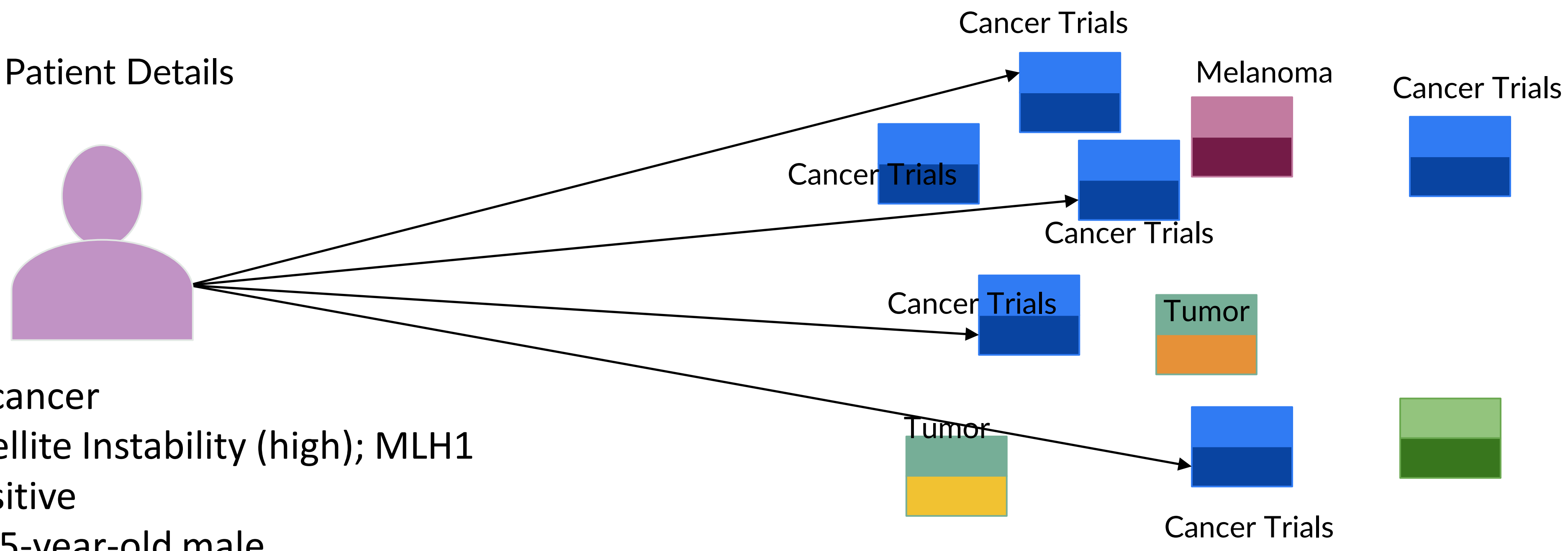
Text REtrieval Conference (TREC) task1

Task-1: Given a clinical trial, find eligible patients



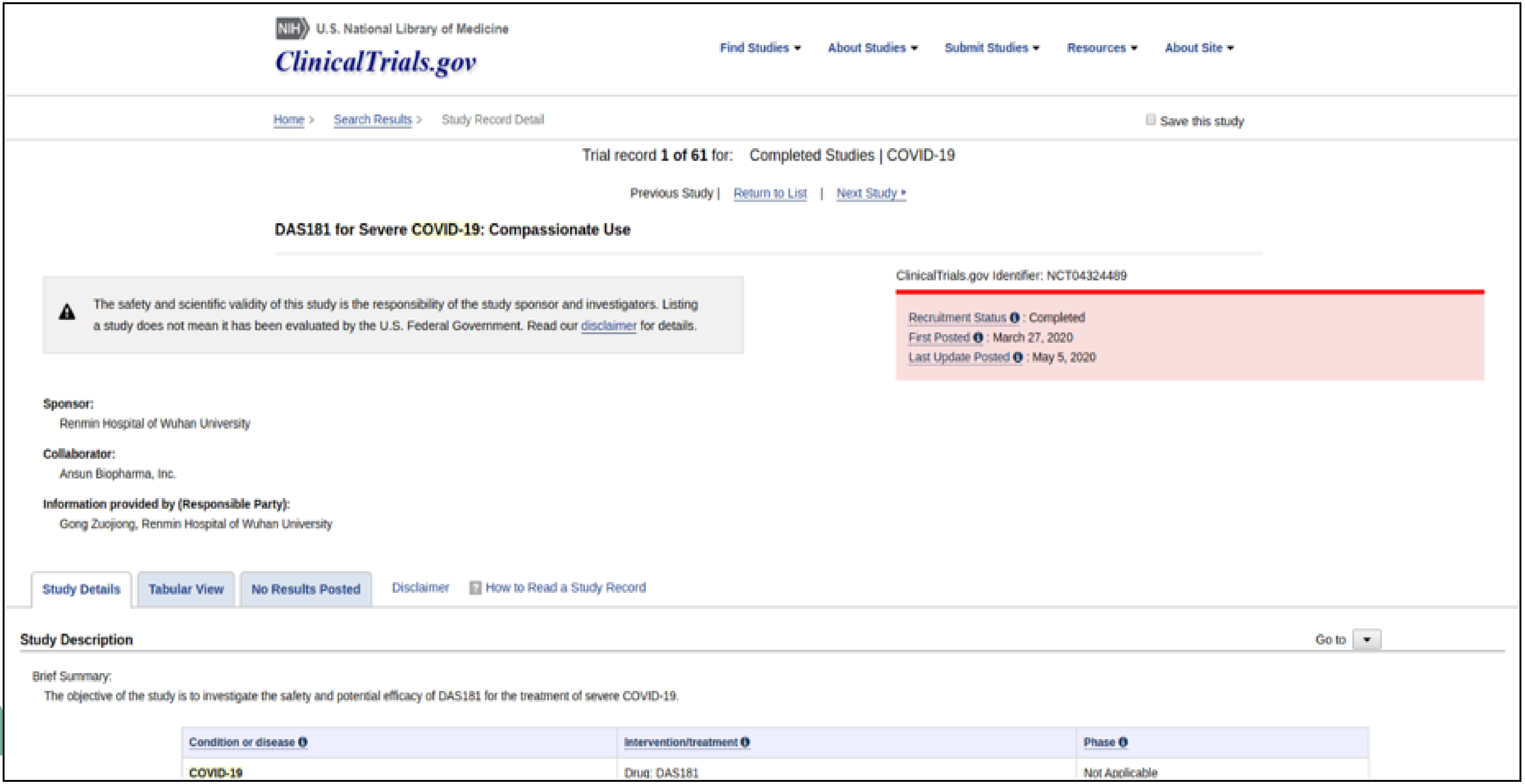
Text REtrieval Conference (TREC) task2

Task-2: **Given a patient detail, find its relevant trials**



Disease: Colon cancer
Gene: Microsatellite Instability (high); MLH1 methylation positive
Demographic: 65-year-old male
Other: Hypercholesterolemia, hypertension

Clinical Trial Search



U.S. National Library of Medicine

ClinicalTrials.gov

Find Studies ▾ About Studies ▾ Submit Studies ▾ Resources ▾ About Site ▾

Home > Search Results > Study Record Detail

Save this study

Trial record **1 of 61** for: Completed Studies | COVID-19

Previous Study | Return to List | Next Study ▸

DAS181 for Severe COVID-19: Compassionate Use

The safety and scientific validity of this study is the responsibility of the study sponsor and investigators. Listing a study does not mean it has been evaluated by the U.S. Federal Government. Read our disclaimer for details.

ClinicalTrials.gov Identifier: NCT04324489

Recruitment Status : Completed
First Posted : March 27, 2020
Last Update Posted : May 5, 2020

Sponsor:
Renmin Hospital of Wuhan University

Collaborator:
Ansun Biopharma, Inc.

Information provided by (Responsible Party):
Gong Zuojiong, Renmin Hospital of Wuhan University

Study Details | Tabular View | No Results Posted | Disclaimer | How to Read a Study Record

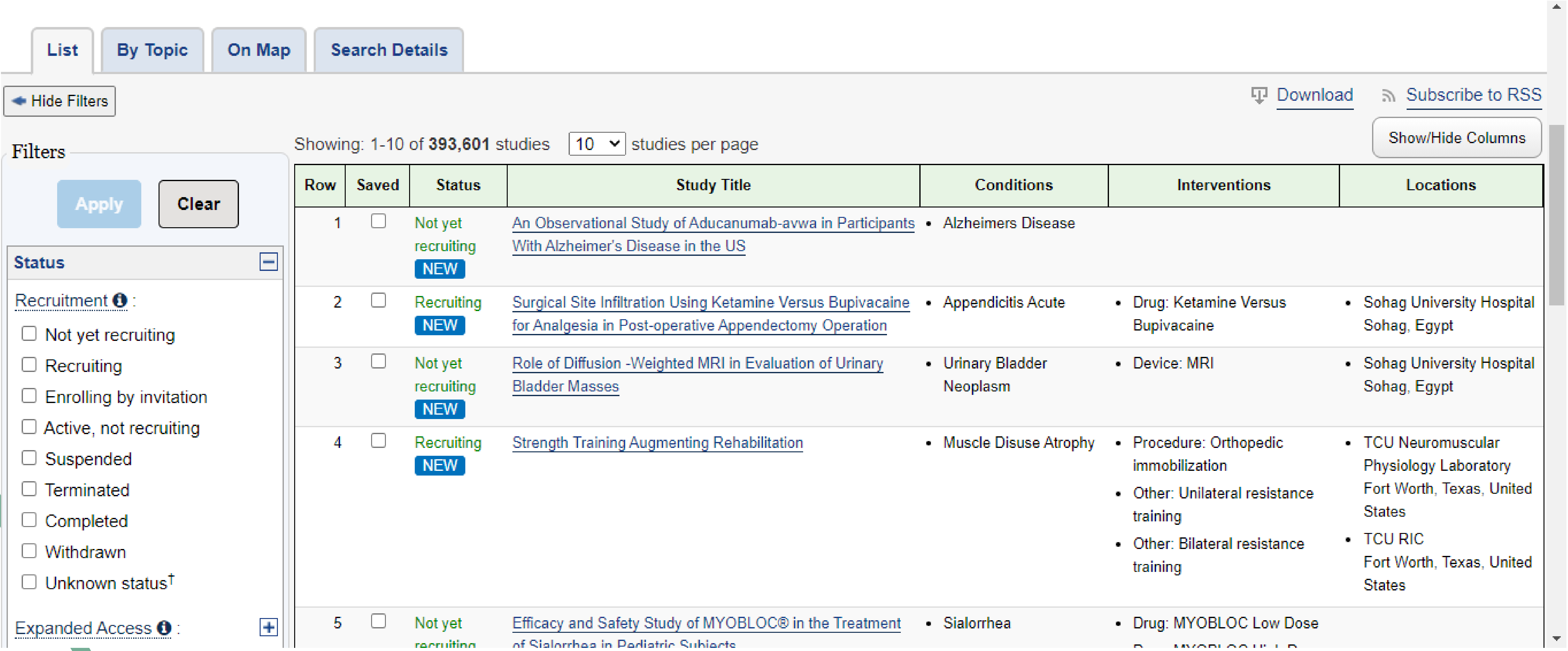
Study Description

Go to ▾

Brief Summary:
The objective of the study is to investigate the safety and potential efficacy of DAS181 for the treatment of severe COVID-19.

Condition or disease	Intervention/treatment	Phase
COVID-19	Drug: DAS181	Not Applicable

Existing Search Engines/Filters



List | By Topic | On Map | Search Details

Hide Filters

Download | Subscribe to RSS

Filters

Apply | Clear

Status

Recruitment :

- Not yet recruiting
- Recruiting
- Enrolling by invitation
- Active, not recruiting
- Suspended
- Terminated
- Completed
- Withdrawn
- Unknown status†

Expanded Access :

Showing: 1-10 of **393,601** studies | 10 studies per page

Show/Hide Columns

Row	Saved	Status	Study Title	Conditions	Interventions	Locations
1		Not yet recruiting NEW	An Observational Study of Aducanumab-avwa in Participants With Alzheimer's Disease in the US	• Alzheimers Disease		
2		Recruiting NEW	Surgical Site Infiltration Using Ketamine Versus Bupivacaine for Analgesia in Post-operative Appendectomy Operation	• Appendicitis Acute	• Drug: Ketamine Versus Bupivacaine	• Sohag University Hospital Sohag, Egypt
3		Not yet recruiting NEW	Role of Diffusion -Weighted MRI in Evaluation of Urinary Bladder Masses	• Urinary Bladder Neoplasm	• Device: MRI	• Sohag University Hospital Sohag, Egypt
4		Recruiting NEW	Strength Training Augmenting Rehabilitation	• Muscle Disuse Atrophy	• Procedure: Orthopedic immobilization • Other: Unilateral resistance training • Other: Bilateral resistance training	• TCU Neuromuscular Physiology Laboratory Fort Worth, Texas, United States • TCU RIC Fort Worth, Texas, United States
5		Not yet recruiting	Efficacy and Safety Study of MYOBLOC® in the Treatment of Sialorrhea in Pediatric Subjects	• Sialorrhea	• Drug: MYOBLOC Low Dose	

Problem Statement 1

TITLE: Randomized Trial of Acetazolamide for Uveitis-Associated Cystoid Macular Edema
CONDITION: Macular Edema, Cystoid
INTERVENTION: Acetazolamide
SUMMARY: To test the efficacy of acetazolamide for the treatment of uveitis-associated cystoid macular edema.
DETAILED DESCRIPTION: Uveitis, an intraocular inflammatory disease, is the cause of about 10 percent of visual impairment in the United States. Uveitis may lead to many sight-threatening conditions including cataract, vitreal opacities, glaucoma, and, most commonly, cystoid macular edema. Reduction of swelling or edema within the retina depends on the movement of fluid from the retina through the choroid. A number of studies indicate that this process requires active transport of fluid ions by the retinal pigment epithelium and may involve the carbonicm anhydrase system.
ELIGIBILITY Gender: All Age: 8 Years to N/A

Inclusion

Males and females 8 years of age or older and weighing at least 35 kg (77 lb) were eligible for the study. Patients had to have a best corrected visual acuity of 20/40 or worse in at least one eye with cystoid macular edema demonstrable on fluorescein angiography. Patients were allowed to receive systemic therapy for their uveitis. Exclusion criteria included current use of acetazolamide as part of a therapeutic regimen; a history of hypersensitivity reactions to acetazolamide, sulfonamides, or angiography dye; unclear ocular media that would obscure fluorescein angiography. Also, patients with macular subretinal for medical reasons may be excluded.

Exclusion

Matching Pipeline

1. Find Inclusion – Exclusion segments within a description
2. Within each segment – label text sequences as clinical aspects
3. Within each clinical aspect – identify Named Entities with their labels
4. Detect paraphrases – different ways of saying the same thing
5. Derive match score using Named Entities

Inclusion - Patients had to have a best corrected visual acuity of 20/40 or worse in at least one eye with cystoid macular edema

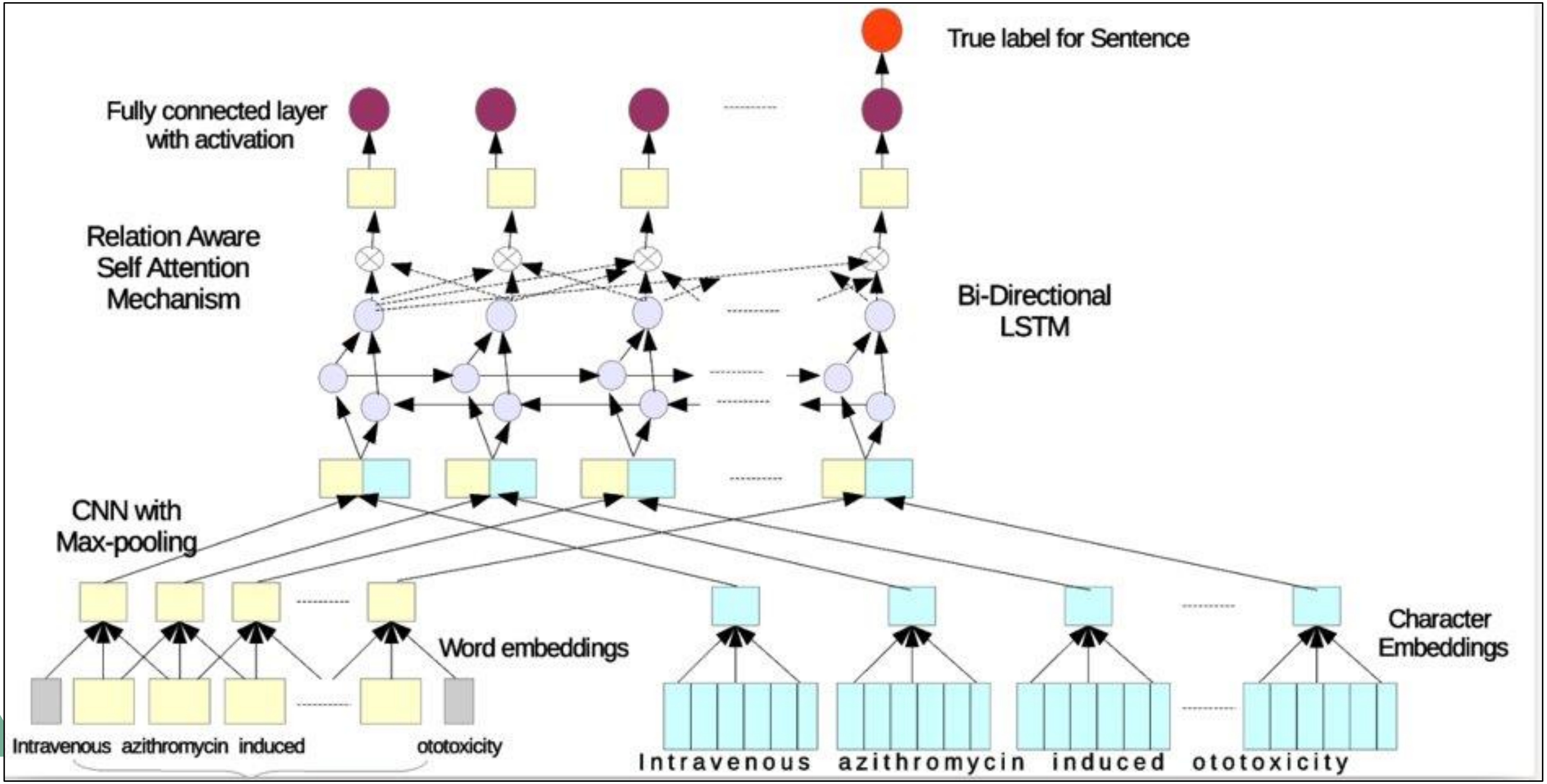
Exclusion - a history of hypersensitivity reactions to acetazolamide, **sulfonamides**, or angiography dye

Look up appropriate content in EHR

Automatic Segregation and Classification of Inclusion and Exclusion Criteria – Text Sequence Labeling task

Sentences	Prediction
Patients with active infections, including HIV, will be excluded, due to unknown effects of the vaccine on lymphoid precursors.	exclusion
Patients with HIV infection (but not AIDS) are eligible for this trial. Therefore, no HIV testing will be required.	Inclusion
HIV-1 infection as documented by any licensed ELISA (enzyme-linked immunosorbent assay) test kit and confirmed by Western blot at any time prior to study entry; HIV-1 culture, HIV-1 antigen, plasma HIV-1 ribonucleic acid (RNA), or a second antibody test by a method other than ELISA is acceptable as an alternative confirmatory test.	Inclusion
Patients who are HIV seropositive can have decreased immune competence and thus be less responsive to the experiment and more susceptible to its toxicities).	exclusion

Sequence Labeling Task



Dataset Available for Research

Source: TREC 2018 Precision Medicine Task, Clinical trials

Dataset-1	Inclusion instances	Exclusion instances	Total instances
Training Data	10,000	12,280	22,280
Test Data	8,000	5,700	13,700

https://www.kaggle.com/auriml/eligibilityforcancerclinicaltrials

Dataset-1	Inclusion instances	Exclusion instances	Total instances
Training Data	29,855	32,280	61865
Test Data	11,000	12,000	23,000

Results

(AAAI (W) on Health Intelligence, 2020)

	Dataset-I (TREC-2019)						Dataset-II (Kaggle)					
Word Embedding	Inclusion			Exclusion			Inclusion			Exclusion		
	W	FT	E	W	FT	E	W	FT	E	W	FT	E
BiLSTM	0.70	0.74	0.78	0.72	0.72	0.76	0.70	0.73	0.80	0.71	0.75	0.76
BiLSTM*	0.73	0.76	0.80	0.74	0.75	0.78	0.72	0.77	0.84	0.73	0.76	0.78
CNN	0.68	0.67	0.68	0.65	0.65	0.70	0.71	0.73	0.62	0.61	0.65	0.70
CNN*	0.69	0.71	0.73	0.67	0.66	0.74	0.79	0.77	0.67	0.63	0.68	0.71
C-BiLSTM	0.70	0.72	0.76	0.71	0.71	0.74	0.73	0.74	0.81	0.70	0.72	0.79
C-BiLSTM*	0.77	0.75	0.81	0.73	0.73	0.78	0.75	0.77	0.86	0.72	0.75	0.80
$C-BiLSTM-S_{att}$	0.75	0.78	0.82	0.72	0.73	0.80	0.82	0.83	0.89	0.78	0.85	0.88
$C-BiLSTM-S_{att}*$	0.80	0.82	0.83	0.78	0.76	0.81	0.84	0.86	0.90	0.82	0.88	0.90
$C-BiLSTM-Re_{att}$	0.81	0.79	0.84	0.79	0.80	0.81	0.85	0.87	0.91	0.86	0.85	0.91
$C-BiLSTM-Re_{att}*$	0.84	0.84	0.89	0.82	0.81	0.85	0.87	0.87	0.95	0.88	0.90	0.93

BERT-base	Dataset-I		Dataset-II	
	Inclusion	Exclusion	Inclusion	Exclusion
BERT-base	0.86	0.80	0.91	0.92

Problem-2: Matching EHRs against trial – more complex problem

- Medical Literature proposed five different clinical aspects
  - Lab Test Results
  - Health Status
  - Treatment status
  - Demography
  - Lifestyle

Total bilirubin less than or equal to 1.5 mg/dl, except in patients with history of anaemia. Have had their ileostomy or colostomy for at least 3 months. Subjects must be between the ages of 18 to 65 years old and must not intake alcohol. Pregnant women of stage 3 and age 18 years and older attending delivery room. Life expectancy of at least 6 months and willing to provide informed consent. Live vaccine within 4 weeks prior to therapy or potential need for a live vaccine. Serum ALAT or serum ASAT > 5 x upper limit of normal (ULN) at screening. Current alcohol abuse or drug addiction that in the opinion of the investigator

Health status

Lab Test Results

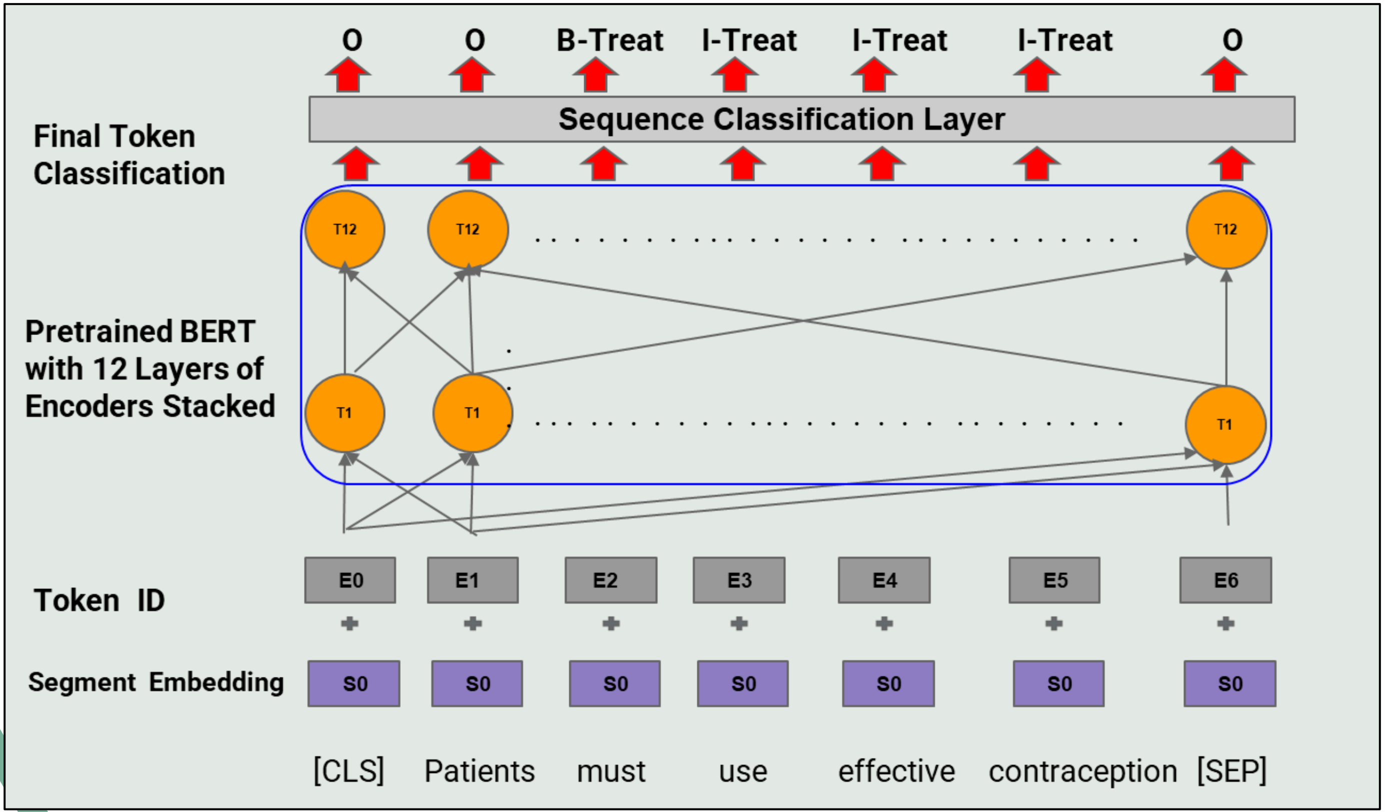
Treatment status

Demography status

Life style

(Lou et al., 2011)

Transformer based Architecture for Clinical Aspect Extraction – multi-class classification problem



	Baseline (LSTM)			Our Model		
	Precision	Recall	F1	Precision	Recall	F1
Health	0.67	0.68	0.68	0.78	0.75	0.77
Demography	0.92	0.97	0.94	0.94	0.96	0.95
Treatment	0.69	0.71	0.70	0.71	0.79	0.74
Lab Test	0.79	0.83	0.81	0.82	0.83	0.82
Lifestyle	0.62	0.76	0.69	0.72	0.83	0.77

(Clinical Natural Language Processing, 2020)

Problem-3: Different manifestations of the same criteria- Paraphrase detection

history of traumatic brain or head injury with loss of consciousness
diagnosis of traumatic brain injury
history of tbi
history of known brain metastases
subjects with a history of stroke or traumatic brain injury
volunteers must have history of at least one mild traumatic brain injury

- The ability to provide informed consent before any trial-related activities.
- written informed consent to participate in the study
  Ability and willingness to give written informed consent
- The child and parent or legal guardian is able to provide assent and/or consent.
- Willingness and ability to sign informed consent document
- Provide assent and have a legal guardian that will participate and provide parental permission.
- Able to comprehend and willing to provide written consent

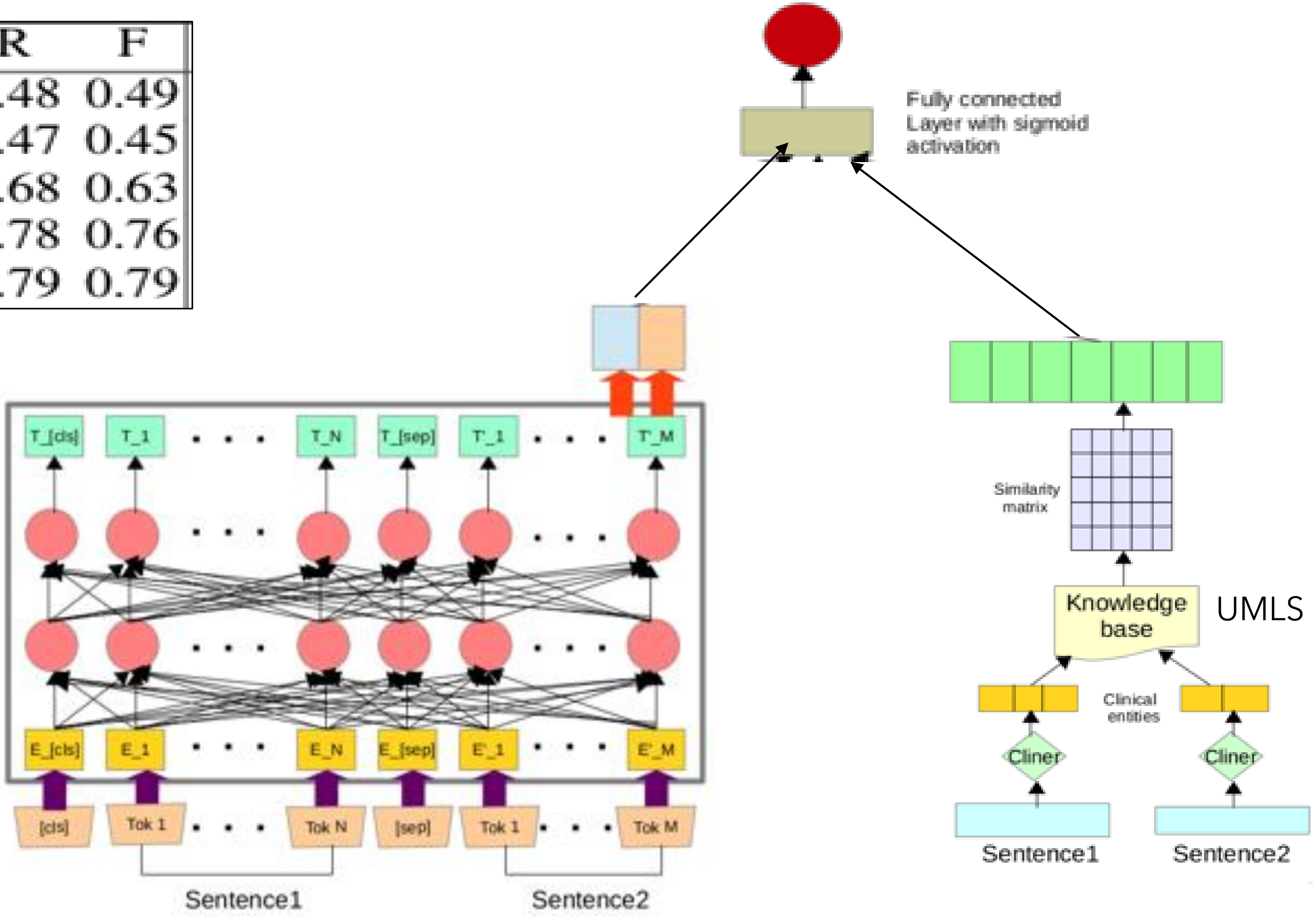
Obese with Hyperinsulinemia can be included in the study

Patients with Type-II Diabetes can be included

Detecting different manifestations of the same criteria – classification – similar / dissimilar

(ICDMAI, 2020)

	P	R	F
BiLSTM Siamese	0.51	0.48	0.49
CNN	0.45	0.47	0.45
C-BiLSTM Siamese	0.60	0.68	0.63
BERT	0.75	0.78	0.76
BERT+KB	0.81	0.79	0.79



Matching Patients to Trials – work in progress

Disease: Colon cancer
Gene: Microsatellite Instability (high);
Demographic: 25-year-old female
Other: Early pregnancy

Disease: Colon cancer
Gene: MLH1 methylation positive
Demographic: 65-year-old male
Other: Hypercholesterolemia, SGOT=73.0, history of hypertension

Involves aspect resolution followed by matching.

Total bilirubin less than or equal to 1.5 mg/dl, except in patients with history of anaemia. Have had their ileostomy or colostomy for at least 3 months. Subjects must be between the ages of 18 to 65 years old, able to provide informed consent and must not intake alcohol. Life expectancy of at least 6 months and willing to provide informed consent. Live vaccine within 4 weeks prior to therapy or potential need for a live vaccine. SGOT and SGPT < 3x the normal, Serum ALAT or serum ASAT > 5 x upper limit of normal (ULN) at screening. Current alcohol abuse or drug addiction that in the opinion of the investigator, Hypertensive patients Pregnant women of stage 3 and age 18 years and older must be excluded.

Use Case 3 – Predicting ICU Length of stay

Using Clinical Notes for ICU stay prediction

- Predicting *length of stay* in ICU helps in better logistics planning – ensures ***better resource usage*** for critically ill patients
- MIMIC Dataset - a publicly available dataset was developed by the Laboratory for Computational Physiology
  - Comprises deidentified health data associated with thousands of intensive care unit admissions
  - The dataset is widely used by investigators and engineers around the world - to drive research in clinical informatics, epidemiology, and machine learning
- During hospitalization a nursing notes contain information about patient's condition, nursing assessments, care provided to a patient.
- Objective
  - Predict length of stay of a patient in Intensive care unit ***at the time of hospital admission*** from their nursing notes of first 24 hours .

	Dataset	Feature used	Method	Best Result
Alghatani et al., 2021	44,000 ICU stays from MIMIC	patient's vital signs like, heart rate, BP, temp., resp. etc	Random Forest	65% accuracy
Su et al., 2021	2224 Sepsis patients PICMISD	Age, $P(v\text{-}a)CO_2$ /C(a-v)O, SO,wbc etc.	XG-Boost model	F1: 0.69, AUC-ROC:0.76
Rocheteau,Liò, et al., 2020	eICUcritical care dataset	medical features, Gender, Age, Ethnicity, etc.	Temporal convolution	Kappa score = 0.58
Harutyunyan et al., 2019	42276 ICU stays of 33798 uniquepatients from mimic database	17 clinical variables like, Capillary refill rate, Diastolicblood pressure etc. from first 24 hours of admission.	LSTM	AUC-ROC : 0.84
van Aken et al., 2021	38013 admission notes from MIMIC III	Created admission notes from discharge summaries	Pretrained CORe +BioBERT	AUC-ROC : 0.72%
Improved upon SOTA	**22789 Nursing Notes from MIMIC III**	**nursing notes + TF-IDF Vector+ SOI scores**	**BlueBERT+LSTM+TF-IDF+SOI**	**Acc: 79%; AUC-ROC:0.87; Kappa: 0.59**

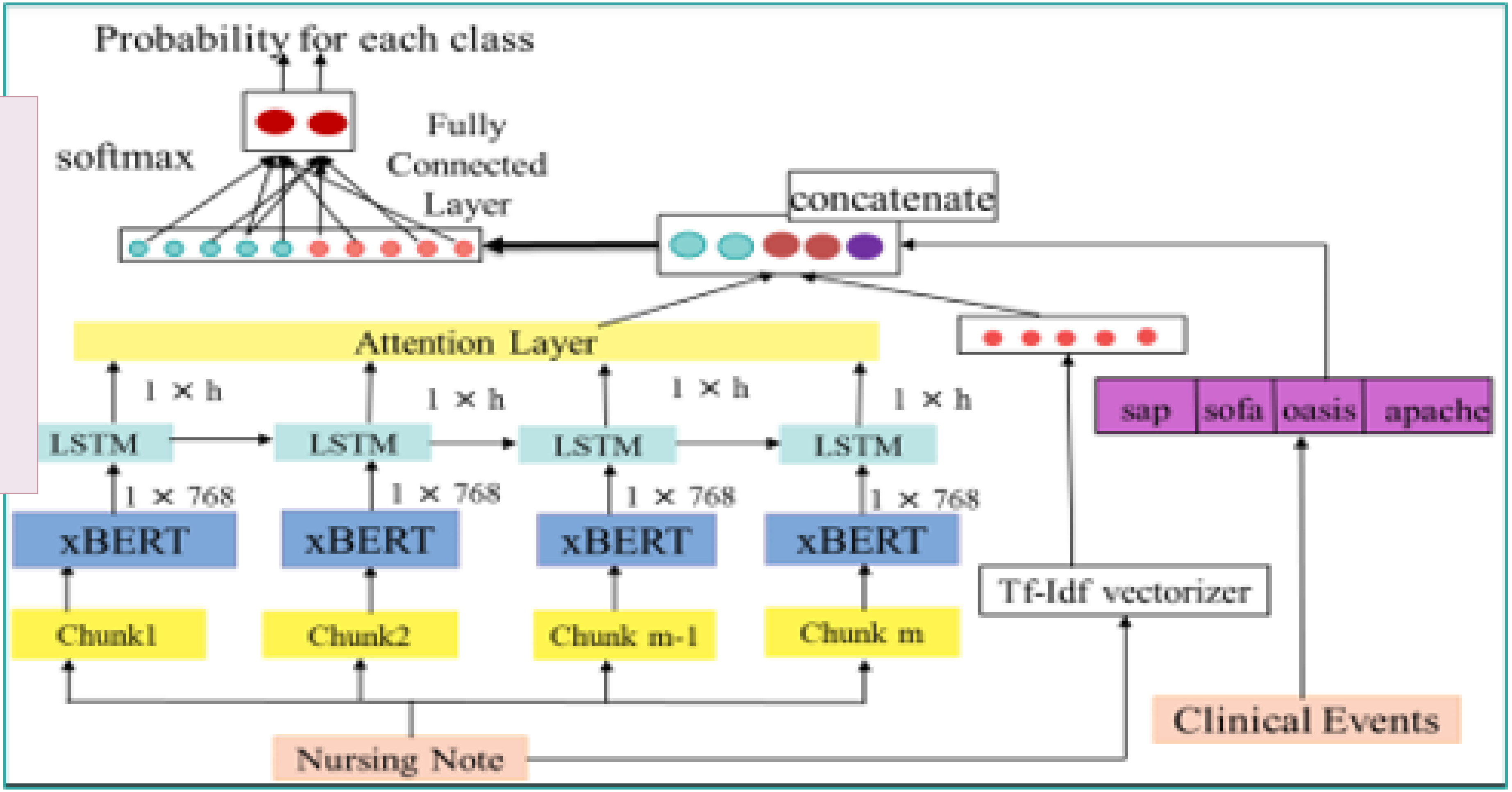
Prediction Architecture

Explainability

- Attempts to understand the model - by perturbing the input of data samples and assess how the predictions change

- Weighted average from multi-head attention models

Long notes



Indicative Phrases

•Phrases like "HR dropping", "requiring mask ventilation for resp. failure", "couldn't breathe" ⮕ indicative of **high ri**sk patients needing longer ICU stays

•"good effect from Ativan", "comfortable breathing", "hemodynamically stable" ⮕ **short ICU stays**.

- Many more problems
  - Trigger alerts
  - More exact predictions
  - Correlations between vital parameters and clinical notes
  - Contra-indications

Summary

- Clinical texts contain a wealth of information about state of patients, healthcare
  - Long Covid
  - Precision Medicine
- Automation possibilities – to reduce risks
  - Physician's Aid – Expert insights
- Explainable, Secure models are needed
- Resource-light models are needed


© Randy Glasbergen
www.glasbergen.com
GLASBERGEN
"Doctor and physician are outdated terms.
I'm your biological tech support specialist."

Thank You

References

- **Perera et al. Named Entity Recognition and Relation Detection for Biomedical Information Extraction -** https://internal-journal.frontiersin.org/articles/10.3389/fcell.2020.00673/full