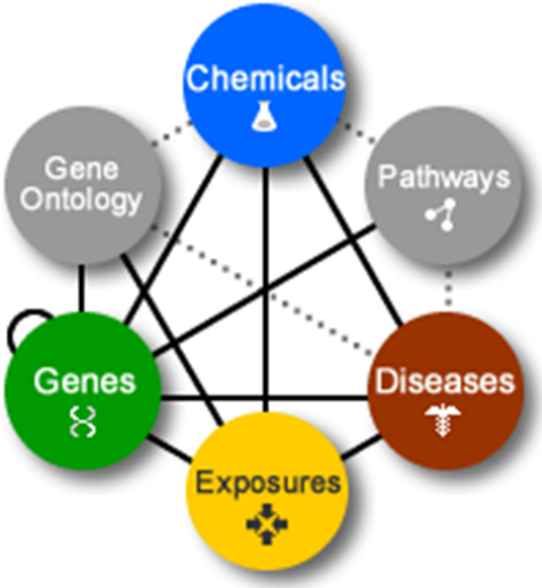
BioCreative V Track 3: Chemical-Disease Relation and Disease Named Entity Recognition and Normalization



Zhiyong Lu, 陆致用
Earl Stadtman Investigator
Head, Biomedical Text Mining Group
National Center for Biotechnology Information

Thomas C. Wiegers
Research Bioinformatician
Comparative Toxicogenomics Database
North Carolina State University

September 10, 2015

Agenda

Introduction and Background

- Introduction to CDR/DNER-based curation
- Explanation of why we chose these particular tasks:
  - Why is CDR/DNER text mining so important?
- Case study of how CDR/DNER curation is used

Task Design

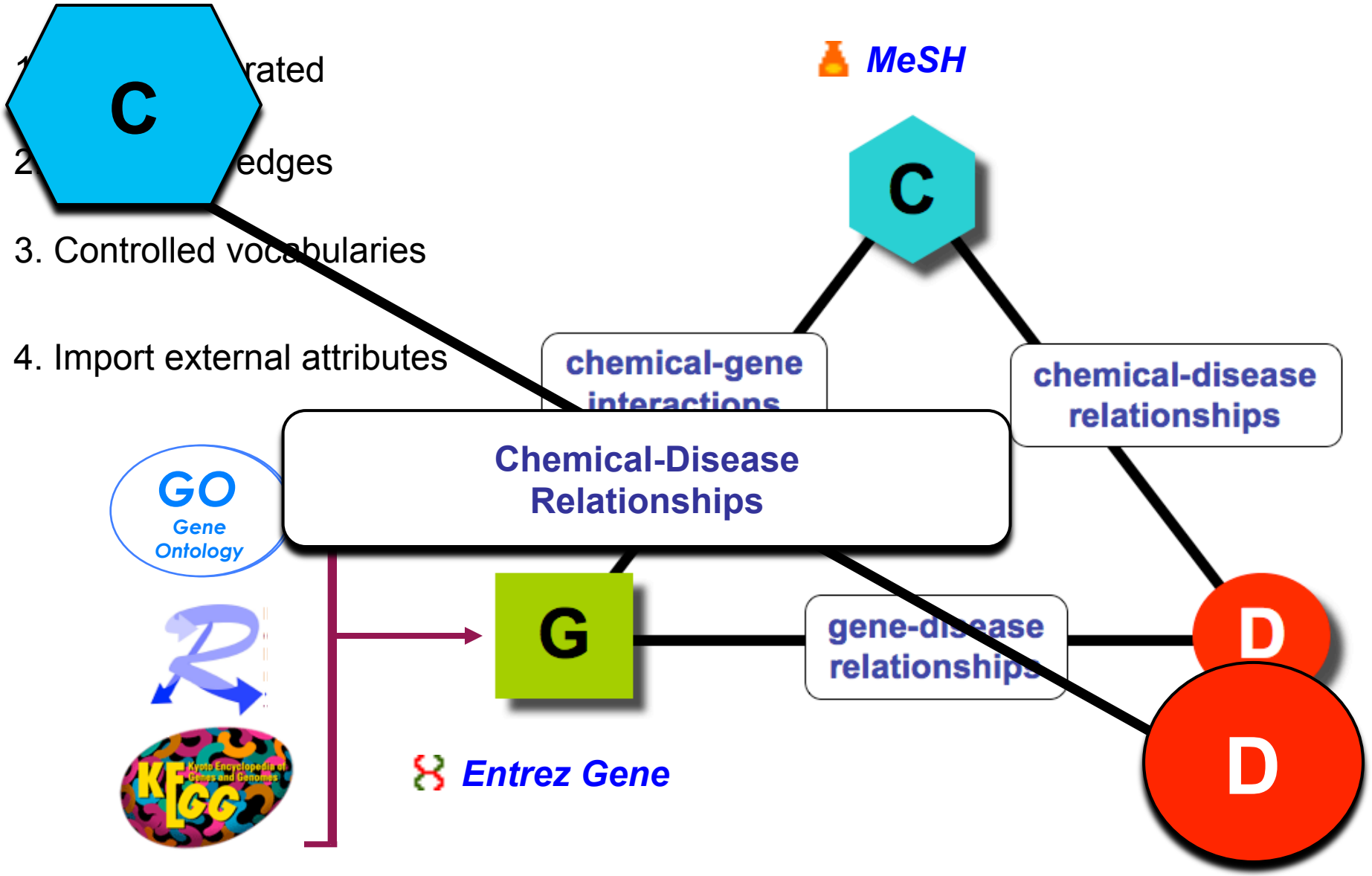
- DNER task examples
- CDR task examples
- Corpora descriptions
- Supporting resources
- Task evaluation

Results

- Challenge results
- Survey results

CTD is a Curated Knowledgebase

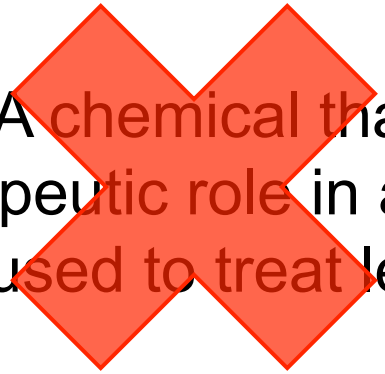
1. [illegible]rated
2. [illegible]edges
3. Controlled vocabularies
4. Import external attributes



How Does CTD Curate Chemical/Disease Relationships (CDRs)?

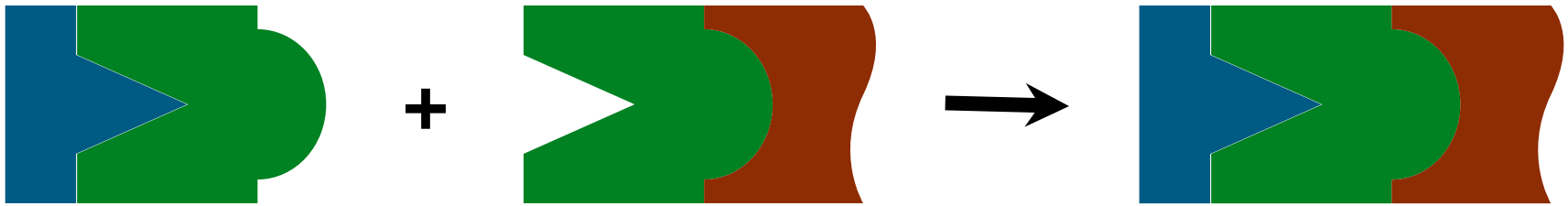
- ***Marker***: A chemical that correlates with a disease (e.g., increased abundance in the brain of chemical X correlates with Alzheimer disease) or may play a role in the etiology of a disease (e.g., exposure to chemical X may play a role in causing lung cancer).

- ***Therapeutic:*** A chemical that has a known or potential therapeutic role in a disease (e.g., chemical X is used to treat leukemia).



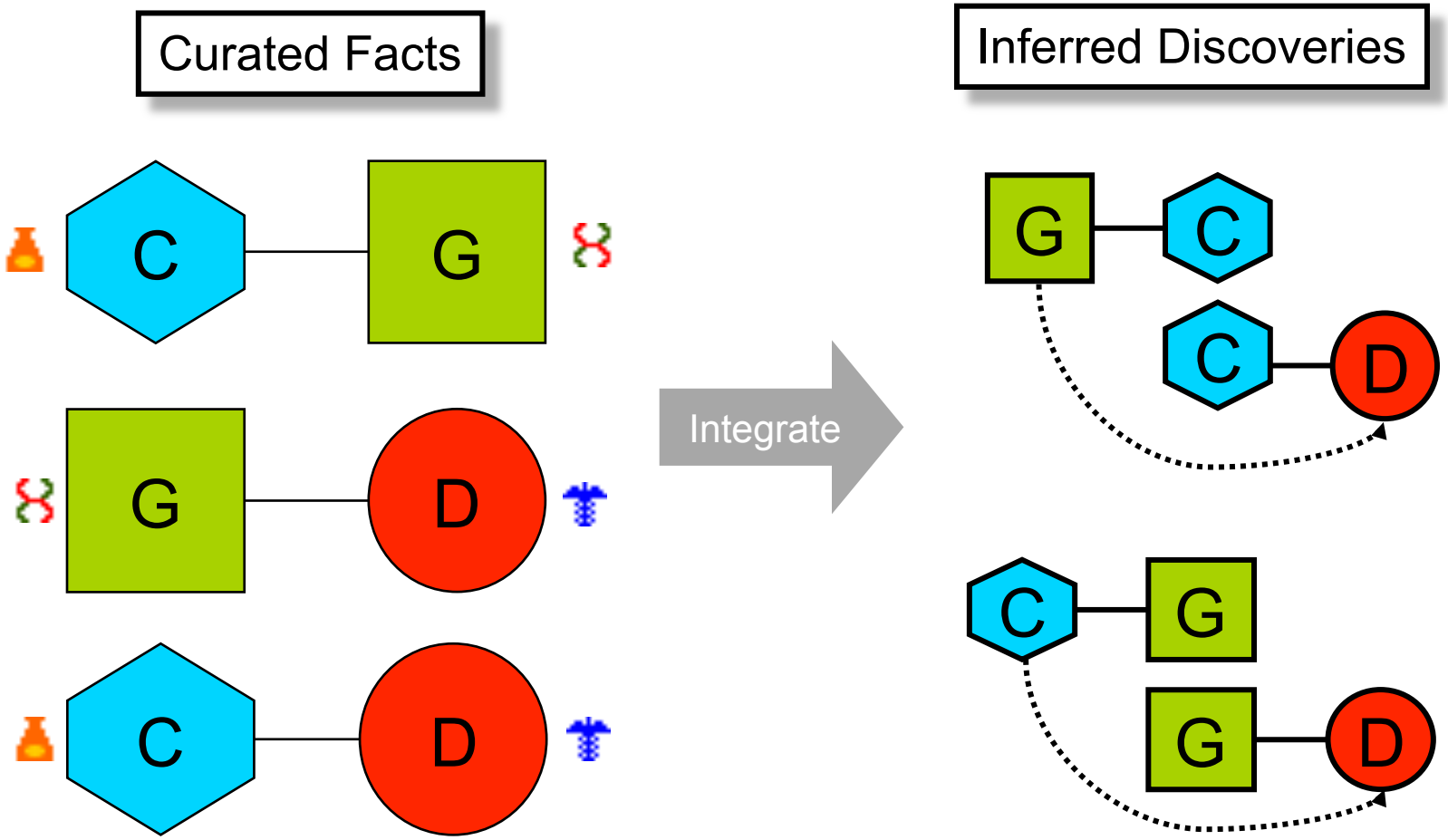
Why are these relationships important?

Transitive Inference
ABC Model of Hypothesis Generation



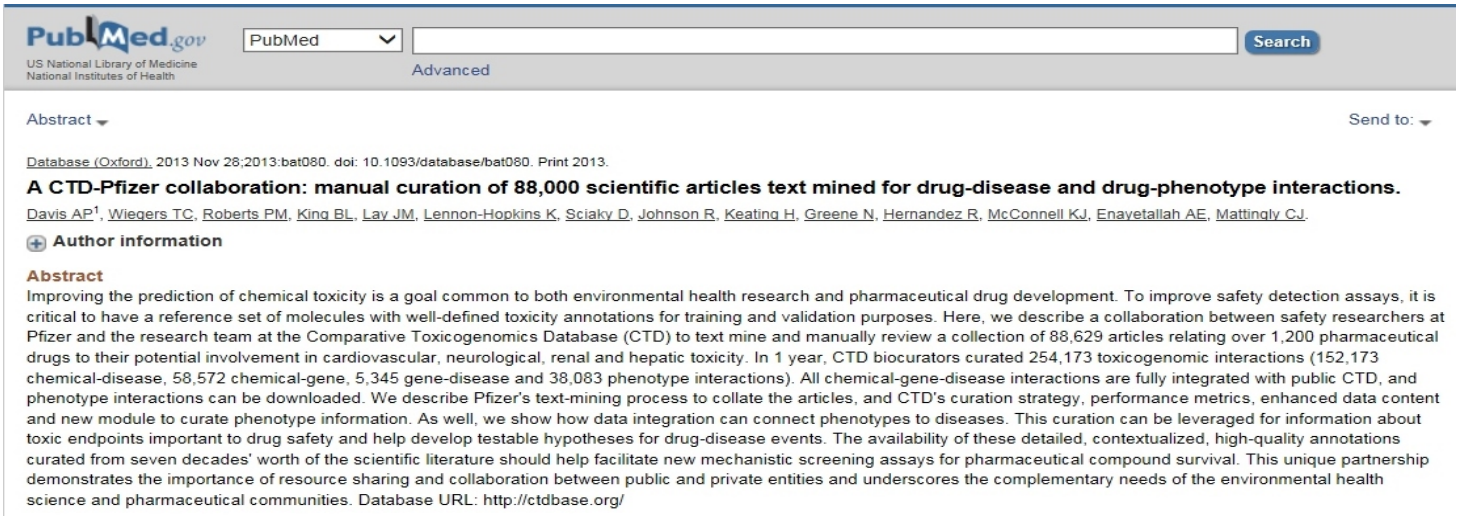
Evans and Rzhetsky, Science, 329:399 (2010)

Transitive Inference at CTD



CTD: Over 27 million Toxicogenomic Relationships

Case Study: CTD/Pfizer Collaboration



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Search | Advanced

Abstract | Send to:

Database (Oxford). 2013 Nov 28;2013:bat080. doi: 10.1093/database/bat080. Print 2013.

A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions.

Davis AP[1], Wiegers TC, Roberts PM, King BL, Lay JM, Lennon-Hopkins K, Sciaky D, Johnson R, Keating H, Greene N, Hernandez R, McConnell KJ, Enayetallah AE, Mattingly CJ.

Author information

Abstract

Improving the prediction of chemical toxicity is a goal common to both environmental health research and pharmaceutical drug development. To improve safety detection assays, it is critical to have a reference set of molecules with well-defined toxicity annotations for training and validation purposes. Here, we describe a collaboration between safety researchers at Pfizer and the research team at the Comparative Toxicogenomics Database (CTD) to text mine and manually review a collection of 88,629 articles relating over 1,200 pharmaceutical drugs to their potential involvement in cardiovascular, neurological, renal and hepatic toxicity. In 1 year, CTD biocurators curated 254,173 toxicogenomic interactions (152,173 chemical-disease, 58,572 chemical-gene, 5,345 gene-disease and 38,083 phenotype interactions). All chemical-gene-disease interactions are fully integrated with public CTD, and phenotype interactions can be downloaded. We describe Pfizer's text-mining process to collate the articles, and CTD's curation strategy, performance metrics, enhanced data content and new module to curate phenotype information. As well, we show how data integration can connect phenotypes to diseases. This curation can be leveraged for information about toxic endpoints important to drug safety and help develop testable hypotheses for drug-disease events. The availability of these detailed, contextualized, high-quality annotations curated from seven decades' worth of the scientific literature should help facilitate new mechanistic screening assays for pharmaceutical compound survival. This unique partnership demonstrates the importance of resource sharing and collaboration between public and private entities and underscores the complementary needs of the environmental health science and pharmaceutical communities. Database URL: http://ctdbase.org/

- Objective: Leverage Published Literature on Toxicity
  - Begin building comprehensive drug-event relationships database
  - Target: Cardiovascular, Neurological, Renal, and Hepatic Toxicity
- CTD Manually Curated 88,629 Articles:
  - 254,173 Toxicogenomic Interactions
  - 152,173 Chemical-Disease
  - 58,572 Chemical-Gene
  - 5,345 Gene-Disease
  - 38,083 Phenotypic

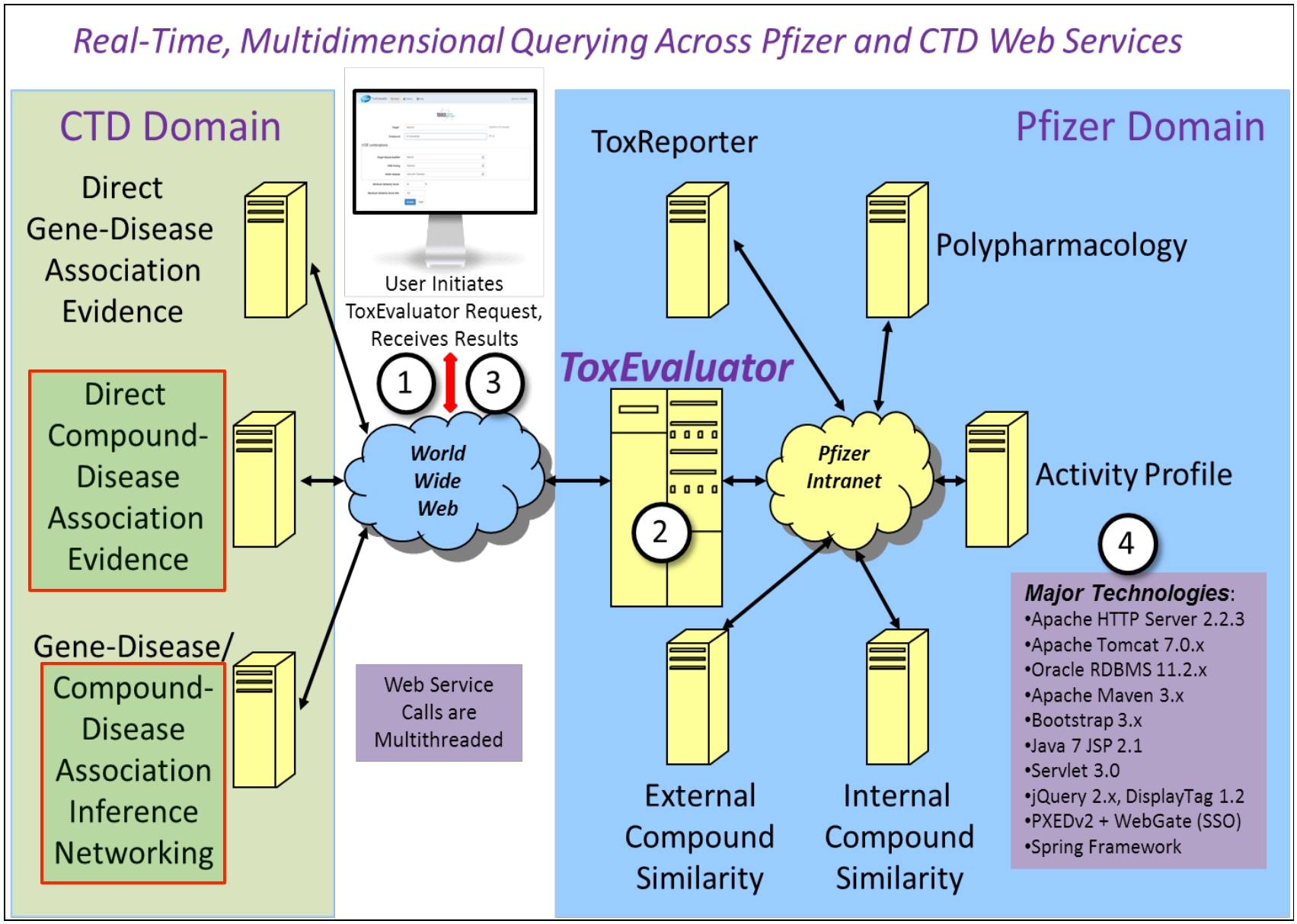
Why? **Example: Investigation of Adverse Study-Related Findings**

Investigation of Adverse Findings Clinical and Pre-Clinical Studies

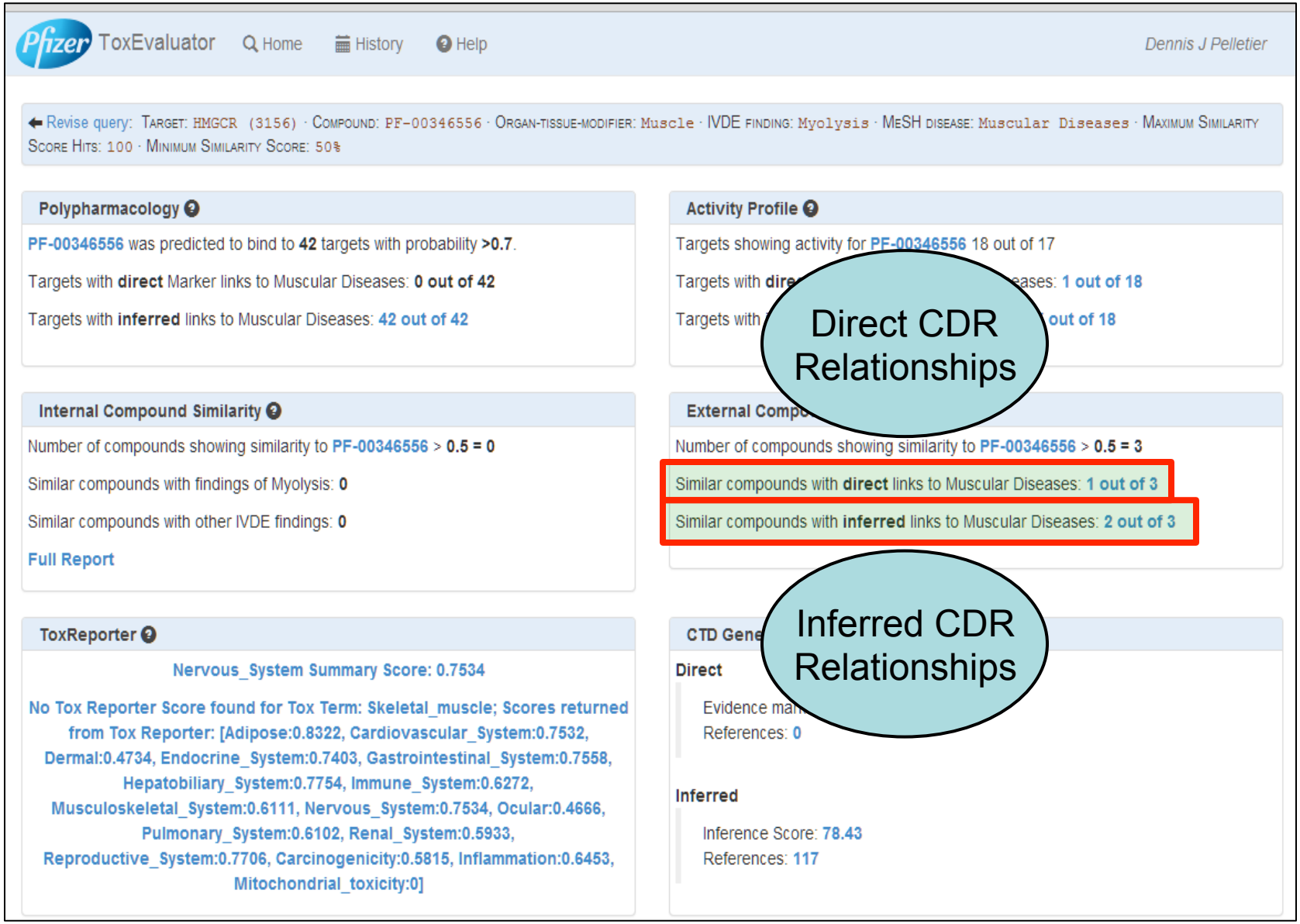
Pathologists and Toxicologists Investigate Adverse Findings in Animal Studies:

- Investigations Typically Begin With:
  - Chemical Structure
  - Intended Biological Target
  - Description of Observed Pathology:
    - Organ
    - Type of lesion
    - Species
- First Question: Is Lesion Related to the Study?
- If So, Is Lesion Related to:
  - The Chemical Structure Itself, i.e., the ***chemotype***
  - Pharmacological Response to Modulation of Target, i.e., ***mechanism***
  - Interaction with Related or Unrelated Biological Systems, i.e., ***off-target***
- Bottom Line: Are there Connections Between the:
  - Chemical Structure
  - Targeted Modulation
  - Observed Pathology

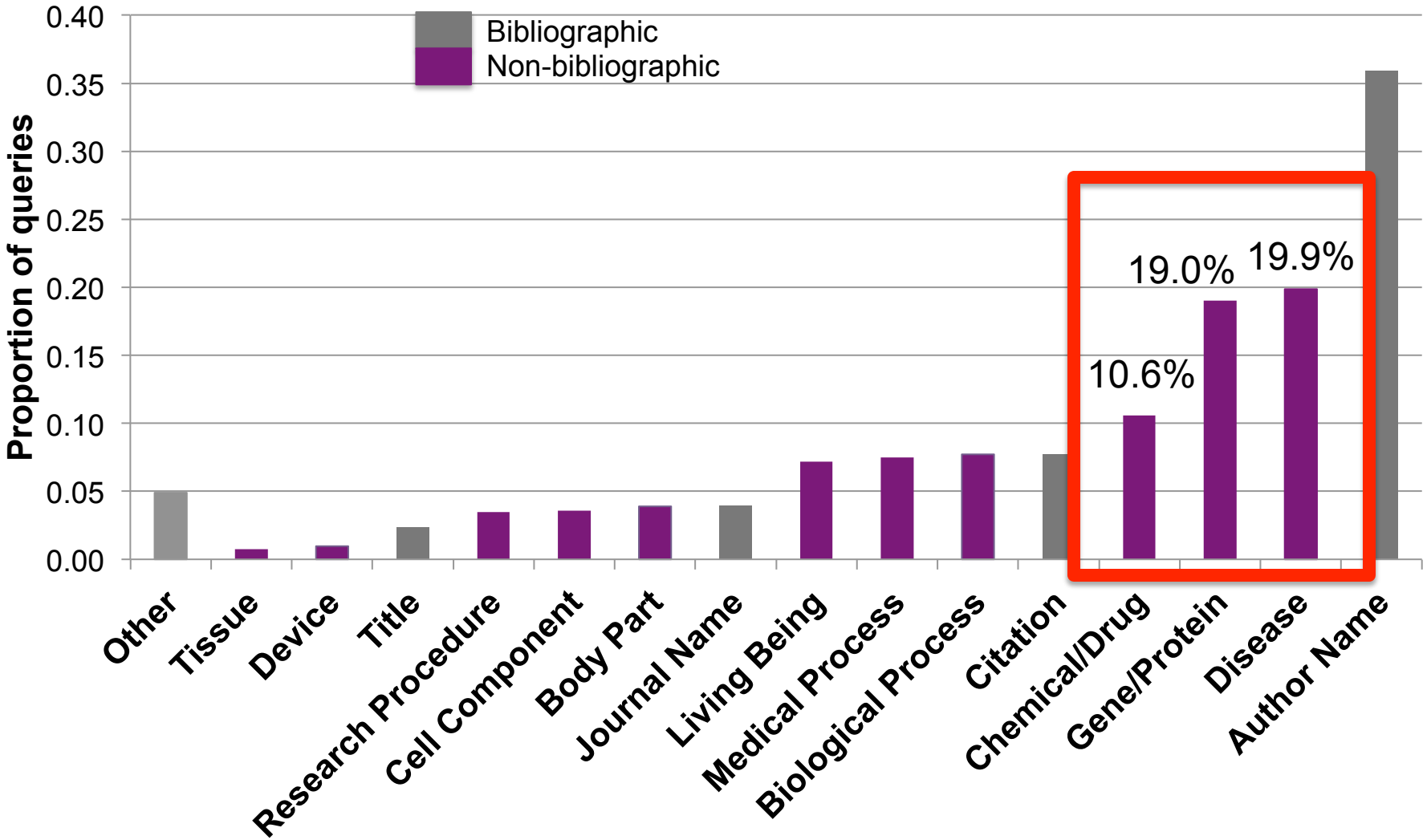
Pfizer's ToxEvaluator



Pfizer's ToxEvaluator



Most Searched Bio-Concepts in PubMed Queries



Dogan et al., **Understanding PubMed user search behavior through log analysis.** Database, 2009

Disease Named Entity Recognition (DNER) Task

Clin Oncol (R Coll Radiol). 1993;5(6):367-71.

Liposomal daunorubicin in advanced Kaposi's sarcoma: a phase II study.

Money-Kyrle JF[1], Bates F, Ready J, Gazzard BG, Phillips RH, Boag FC.

Author information

MeSH: D012514
Sarcoma, Kaposi

Abstract

We report a non-randomized Phase II clinical trial to assess the efficacy and safety of liposomal daunorubicin (DaunoXome) in the treatment of AIDS related Kaposi's sarcoma. Eleven homosexual men with advanced Kaposi's sarcoma were entered in the trial. Changes in size, colour and associated oedema of selected 'target' lesions were measured. Clinical, biochemical and haematological toxicities were assessed. Ten subjects were evaluated. A partial response was achieved in four, of whom two subsequently relapsed. Stabilization of Kaposi's sarcoma occurred in the remaining six, maintained until the end of the trial period in four. The drug was generally well tolerated, with few mild symptoms of toxicity. The main problem encountered was haematological toxicity, with three subjects experiencing severe neutropenia (neutrophil count < 0.5 x 10(9)/l). There was no evidence of cardiotoxicity. In this small patient sample, liposomal daunorubicin was an effective and well tolerated agent in the treatment of Kaposi's sarcoma.

PMID: 8305357 [PubMed - indexed for MEDLINE]

- Input: Raw PubMed Abstracts
- Output: Disease Mentions Normalized to MeSH IDs

Chemical-induced Diseases (CID) Relation Extraction Task

Clin Oncol (R Coll Radiol). 1993;5(6):367-71.

Liposomal daunorubicin in advanced Kaposi's sarcoma: a phase II study.

Money-Kyrle JF[1], Bates F, Ready J, Gazzard BG, Phillips RH, Boag FC.

Author information

Abstract

We report a non-randomized Phase II clinical trial to assess the efficacy and safety of liposomal daunorubicin (DaunoXome) in the treatment of AIDS related Kaposi's sarcoma. Eleven homosexual men with advanced Kaposi's sarcoma were entered in the trial. Changes in size, colour and associated oedema of selected 'target' lesions were measured. Clinical, biochemical and haematological toxicities were assessed. Ten subjects were evaluated. A partial response was achieved in four, of whom two subsequently relapsed. Stabilization of Kaposi's sarcoma occurred in the remaining six, maintained until the end of the trial period in four. The drug was generally well tolerated, with few mild symptoms of toxicity. The main problem encountered was haematological toxicity, with three subjects experiencing severe neutropenia (neutrophil count < 0.5 x 10(9)/l). There was no evidence of cardiotoxicity. In this small patient sample, liposomal daunorubicin was an effective and well tolerated agent in the treatment of Kaposi's sarcoma.

PMID: 8305357 [PubMed - indexed for MEDLINE]

- Input: Raw PubMed Abstract
- Output: Chemical-Disease Pairs

Relation name	Relation type	Bio-entities
CID	Chemical_Disease	D003630(daunorubicin) | D009503(neutropenia)

Challenges in Relation Extraction

Clin Oncol (R Coll Radiol). 1993;5(6):367-71.

Liposomal daunorubicin in advanced Kaposi's sarcoma: a phase II study.

Money-Kyrle JF[1], Bates F, Ready J, Gazzard BG, Phillips RH, Boag FC.

Author information

Abstract

We report a non-randomized Phase II clinical trial to assess the efficacy and safety of liposomal daunorubicin (DaunoXome) in the treatment of AIDS related Kaposi's sarcoma. Eleven homosexual men with advanced Kaposi's sarcoma were entered in the trial. Changes in size, colour and associated oedema of selected 'target' lesions were measured. Clinical, biochemical and haematological toxicities were assessed. Ten subjects were evaluated. A partial response was achieved in four, of whom two subsequently relapsed. Stabilization of Kaposi's sarcoma occurred in the remaining six, maintained until the end of the trial period in four. The drug was generally well tolerated, with few mild symptoms of toxicity. The main problem encountered was haematological toxicity, with three subjects experiencing severe neutropenia (neutrophil count < 0.5 x 10(9)/l). There was no evidence of cardiotoxicity. In this small patient sample, liposomal daunorubicin was an effective and well tolerated agent in the treatment of Kaposi's sarcoma.

PMID: 8305357 [PubMed - indexed for MEDLINE]

Not relevant; Therapeutic Relationship

Relation Across Sentence Boundaries

Negation

Previous Work

- Related Resources
  - Corpora: NCBI Disease Corpus; CHEMDNER; ADE; EU-ADR
  - Ontologies: MeSH; OMIM; MEDIC; UMLS; Disease Ontology
  - Tools: DNorm; MetaMap; Peregrine

- Previous Challenges:
  - BioCreative 2012 and IV CTD Tasks (2012, 13)
  - ShARe/ CLEF eHealth Shared Task (2013-)
  - PPI, DDI

Task Corpora

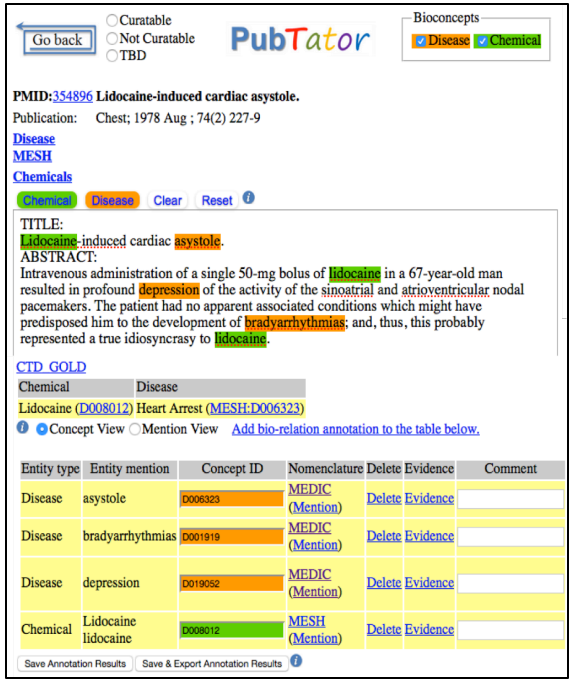
- Selected from CTD-Pfizer Set
  - Training Set : 500 Articles
  - Development Set: 500 Articles
  - Test Set : 500 Articles

- For Each Article, Annotated Data Includes:
  - Disease Named-Entity Annotations
  - Chemical Named-Entity Annotations
  - Relation Annotations Via CTD Curation

- Data Formats:
  - BioC XML
  - PubTator Text

Chemical and Disease Annotations

- Task: Mark Up Chemicals & Diseases
  - Text spans
  - Normalized MeSH ids
  - All 1,500 articles
- Annotators
  - MeSH Indexers
  - Double Annotation performed
- Annotation Guidelines
  - What TO and NOT to Annotate
  - Largely Consistent with Previous Guidelines…
  - …i.e., CHEMDNER and NCBI Disease Corpus
- PubTator-Assisted Annotation
  - For Improved Productivity
  - Used Pre-Annotations by DNorm & tmChem



Wei et al., **PubTator: a Web-based text-mining tool for assisting biocuration**, Nucleic Acids Research, 2013

Relation Annotation Overview

- Leveraged CTD-Pfizer Curation…Except:
  - Removed Relations Involving Entities Not in Abstracts
  - Removed General Relations, e.g. Adverse Drug Event
  - Updated Some Relations Due to MeSH Update

- Performed new annotations for 100 test articles
  - Selected Using NCBI Similarity Indexing
  - Similar C/D/CID to Training and Development Sets
  - Done by CTD Curators
  - Used Normal CTD Curation Protocol
  - Not Released to Public until Testing Complete

Annotation Results

- Corpus Overall Statistics

Data Set	# of articles	Chemical		Disease		CID relation
		Mention	ID	Mention	ID	
Training Set	500	5203	1467	4182	1946	1039
Development Set	500	5347	1507	4244	1838	1012
Test Set	500	5385	1435	4424	1988	1066

- Inter-Annotator Agreements in Jaccard Index

Data Set	IAA - Diseases	IAAs - Chemicals
Training Set	0.8600	0.9523
Development Set	0.8742	0.9577
Test Set	**0.8875**	0.9630
All Sets	0.8747	0.9605

Task Supporting Resources

- BioC Library (bioc.sourceforge.net)
- Data Visualization & Comparison in PubTator
- Freely Available NER Taggers
  - Disease: DNorm
    - 1st in ShARe/CLEF Disease Normalization Task (2013)
    - 80.9% in F_1 Score
  - Chemical: tmChem
    - 1st in BioCreative IV CHEMDNER-CEM Task (2013)
    - 88.3% in F_1 Score
  - Customized for CDR-task
    - Trained models with CDR data sets
    - Folder-pooling mechanism for instant response
- Evaluation Kit
- FAQs

Leaman et al., **DNorm: disease name normalization with pairwise learning to rank.** Bioinformatics, 2013
Leaman et al., **tmChem: a high performance approach for chemical named entity recognition and normalization.** Cheminformatics, 2015

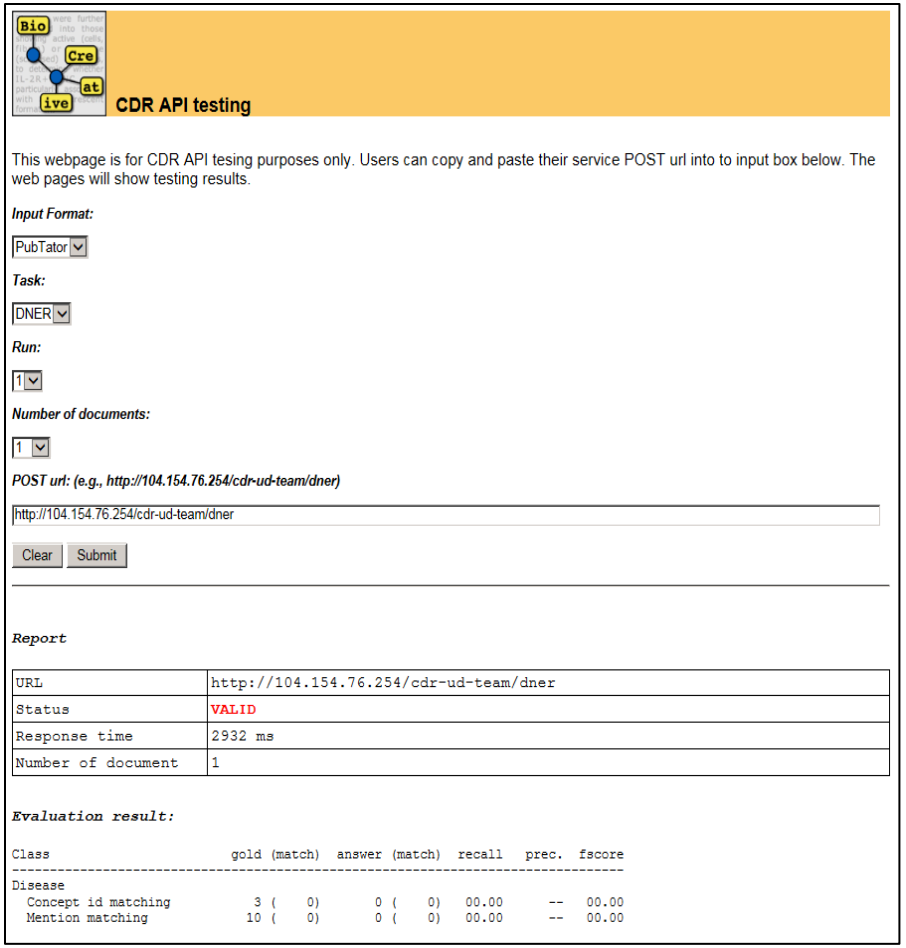
CDR Task Evaluation

- How? Through Web Services:
  - Representational State Transfer (RESTful) APIs
    - Client Sends Text Passages…
    - …Using Simple HTTP Calls
    - Web Service Annotates Text Passages
    - Client Receives Automatic Mark-Ups
  - Requested Responses Within 30 Seconds
- Why Web Services?
  - Quick Access to Text-Mining Services
  - Simpler Text-Mining Pipeline Integration
  - Potential Improvement of System Scalability

RESTful Web Service Support

1. RESTful Set-up Support:
   - Executable programs
   - Step-By-Step Instructions
   - Or Create Your Own

2. API Testing
   - Results Format Validation
   - Response Time



CDR API testing

This webpage is for CDR API tesing purposes only. Users can copy and paste their service POST url into to input box below. The web pages will show testing results.

Input Format:

PubTator

Task:

DNER

Run:

1

Number of documents:

1

POST url: (e.g., http://104.154.76.254/cdr-ud-team/dner)

http://104.154.76.254/cdr-ud-team/dner

Clear Submit

Report

URL	http://104.154.76.254/cdr-ud-team/dner
Status	VALID
Response time	2932 ms
Number of document	1

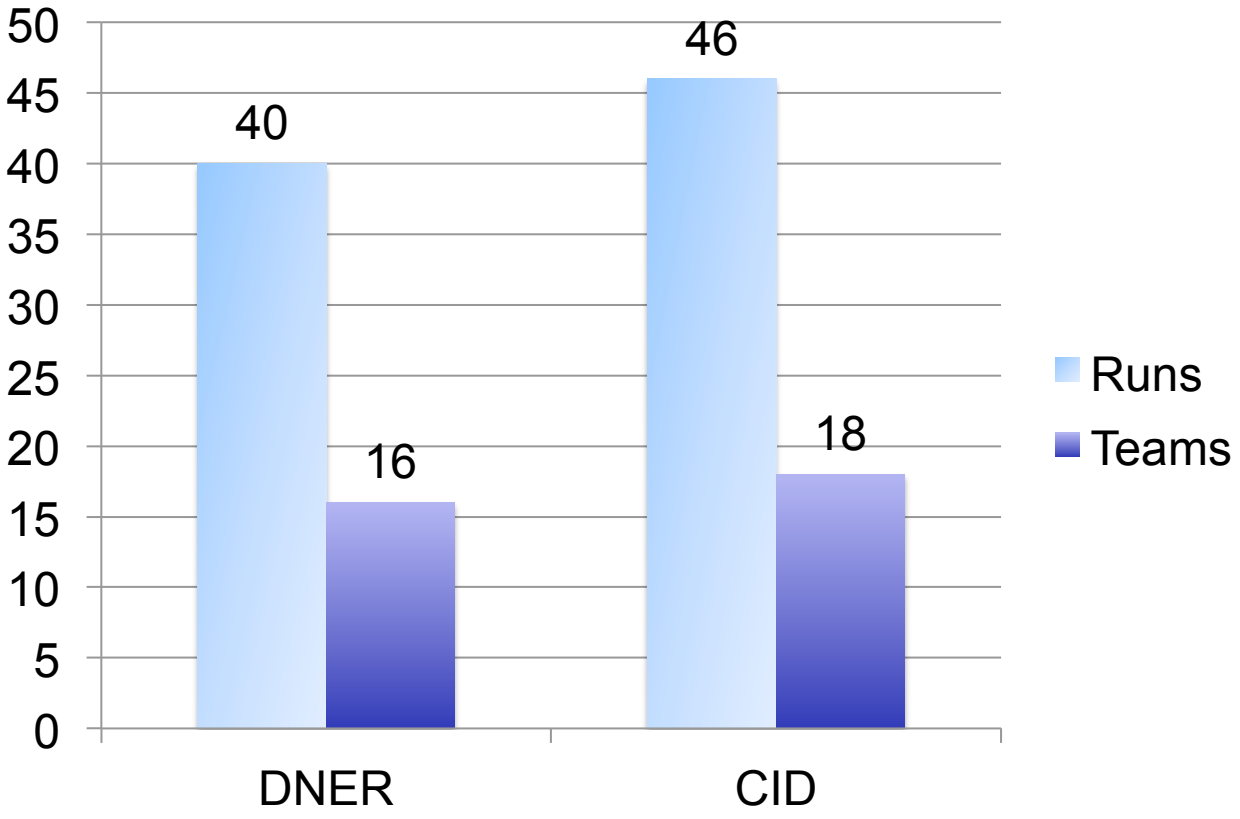
Evaluation result:

```
Class                       gold (match)  answer (match)  recall   prec.  fscore
-------------------------------------------------------------------------------
Disease
  Concept id matching          3 (    0)       0 (    0)   00.00     --    00.00
  Mention matching            10 (    0)       0 (    0)   00.00     --    00.00
```

Baseline & Comparison Systems

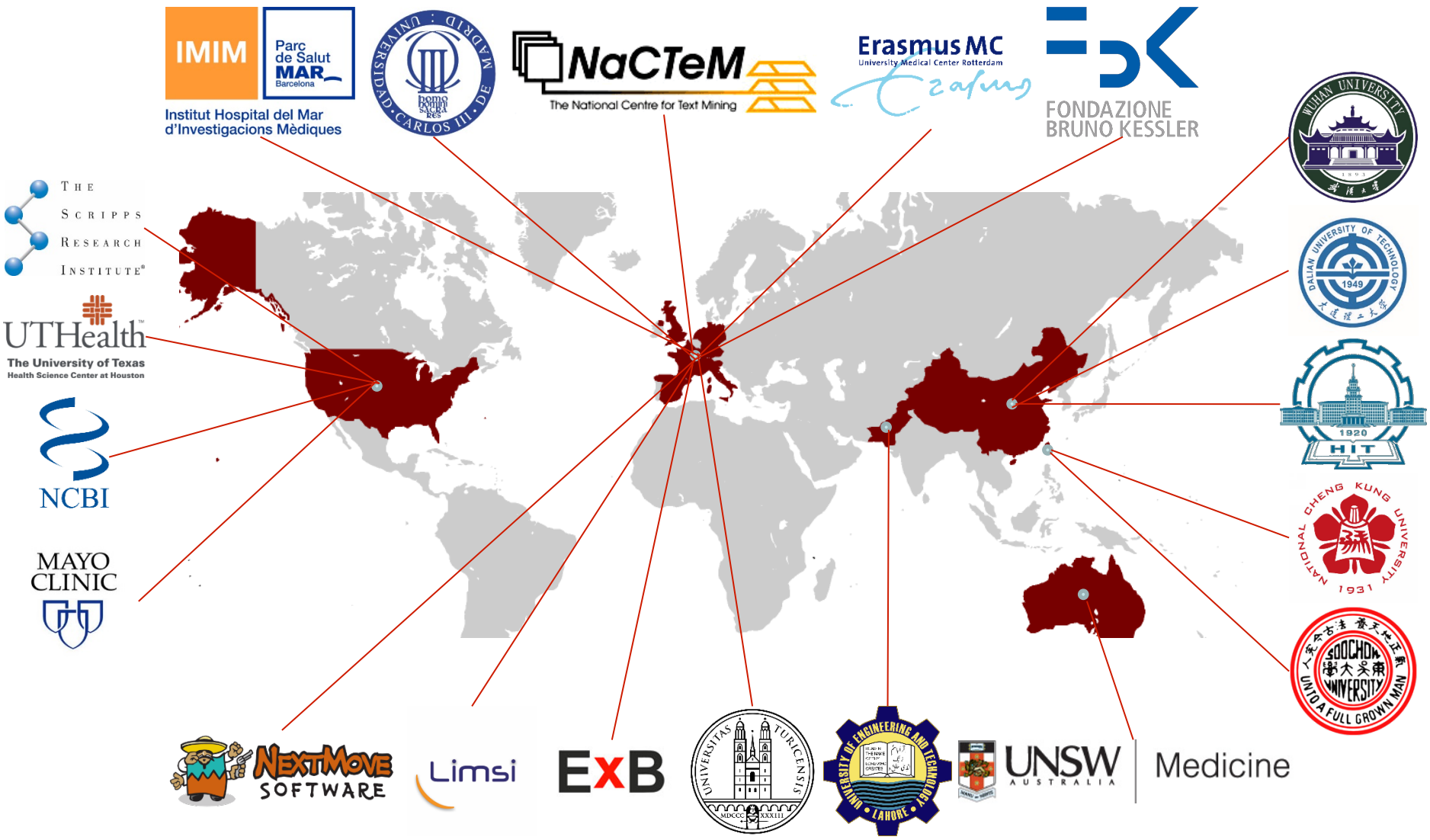
- DNER Task:
  - Dictionary Look-Up - Baseline
  - Out-Of-Box DNorm - Machine Learning
    - Used CDR Data with Default Settings
- CID Task:
  - Co-Occurrence Method
    - Used DNorm and tmChem
    - Two Versions:
      - Abstract-Level
      - Sentence-Level

Participating Teams
Up to 3 Runs Per Team



In addition, a CID team submitted two manual runs via crowdsourcing

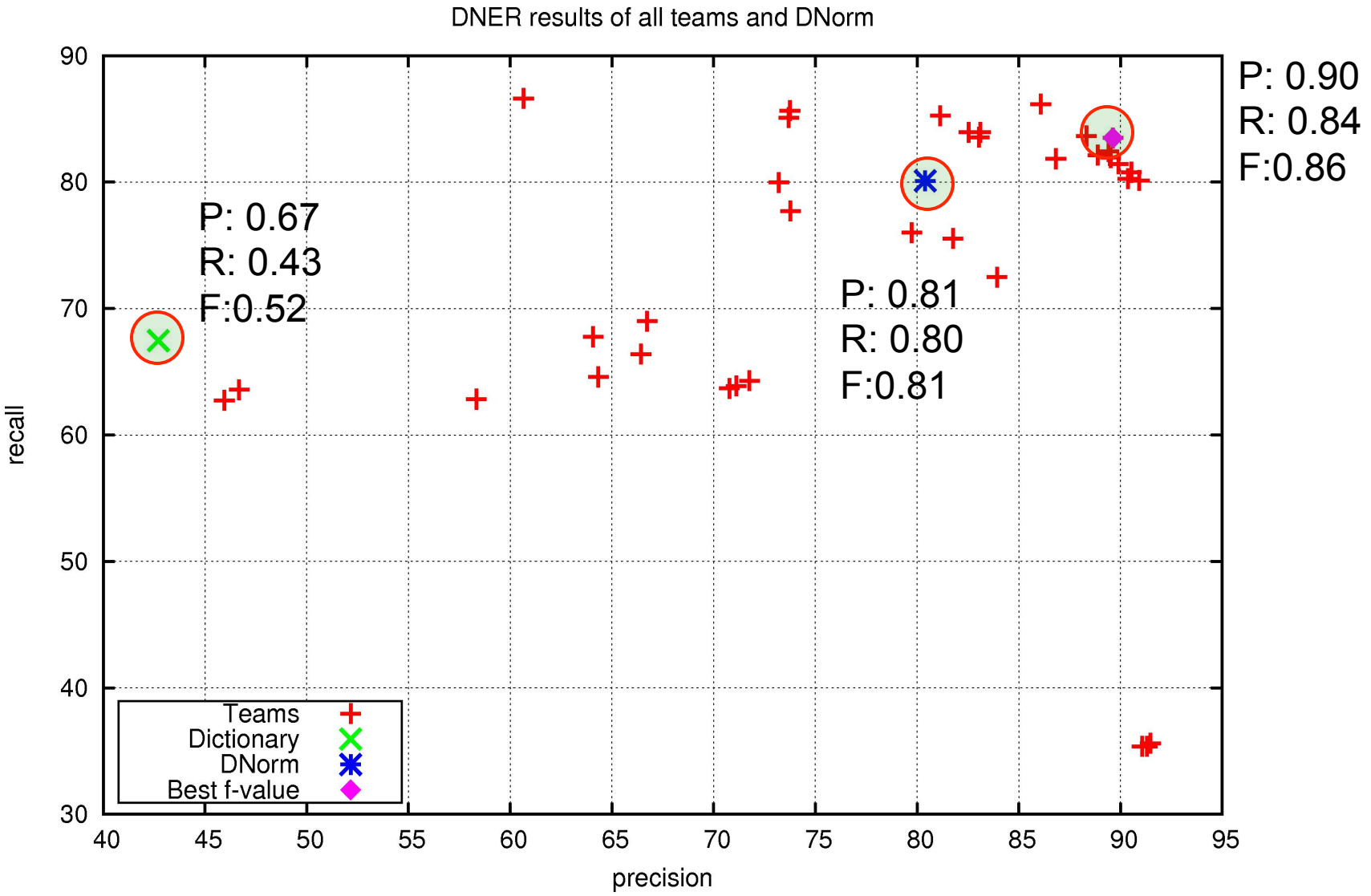
Teams: 12 Countries/4 Continents



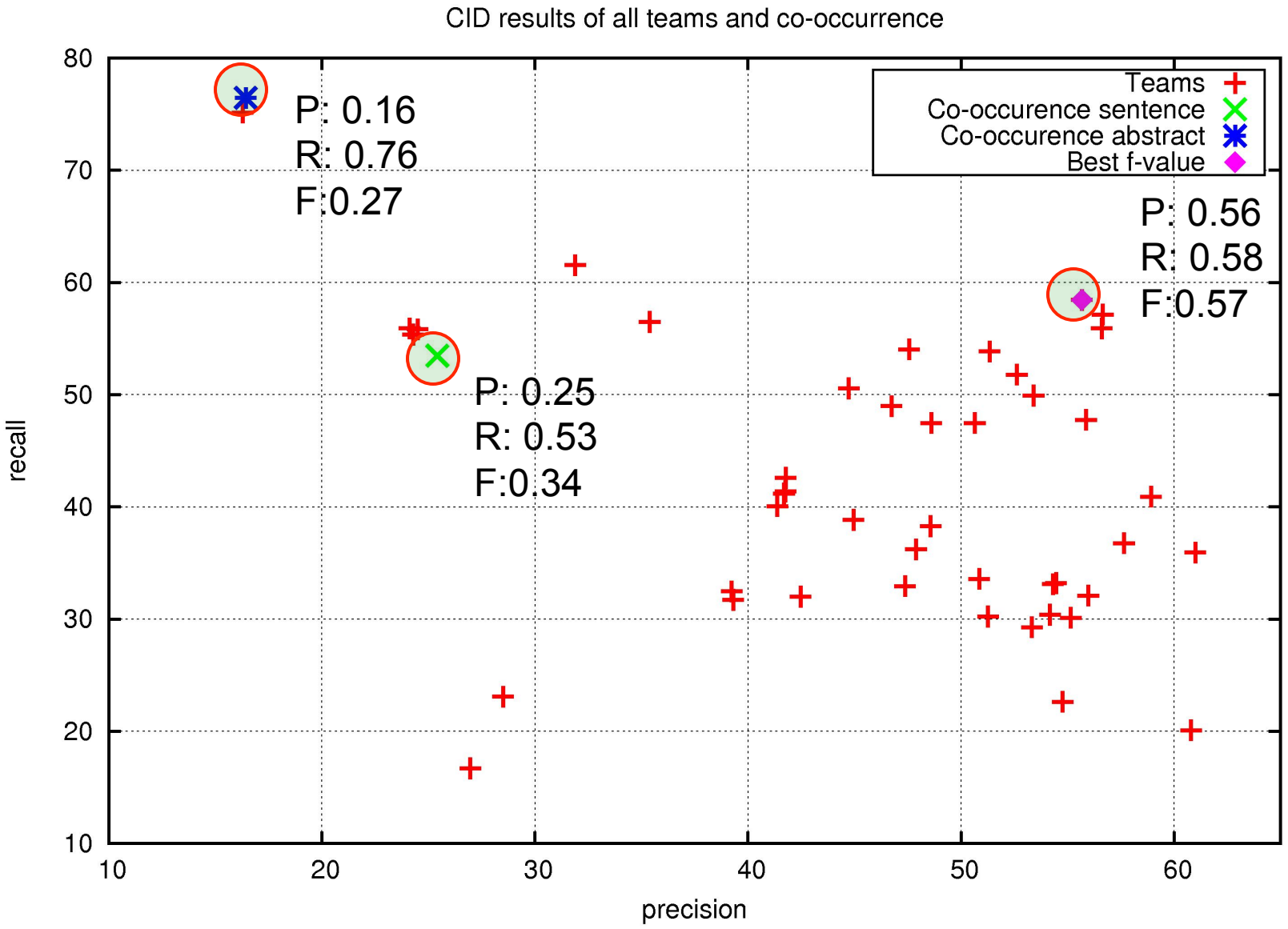
Evaluation Metrics

- DNER Task
  - Data Points: Document ID, Disease Concept ID
  - Precision, Recall, F-score
  - Mention-Level Accuracy Computed
- CID Task
  - Data points:
    - Document ID
    - Disease Concept ID
    - Chemical Concept ID
  - Precision, Recall, F-score
- Refer to Task Overview Paper

DNER Results:16 Teams, 40 Runs

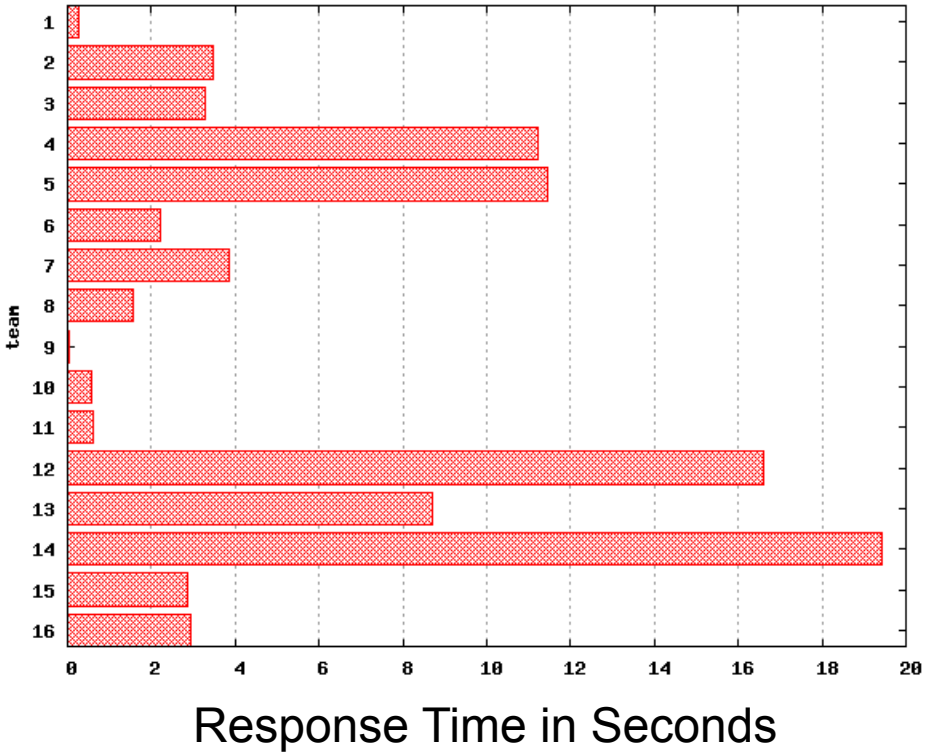


CID Tesults: 18 Teams, 46 Runs



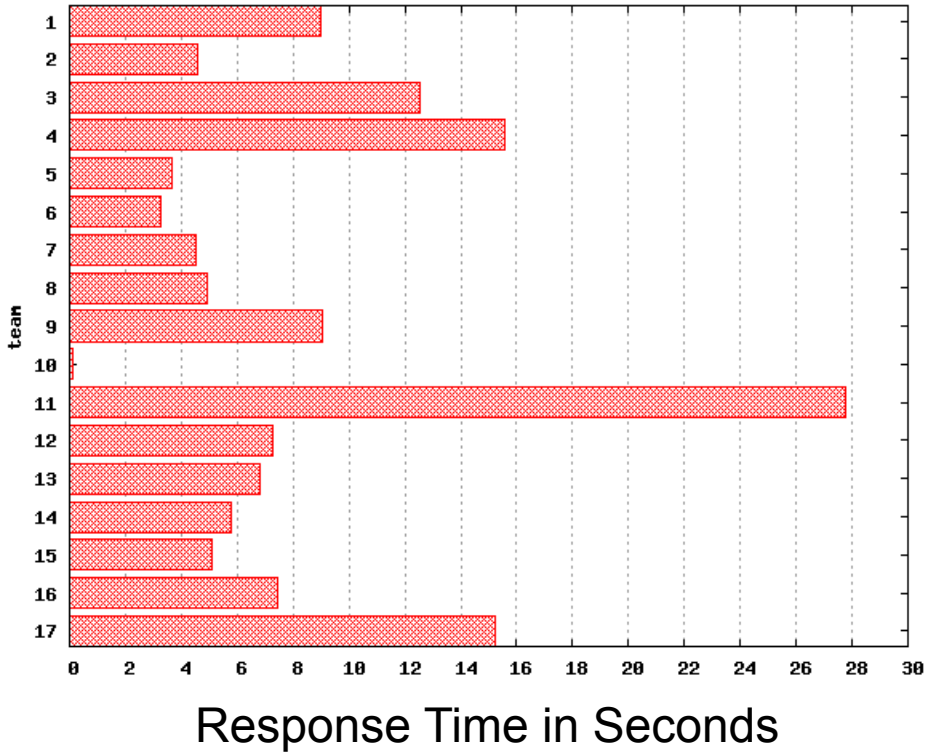
System Response Time by Group

DNER Response Time – Best Run



Response Time in Seconds

CID Response Time – Best Run



Response Time in Seconds

Post-Challenge Survey

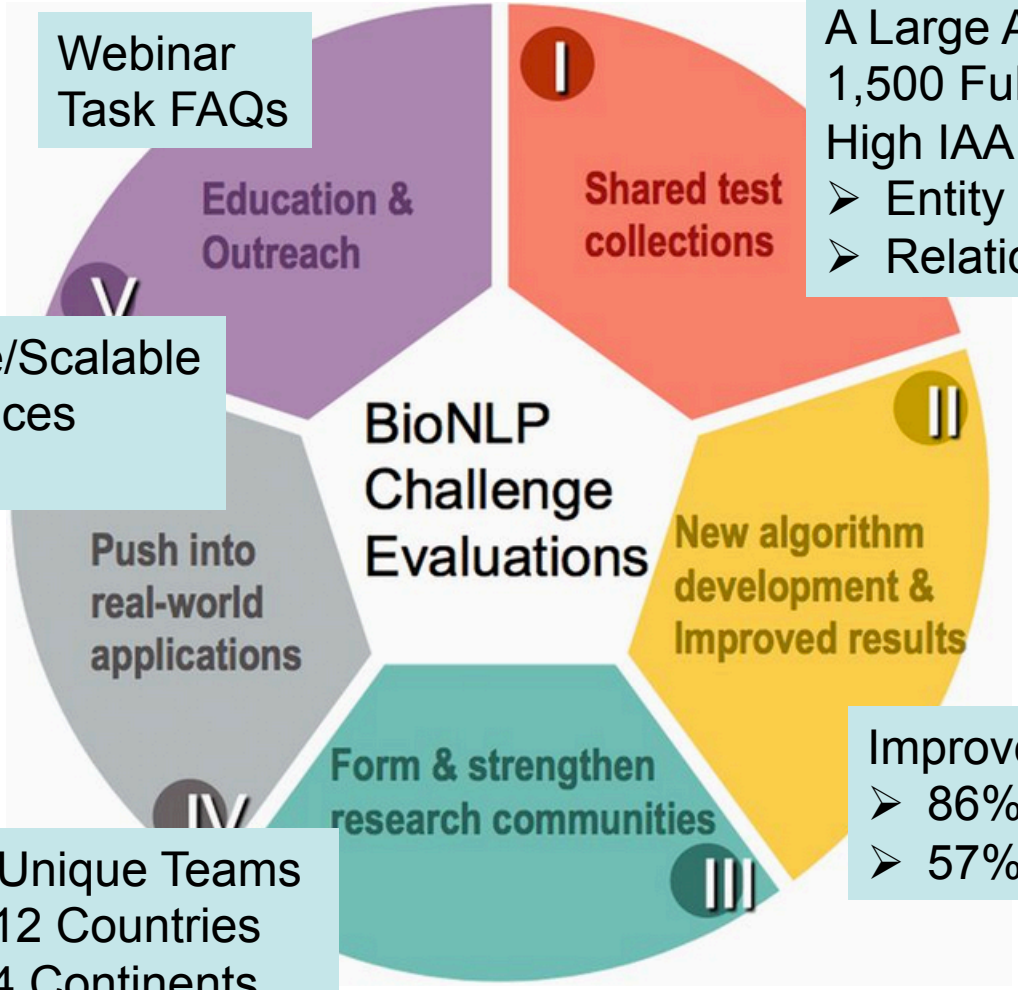
- Use Task Supporting Tools?
- Use Additional Lexicon Resources?
- Use Machine Learning? Features?
- Use Existing BioNLP Tools?
- What General Software Tools?
- What Techniques?
- What Programming Language(s)?
- Pre/Post-Processing?

Post-challenge Survey Results

Supporting Resources	DNER (16 teams)	CID (18 teams)
Web-service setup code	13	14
Evaluation scripts	11	10
DNorm	5	12
tmChem	0	13
PubTator (for data visualization & comparison)	7	4
BioC Library	3	4

Survey of Team Methods

Main Contributions



Webinar
Task FAQs

A Large Annotated Corpus
1,500 Fully Annotated Articles
High IAA scores (~90%)
- Entity Recognition/Normalization
- Relation Extraction Task

Interoperable/Scalable
- Web Services
- BioC

Improved State-of-the-Art
- 86% in DNER
- 57% in CID

25 Unique Teams
- 12 Countries
- 4 Continents

Chung-Chi Huang, and Zhiyong Lu **Community challenges in biomedical text mining over 10 years: success, failure and the future**. Briefings in Bioinformatics; 2015

Limitations

- Relation extraction is difficult!

- Other known issues:
  - Annotation is Imperfect (Human Errors)
  - No New IAAs for Relation Annotations (High Cost)
    - Past CTD IAA Chemical-Gene Curation F1: 77%
  - Difficulty Running Web Services (in China)

What's Next?

- 6 team presentations
  - Performance; Methods; Availability/Scalability
- Release CDR Test Set to Teams
- Invite Teams for Special Issue
- Propose a Meta Web Service for CIDs
  - Merging team results can improve performance
  - Make the results of this challenge more useful for the biocuration community and beyond

Acknowledgements

ncbi.nlm.nih.gov

ctdbase.org

NCBI Team:

Zhiyong Lu
Robert Leaman
Yifan Peng
Chih-Hsuan Wei

CTD/NC State Team:

Carolyn J. Mattingly
Allan Peter Davis
Thomas C. Wiegers
Robin Johnson
Daniela Sciaky

CAM Team:

Jiao Li
Yueping Sun
MeSH Indexers

Funding:

National Center for Biotechnology Information

NCBI

NC STATE UNIVERSITY

NIEHS
National Institute of Environmental Health Sciences

Participants:
Thank You!