華南師範大學 SOUTH CHINA NORMAL UNIVERSITY

文本分析与挖掘 Text Analytics and Mining Lab

iHAVC 智慧健康与可视化计算 Intelligent Health and Visual Computing

Large-scale Clinical Trial Data Mining through Natural Language Processing

Tianyong HAO

School of CS & School of AI

Big Data Center & Text Analytics and Mining Lab

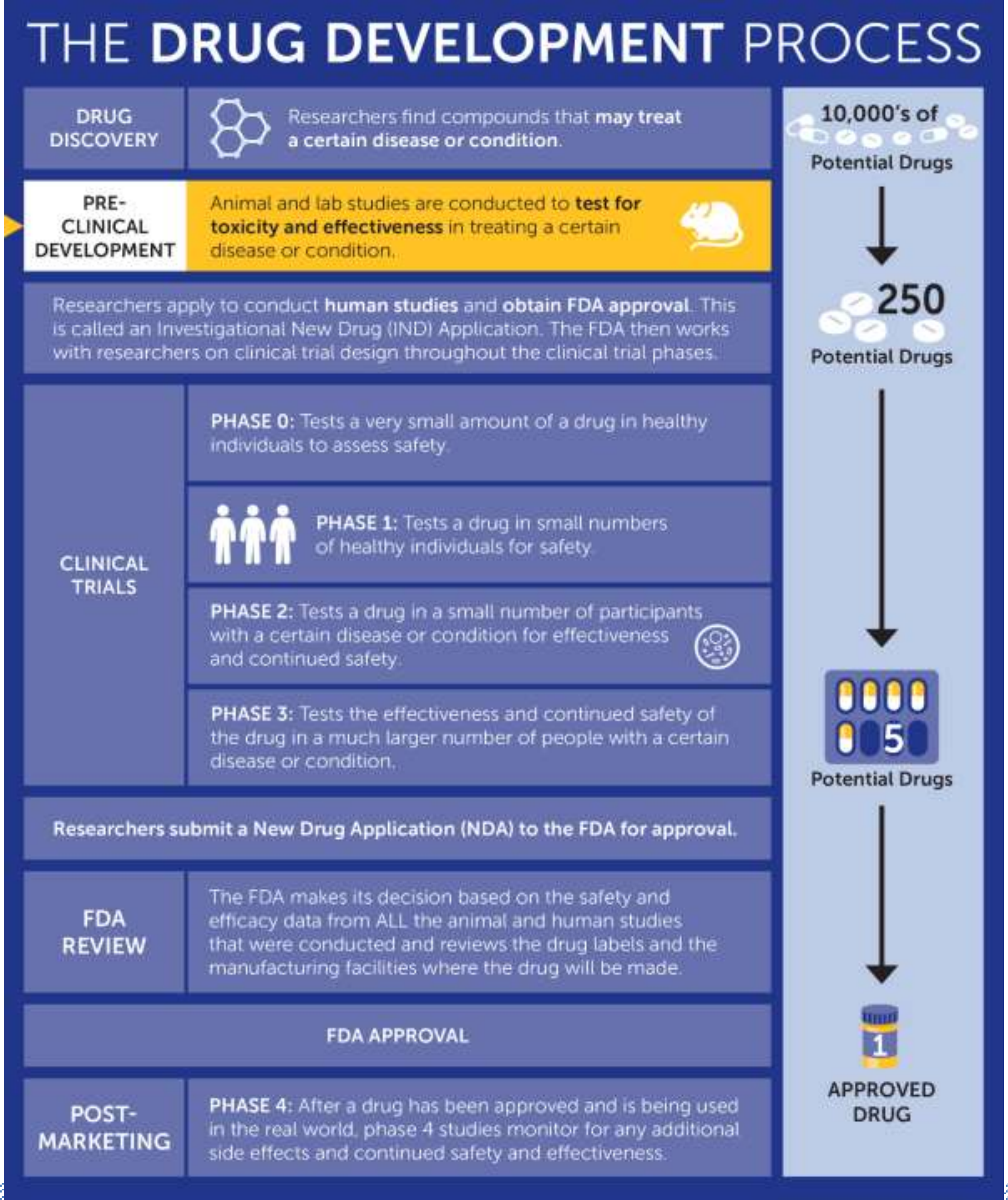
South China Normal University

About Clinical Trials

- prospective research studies on human participants
- designed to answer questions about biomedical or behavioral interventions, including treatment, diagnosis, and prevention of diseases or conditions.
- evaluate the **safety** and **efficacy**.

About Clinical Trials

- An important step in discovering new treatments for diseases as well as new ways to detect, diagnose, and reduce the risk of diseases.



THE DRUG DEVELOPMENT PROCESS

DRUG DISCOVERY — Researchers find compounds that **may treat a certain disease or condition.**

PRE-CLINICAL DEVELOPMENT — Animal and lab studies are conducted to **test for toxicity and effectiveness** in treating a certain disease or condition.

Researchers apply to conduct **human studies** and **obtain FDA approval**. This is called an Investigational New Drug (IND) Application. The FDA then works with researchers on clinical trial design throughout the clinical trial phases.

CLINICAL TRIALS

PHASE 0: Tests a very small amount of a drug in healthy individuals to assess safety.

PHASE 1: Tests a drug in small numbers of healthy individuals for safety.

PHASE 2: Tests a drug in a small number of participants with a certain disease or condition for effectiveness and continued safety.

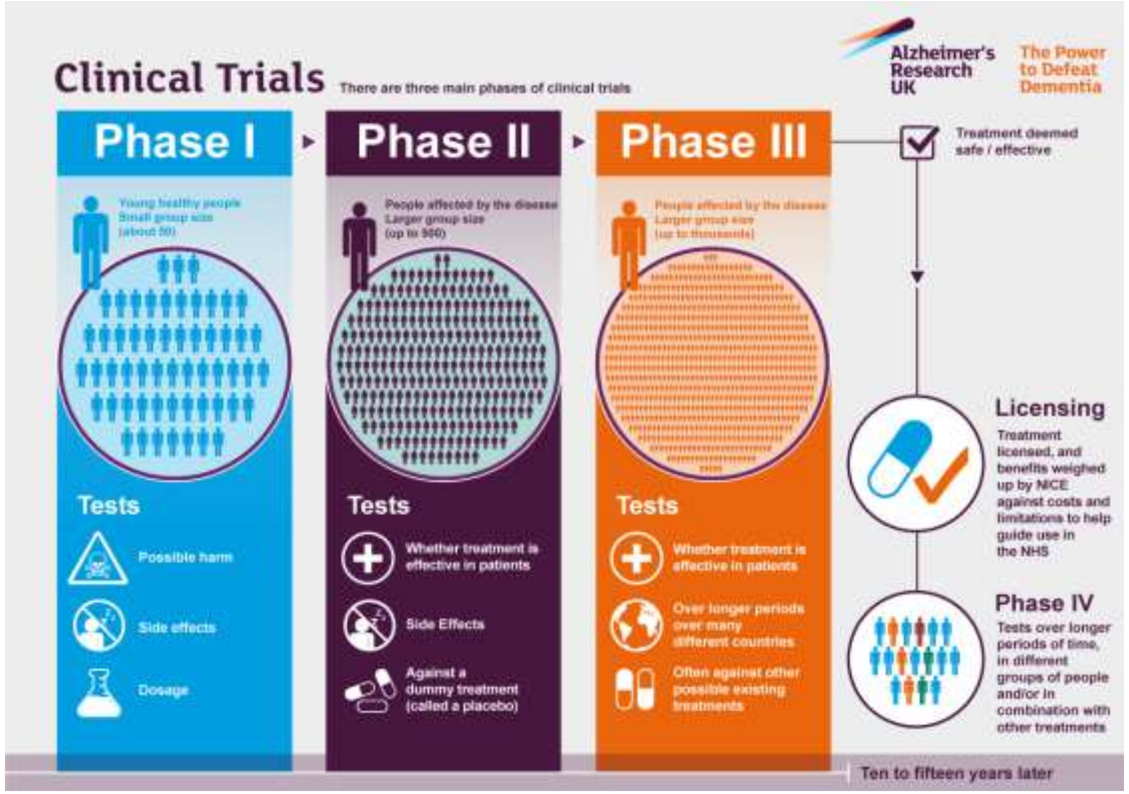
PHASE 3: Tests the effectiveness and continued safety of the drug in a much larger number of people with a certain disease or condition.

Researchers submit a New Drug Application (NDA) to the FDA for approval.

FDA REVIEW — The FDA makes its decision based on the safety and efficacy data from ALL the animal and human studies that were conducted and reviews the drug labels and the manufacturing facilities where the drug will be made.

FDA APPROVAL

POST-MARKETING — **PHASE 4:** After a drug has been approved and is being used in the real world, phase 4 studies monitor for any additional side effects and continued safety and effectiveness.



Clinical Trials There are three main phases of clinical trials

Alzheimer's Research UK — The Power to Defeat Dementia

Phase I
Young healthy people
Small group size
(about 50)

Tests
- Possible harm
- Side effects
- Dosage

Phase II
People affected by the disease
Larger group size
(up to 500)

Tests
- Whether treatment is effective in patients
- Side Effects
- Against a dummy treatment (called a placebo)

Phase III
People affected by the disease
Larger group size
(up to thousands)

Tests
- Whether treatment is effective in patients
- Over longer periods over many different countries
- Often against other possible existing treatments

Treatment deemed safe / effective

Licensing
Treatment licensed, and benefits weighed up by NICE against costs and limitations to help guide use in the NHS

Phase IV
Tests over longer periods of time, in different groups of people and/or in combination with other treatments

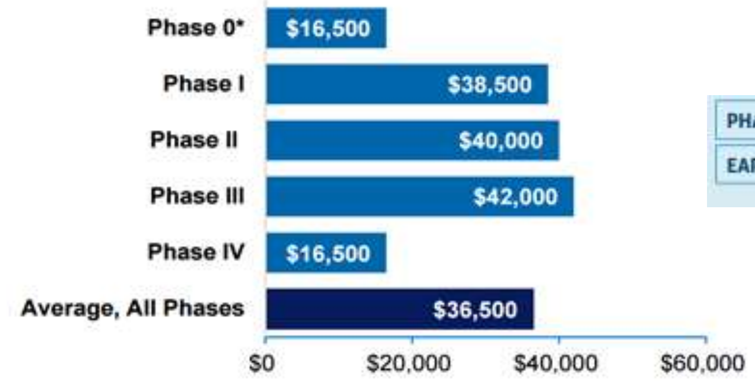
Ten to fifteen years later

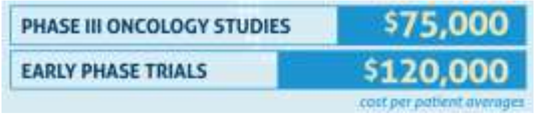
Different numbers of patients needed for different phases

The Difficulties

- **Costly:**

Average Per-Patient Clinical Trial Costs, by Phase, 2013



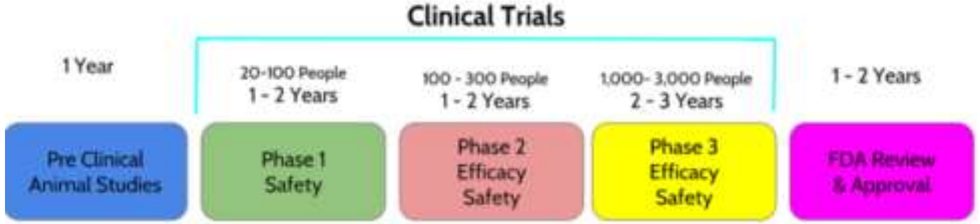


PHASE III ONCOLOGY STUDIES	$75,000
EARLY PHASE TRIALS	$120,000

cost per patient averages

By Tufts Center, the estimated average cost of developing a new medicine was **$2.6 billion** + **$312 million**.

- **Time-consuming:**



6 -10 years on average for oncology studies

WHY ARE CLINICAL TRIALS SO EXPENSIVE?

$2.6B

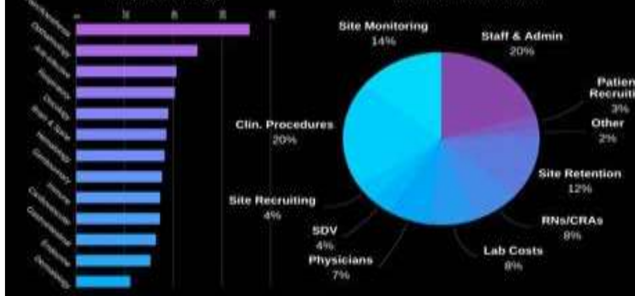
estimated development cost of a single drug (including failed drugs & opportunity cost) (1)

$36,500

cost for just one patient in a trial (2)

TRIAL COST BY INDICATION
in $ millions (3)

WHAT COSTS THE MOST?
within a clinical trial



ESOURCE & CTMS IS NEW TECH THAT CAN ALLEVIATE CLINICAL TRIAL COSTS IN:

Staff & Admin

Source Data Verification (SDV)

Site Monitoring

Patient (PT) Recruiting

ESOURCE

eSource is gaining popularity in clinical studies because it enables efficiencies for both sites and sponsors.

With eSource, site managers have oversight of their site operations and study data instantaneously, allowing them to evaluate their coordinators' work and QC remotely. With this instant access, it facilities a quicker and more cost-effective SDV process.

Ultimately, Source Data Verification may be integrated within the eSource system entirely, allowing sponsors to save money on CRA travel and site monitoring costs.

CTMS

By using a Clinical Trials Management System (CTMS) sites can save costs on recruiting and finances.

A strong CTMS application allows a site to recruit more efficiently by keeping their efforts organized and furthering their reach.

CTMS also works as an efficient financial management tool that tracks receivables, invoices, costs, reconciliation and patient stipends. Automated workflows reduce administrative costs, which total $2.3M (3) for the average Phase III Trial.

Bonus: Some CTMS solutions come integrated with eSource!

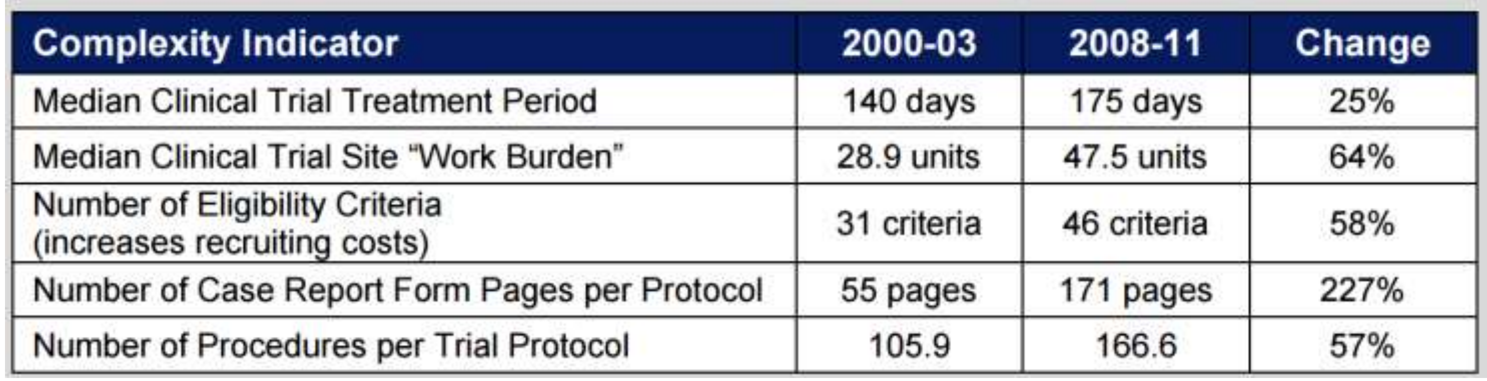
[1] Tufts Center for the Study of Drug Development (CSDD) 2014
[2] Battelle Technology Partnership Practice and PhRMA Biopharmaceutical Industry-Sponsored Clinical Trials: Impact on State Economies, March 2015
[3] Sertkaya et al. Key cost drivers of pharmaceutical clinical trials in the United States, 2016

clinical research

powerful, all-in-one eSource & CTMS

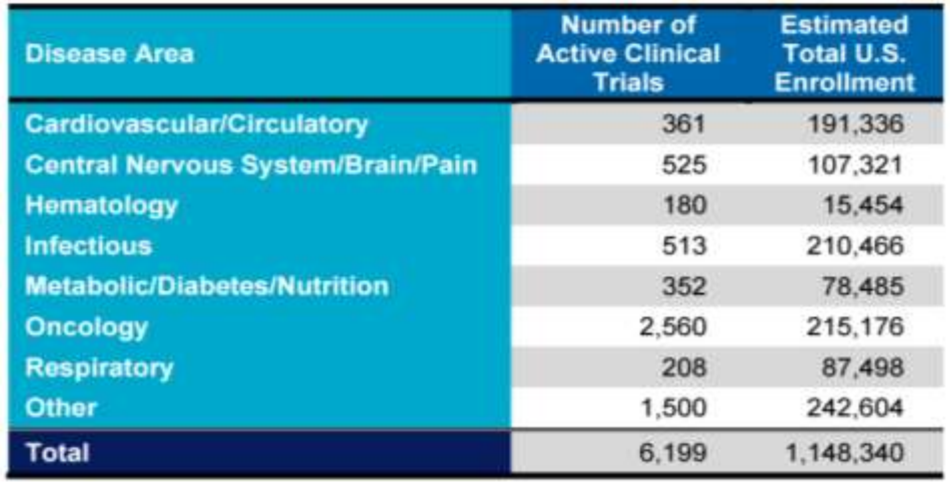
The Difficulties

- **Complex procedure:**



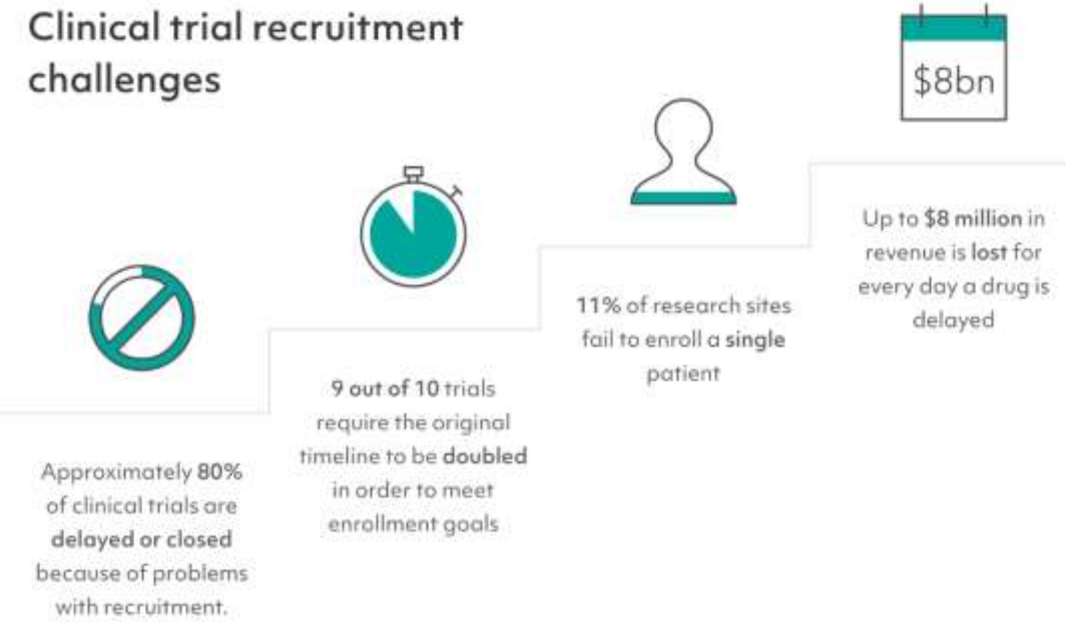
Complexity Indicator	2000-03	2008-11	Change
Median Clinical Trial Treatment Period	140 days	175 days	25%
Median Clinical Trial Site "Work Burden"	28.9 units	47.5 units	64%
Number of Eligibility Criteria (increases recruiting costs)	31 criteria	46 criteria	58%
Number of Case Report Form Pages per Protocol	55 pages	171 pages	227%
Number of Procedures per Trial Protocol	105.9	166.6	57%

- **Hard to recruit**



Disease Area	Number of Active Clinical Trials	Estimated Total U.S. Enrollment
Cardiovascular/Circulatory	361	191,336
Central Nervous System/Brain/Pain	525	107,321
Hematology	180	15,454
Infectious	513	210,466
Metabolic/Diabetes/Nutrition	352	78,485
Oncology	2,560	215,176
Respiratory	208	87,498
Other	1,500	242,604
Total	6,199	1,148,340

Estimated Number of Industry-Sponsored Clinical Trials and Trial Participants



Clinical trial recruitment challenges

Approximately **80%** of clinical trials are **delayed or closed** because of problems with recruitment.

9 out of 10 trials require the original timeline to be **doubled** in order to meet enrollment goals

11% of research sites fail to enroll a **single** patient

Up to **$8 million** in revenue is **lost** for every day a drug is delayed

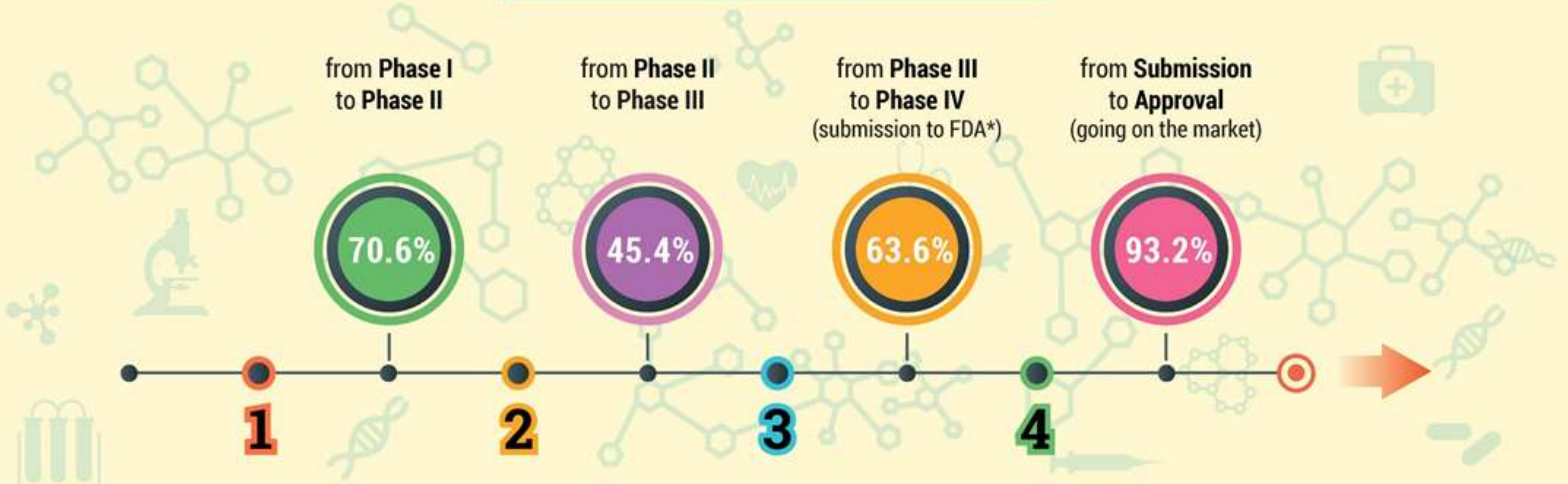
Expensive
Time-consuming

Unfortunately

Most clinical trials failed

Clinical Trials: The **Importance** of **The 4 Phases**

Success Rates by Phase



from **Phase I** to **Phase II**
70.6%

from **Phase II** to **Phase III**
45.4%

from **Phase III** to **Phase IV**
(submission to FDA*)
63.6%

from **Submission** to **Approval**
(going on the market)
93.2%

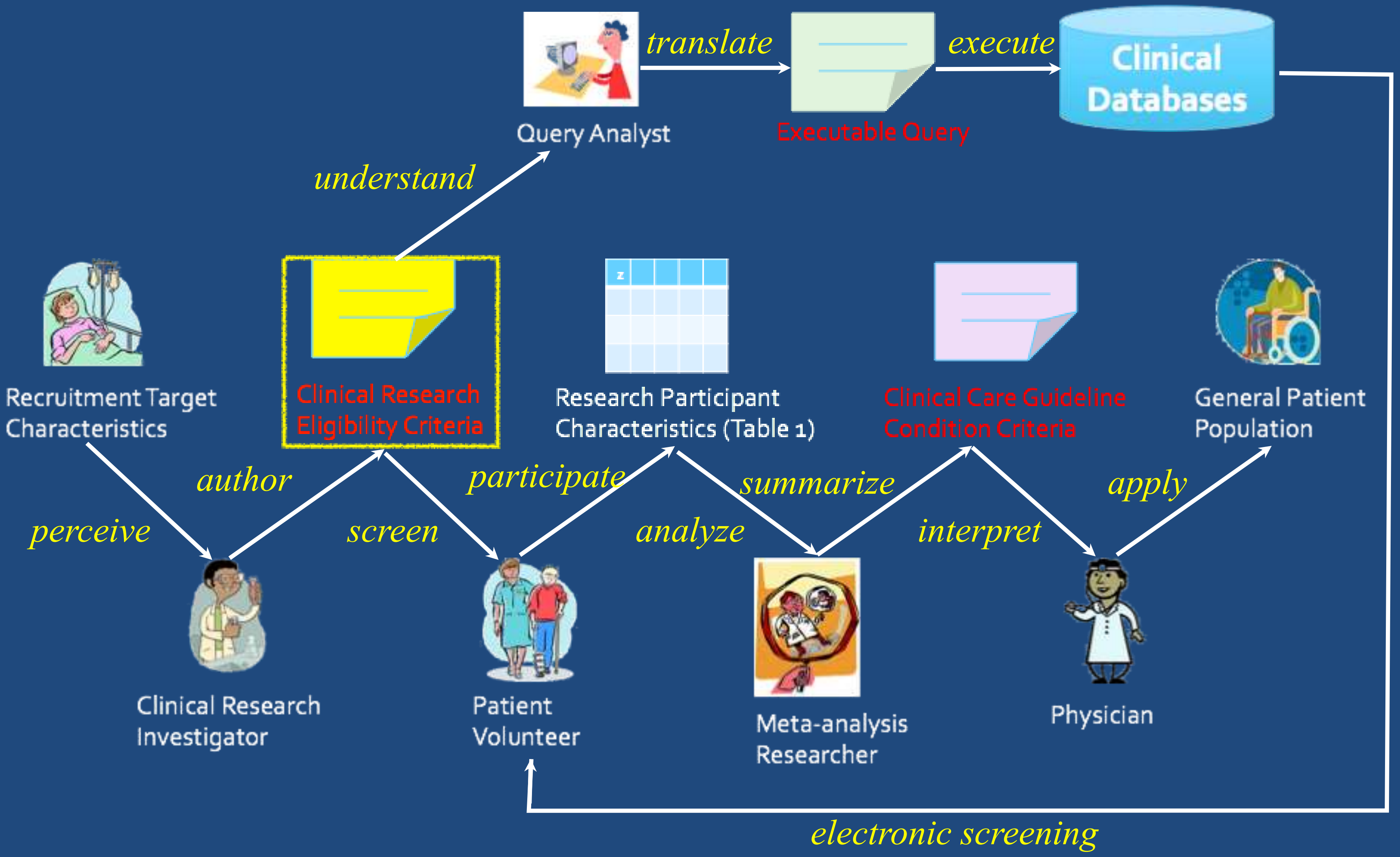
19% - Overall Success Rate from Phase I to FDA Submission

*FDA - The US Food and Drug Administration

www.findmecure.com

source: Tufts, CSDD

Eligibility Criteria: Central to Translational Research


Query Analyst
translate
Executable Query
execute
Clinical Databases
understand
Recruitment Target Characteristics
Clinical Research Eligibility Criteria
Research Participant Characteristics (Table 1)
Clinical Care Guideline Condition Criteria
General Patient Population
perceive
author
screen
participate
analyze
summarize
interpret
apply
Clinical Research Investigator
Patient Volunteer
Meta-analysis Researcher
Physician
electronic screening

Research questions

Q1: ***How to represent key information of eligibility criteria semantically and consistently?***

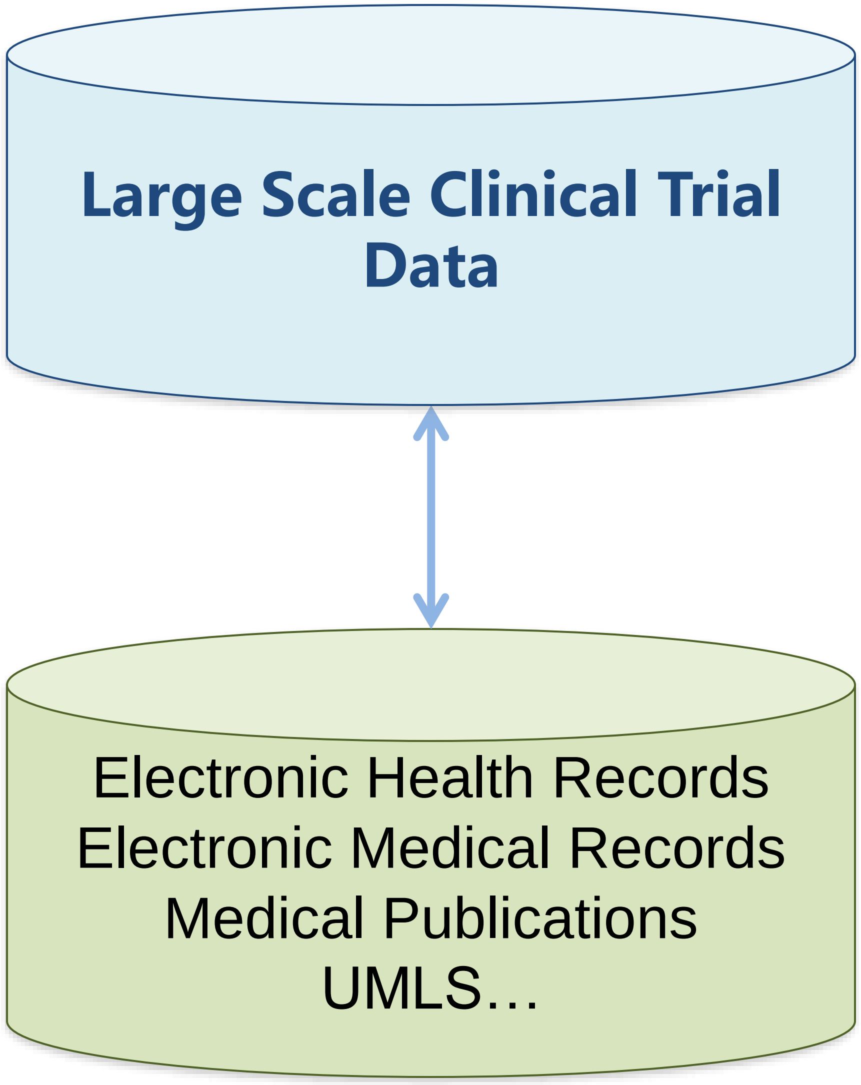
Q2: ***How to extract key information accurately from free eligibility criteria text for patient recruitment?***

Q3: ***How to accurately match study population and real patient population in EMRs from hospitals?***

Q4: ***How to reduce the gap between clinical trial study population and real patient population?***

….

Need real medical data!

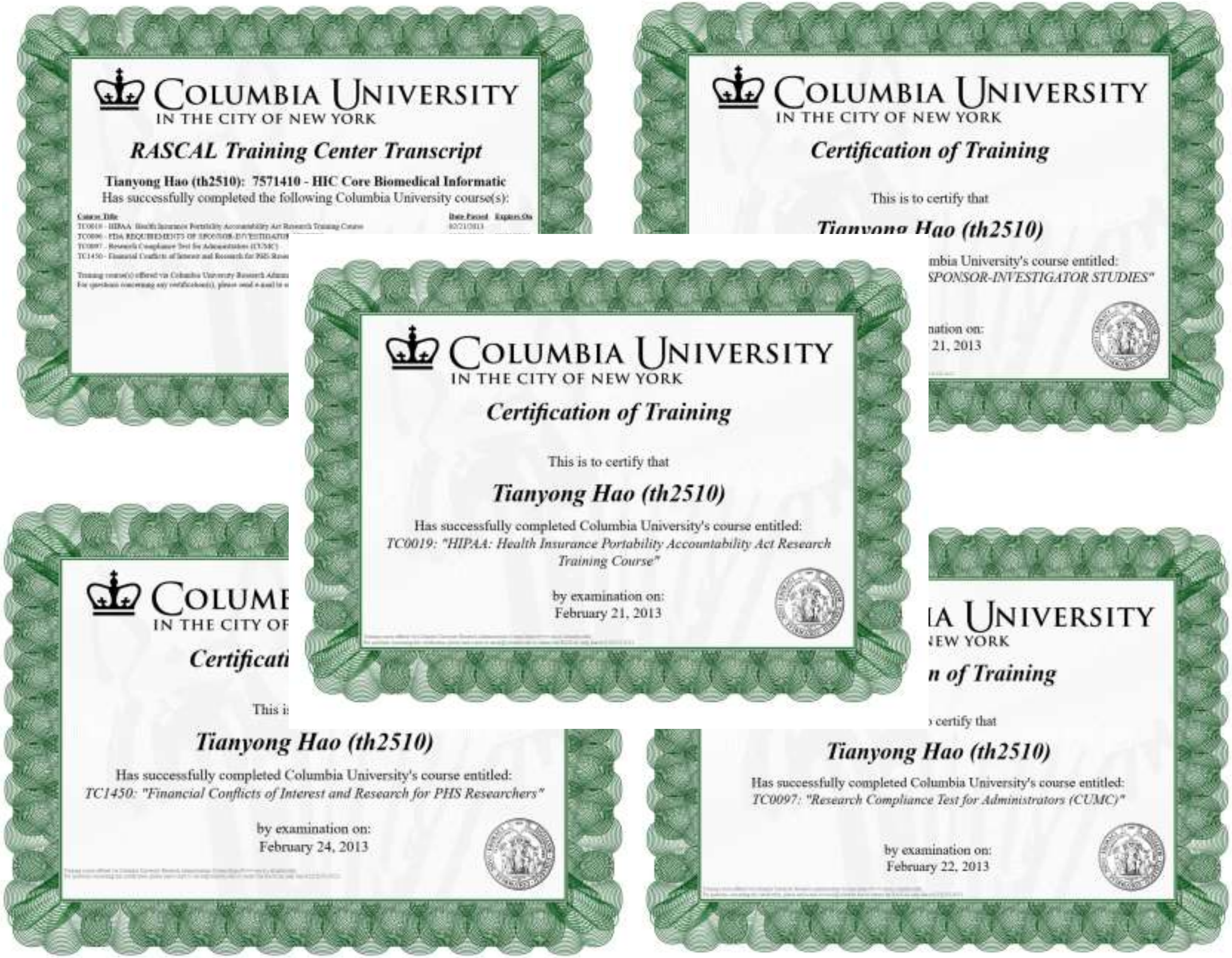

Large Scale Clinical Trial Data

Electronic Health Records
Electronic Medical Records
Medical Publications
UMLS…

Research collaborations

- Columbia U Medical Center
- 广东省中医院
- 中山大学医学院
- 中山大学附属肿瘤医院
- 中山大学附属第三医院
- 南方战区总医院
- 广州医科大学附属第一医院
- 浙江省人民医院
- 重庆医科大学附属儿童医院
- 广州中医药大学(联合博导)
- 深圳市罗湖区人民医院 (联合博后导师)

iHAVC 智慧健康与可视化计算
Intelligent Health and Visual Computing



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

RASCAL Training Center Transcript

Tianyong Hao (th2510): 7571410 - HIC Core Biomedical Informatic
Has successfully completed the following Columbia University course(s):

Course Title	Date Passed	Expires On
TC0019 - HIPAA: Health Insurance Portability Accountability Act Research Training Course	02/21/2013	
TC0096 - FDA REQUIREMENTS OF SPONSOR-INVESTIGATOR		
TC0097 - Research Compliance Test for Administrators (CUMC)		
TC1450 - Financial Conflicts of Interest and Research for PHS Rese		

Training course(s) offered via Columbia University Research Admin
For questions concerning any certification(s), please send e-mail to

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Certification of Training

This is to certify that

Tianyong Hao (th2510)

mbia University's course entitled:
SPONSOR-INVESTIGATOR STUDIES"

nation on:
21, 2013

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Certification of Training

This is to certify that

Tianyong Hao (th2510)

Has successfully completed Columbia University's course entitled:
TC0019: "HIPAA: Health Insurance Portability Accountability Act Research Training Course"

by examination on:
February 21, 2013

COLUMB
IN THE CITY OF

Certificati

This is

Tianyong Hao (th2510)

Has successfully completed Columbia University's course entitled:
TC1450: "Financial Conflicts of Interest and Research for PHS Researchers"

by examination on:
February 24, 2013

A UNIVERSITY
NEW YORK

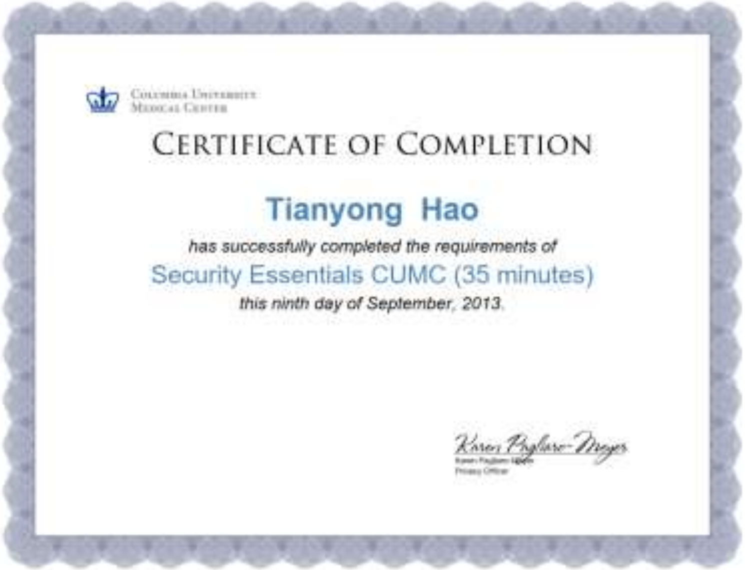
n of Training

certify that

Tianyong Hao (th2510)

Has successfully completed Columbia University's course entitled:
TC0097: "Research Compliance Test for Administrators (CUMC)"

by examination on:
February 22, 2013



COLUMBIA UNIVERSITY
MEDICAL CENTER

CERTIFICATE OF COMPLETION

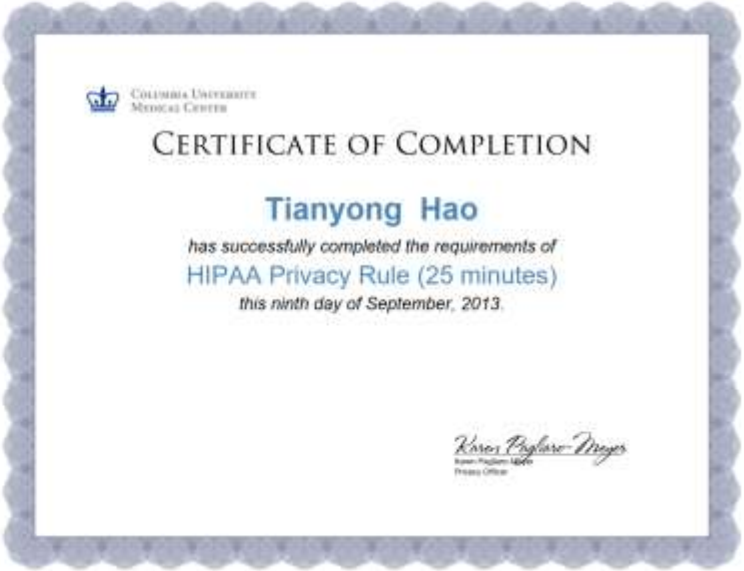
Tianyong Hao

has successfully completed the requirements of

Security Essentials CUMC (35 minutes)

this ninth day of September, 2013.

Karen Pagliaro-Meyer
Karen Pagliaro-Meyer
Privacy Officer



COLUMBIA UNIVERSITY
MEDICAL CENTER

CERTIFICATE OF COMPLETION

Tianyong Hao

has successfully completed the requirements of

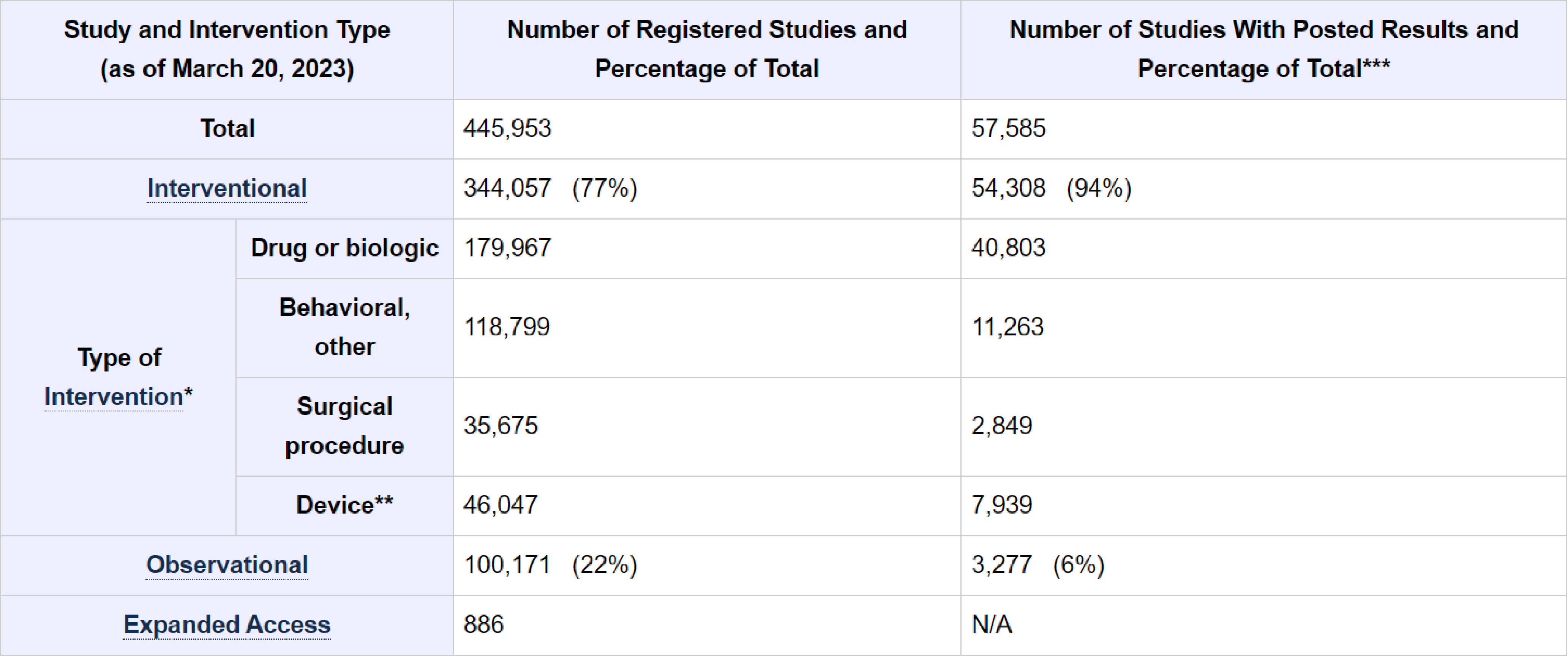
HIPAA Privacy Rule (25 minutes)

this ninth day of September, 2013.

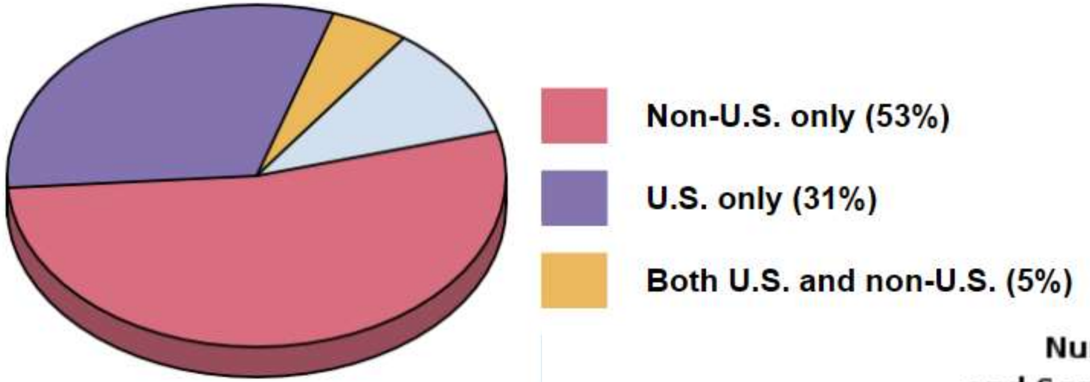
Karen Pagliaro-Meyer
Karen Pagliaro-Meyer
Privacy Officer

Clinical Trial data

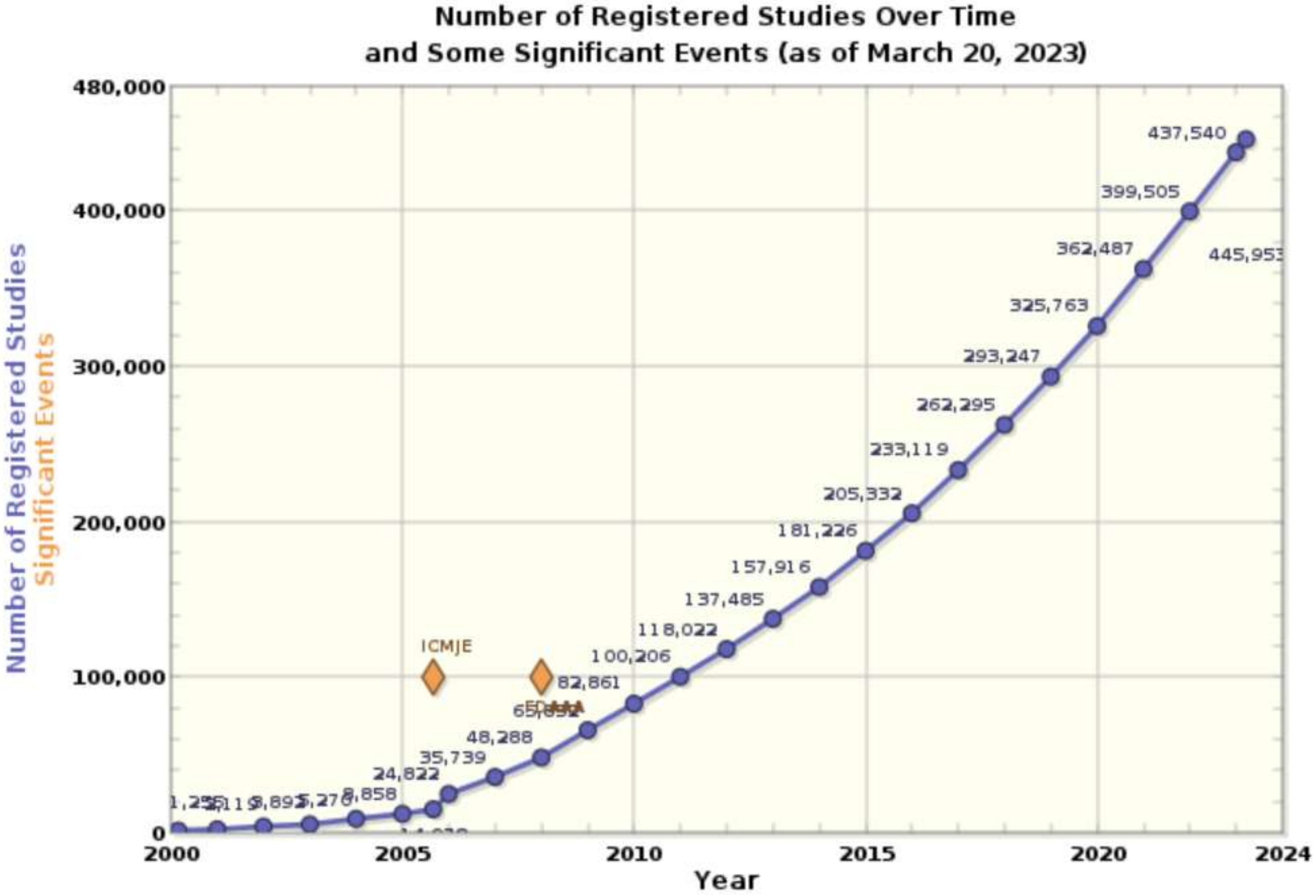
- March 20, 2023
- **445,953 clinical trials; 221 countries**



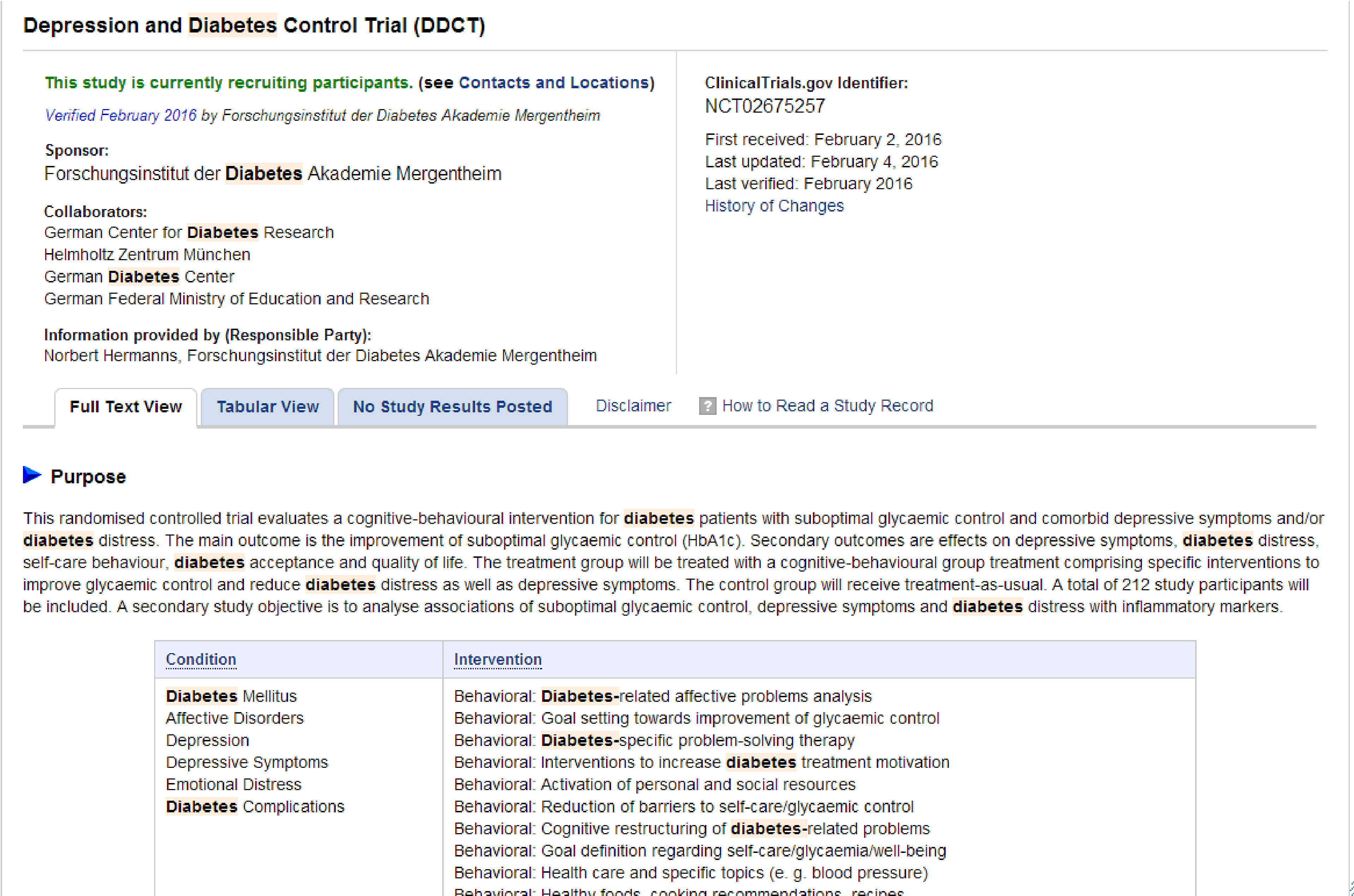
Study and Intervention Type (as of March 20, 2023)		Number of Registered Studies and Percentage of Total	Number of Studies With Posted Results and Percentage of Total***
Total		445,953	57,585
Interventional		344,057 (77%)	54,308 (94%)
Type of Intervention*	**Drug or biologic**	179,967	40,803
	Behavioral, other	118,799	11,263
	Surgical procedure	35,675	2,849
	Device**	46,047	7,939
Observational		100,171 (22%)	3,277 (6%)
Expanded Access		886	N/A



Location	Number
Non-U.S. only	236,247
U.S. only	138,919
Both U.S. and non-U.S.	21,931
Not provided	48,856
Total	445,953


Number of Registered Studies Over Time and Some Significant Events (as of March 20, 2023)

https://www.clinicaltrial.gov/ct2/show/NCT02675257
https://www.clinicaltrial.gov/ct2/show/results/NCT01009138



Depression and Diabetes Control Trial (DDCT)

This study is currently recruiting participants. (see **Contacts and Locations**)

Verified February 2016 by Forschungsinstitut der Diabetes Akademie Mergentheim

Sponsor:
Forschungsinstitut der **Diabetes** Akademie Mergentheim

Collaborators:
German Center for **Diabetes** Research
Helmholtz Zentrum München
German **Diabetes** Center
German Federal Ministry of Education and Research

Information provided by (Responsible Party):
Norbert Hermanns, Forschungsinstitut der Diabetes Akademie Mergentheim

ClinicalTrials.gov Identifier:
NCT02675257

First received: February 2, 2016
Last updated: February 4, 2016
Last verified: February 2016
History of Changes

▶ Purpose

This randomised controlled trial evaluates a cognitive-behavioural intervention for **diabetes** patients with suboptimal glycaemic control and comorbid depressive symptoms and/or **diabetes** distress. The main outcome is the improvement of suboptimal glycaemic control (HbA1c). Secondary outcomes are effects on depressive symptoms, **diabetes** distress, self-care behaviour, **diabetes** acceptance and quality of life. The treatment group will be treated with a cognitive-behavioural group treatment comprising specific interventions to improve glycaemic control and reduce **diabetes** distress as well as depressive symptoms. The control group will receive treatment-as-usual. A total of 212 study participants will be included. A secondary study objective is to analyse associations of suboptimal glycaemic control, depressive symptoms and **diabetes** distress with inflammatory markers.

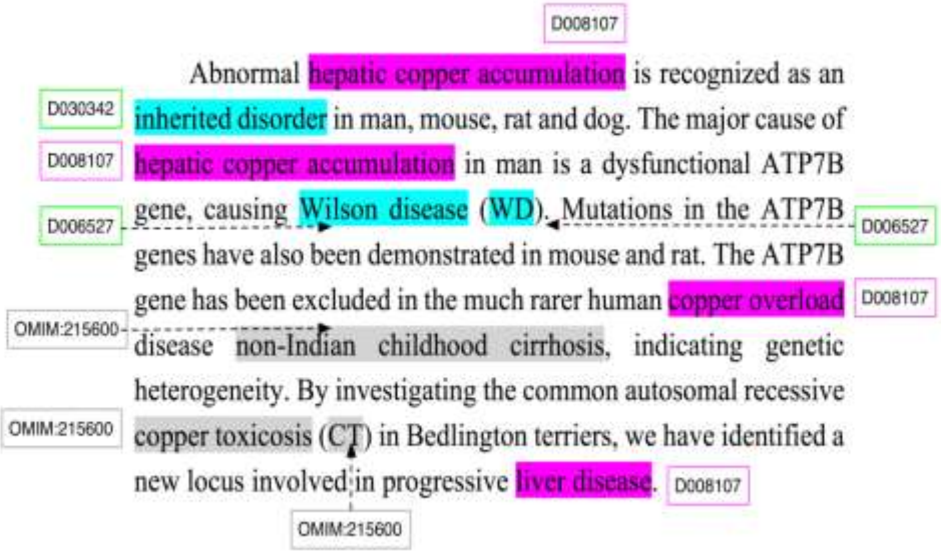
Condition	Intervention
Diabetes Mellitus Affective Disorders Depression Depressive Symptoms Emotional Distress **Diabetes** Complications	Behavioral: **Diabetes-**related affective problems analysis Behavioral: Goal setting towards improvement of glycaemic control Behavioral: **Diabetes-**specific problem-solving therapy Behavioral: Interventions to increase **diabetes** treatment motivation Behavioral: Activation of personal and social resources Behavioral: Reduction of barriers to self-care/glycaemic control Behavioral: Cognitive restructuring of **diabetes-**related problems Behavioral: Goal definition regarding self-care/glycaemia/well-being Behavioral: Health care and specific topics (e. g. blood pressure) Behavioral: Healthy foods, cooking recommendations, recipes

Research topics

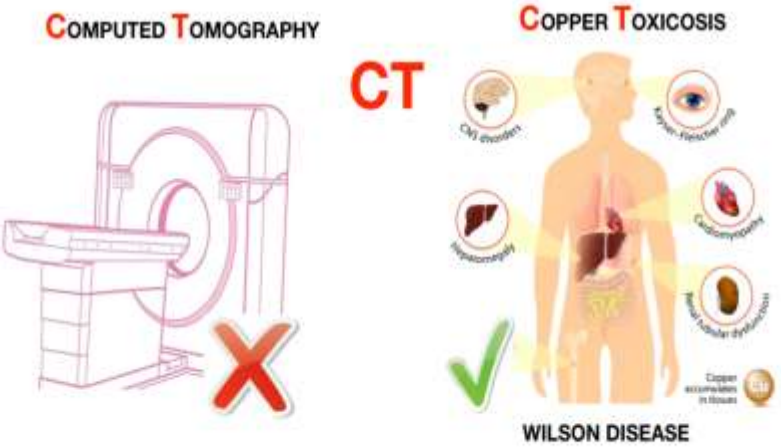
- Semantic tag mining from eligibility criteria text
- Parsing and structuring eligibility criteria text
- Semantic computing and matching of eligibility criteria
- Classification of eligibility criteria text
- Clinical trial clustering
- Personalized clinical trial search and recommendation
- Partnership extraction enhancing clinical trial recruitment
- Gender extraction for enhancing clinical trial recruitment
- Matching eligibility criteria to patient EMRs for automatic recruitment
- Measurable quantitative information and extraction
- …

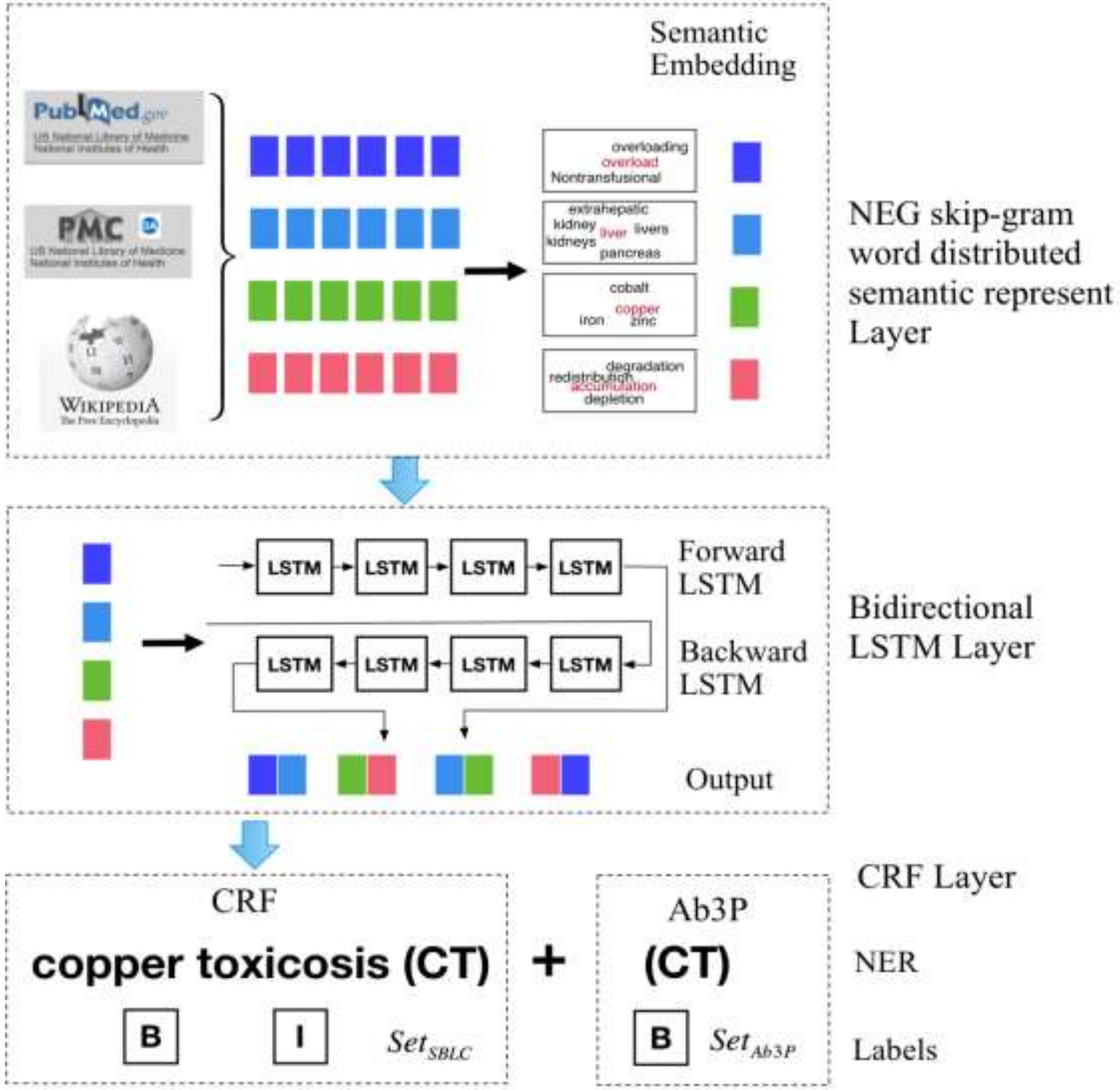
Semantic Tag Mining

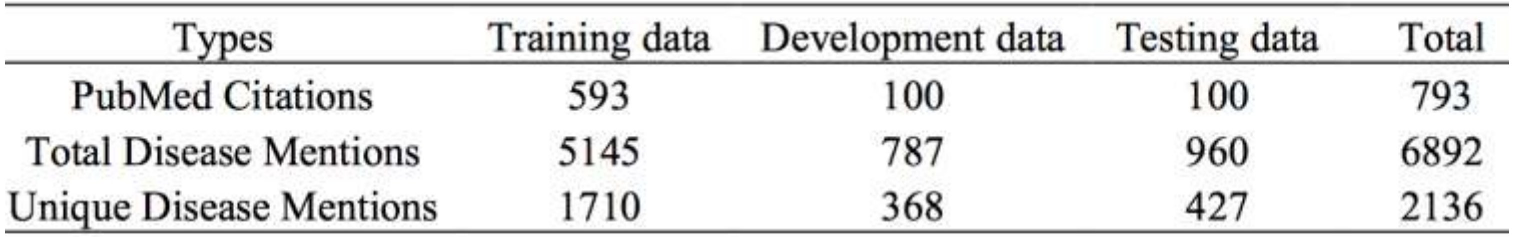
Motivation



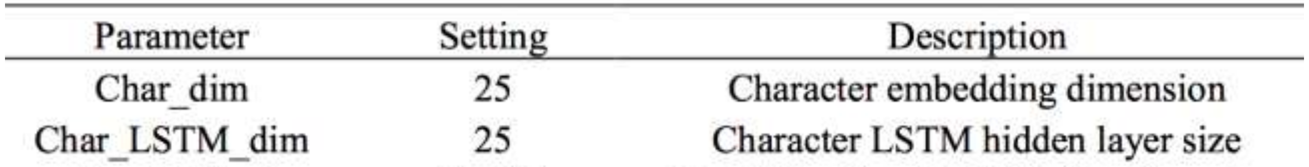
Abnormal hepatic copper accumulation is recognized as an inherited disorder in man, mouse, rat and dog. The major cause of hepatic copper accumulation in man is a dysfunctional ATP7B gene, causing Wilson disease (WD). Mutations in the ATP7B genes have also been demonstrated in mouse and rat. The ATP7B gene has been excluded in the much rarer human copper overload disease non-Indian childhood cirrhosis, indicating genetic heterogeneity. By investigating the common autosomal recessive copper toxicosis (CT) in Bedlington terriers, we have identified a new locus involved in progressive liver disease.



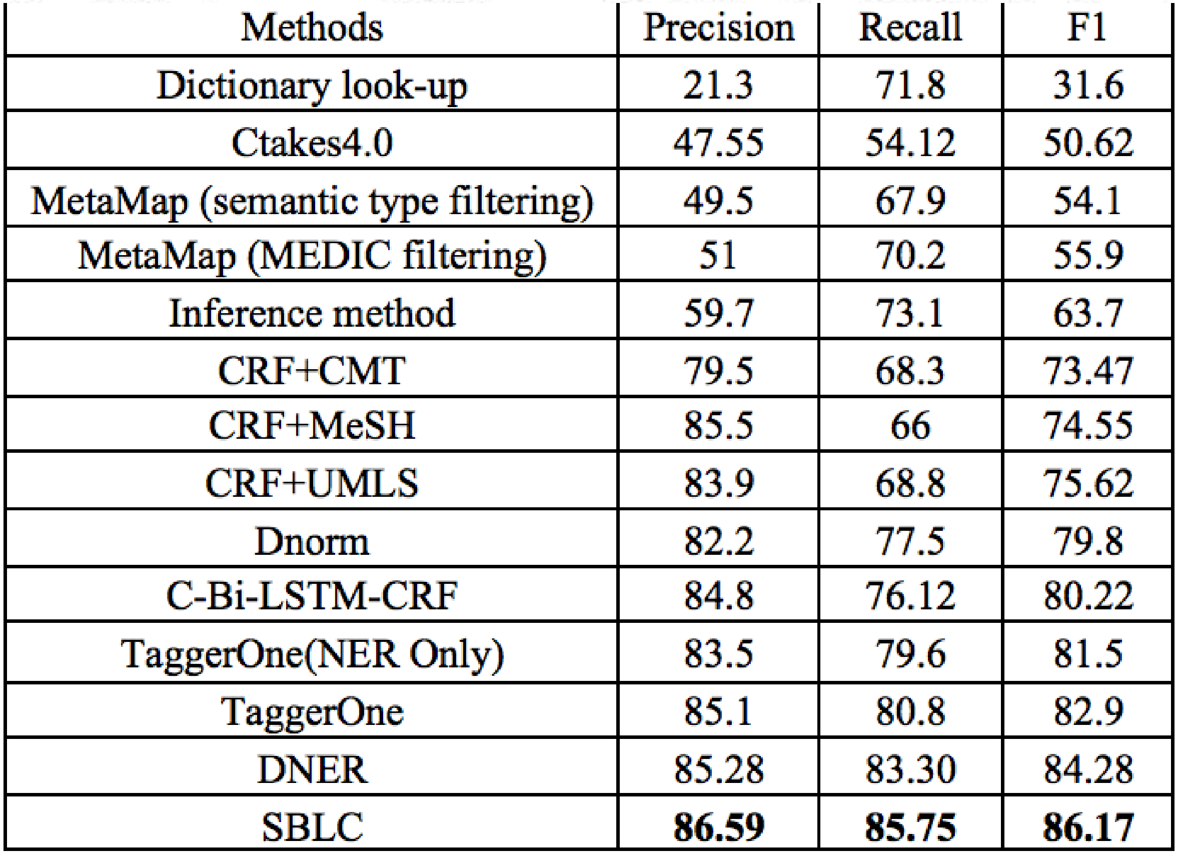




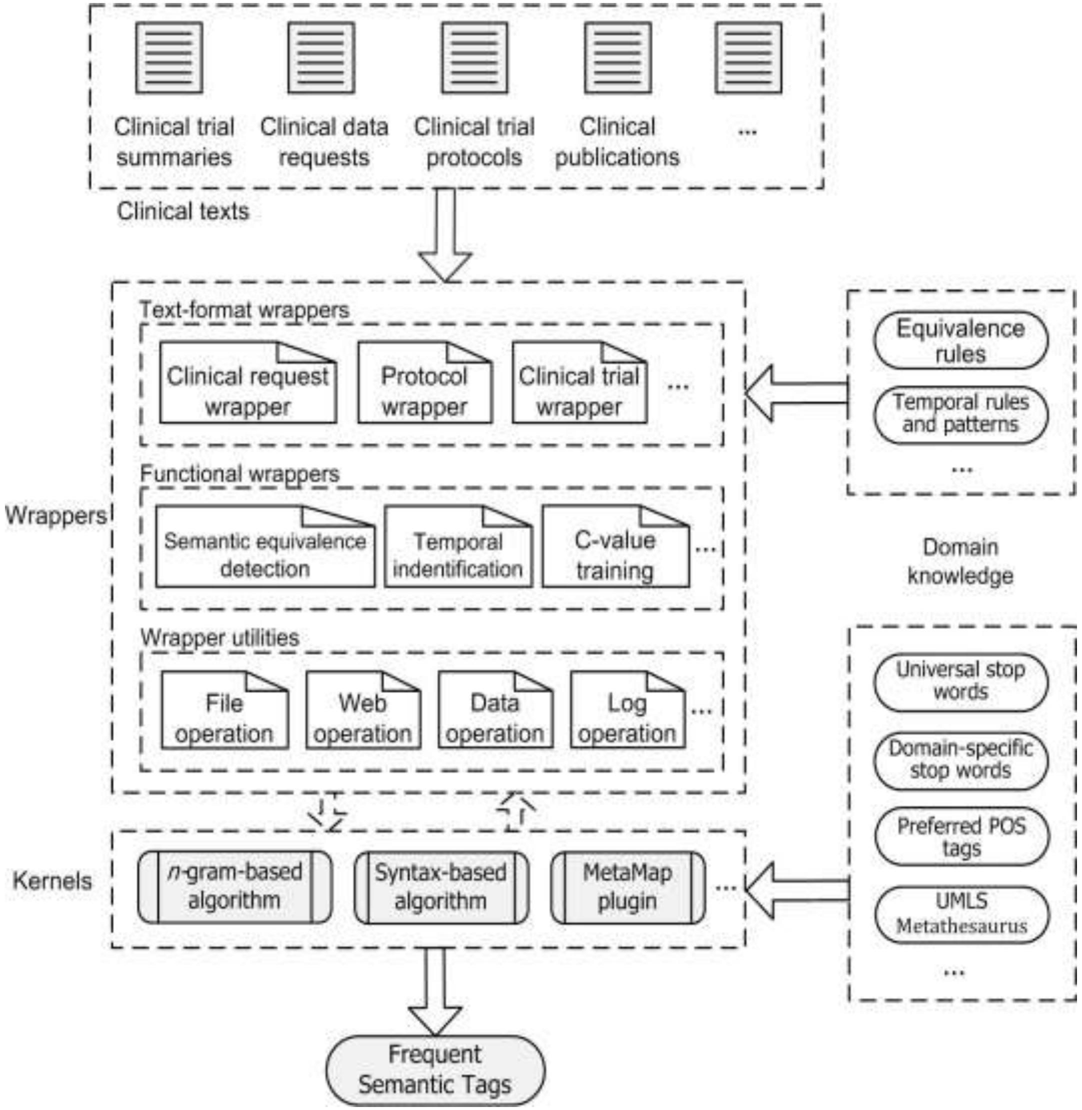
Types	Training data	Development data	Testing data	Total
PubMed Citations	593	100	100	793
Total Disease Mentions	5145	787	960	6892
Unique Disease Mentions	1710	368	427	2136



Parameter	Setting	Description
Char_dim	25	Character embedding dimension
Char_LSTM_dim	25	Character LSTM hidden layer size



Methods	Precision	Recall	F1
Dictionary look-up	21.3	71.8	31.6
Ctakes4.0	47.55	54.12	50.62
MetaMap (semantic type filtering)	49.5	67.9	54.1
MetaMap (MEDIC filtering)	51	70.2	55.9
Inference method	59.7	73.1	63.7
CRF+CMT	79.5	68.3	73.47
CRF+MeSH	85.5	66	74.55
CRF+UMLS	83.9	68.8	75.62
Dnorm	82.2	77.5	79.8
C-Bi-LSTM-CRF	84.8	76.12	80.22
TaggerOne(NER Only)	83.5	79.6	81.5
TaggerOne	85.1	80.8	82.9
DNER	85.28	83.30	84.28
SBLC	**86.59**	**85.75**	**86.17**



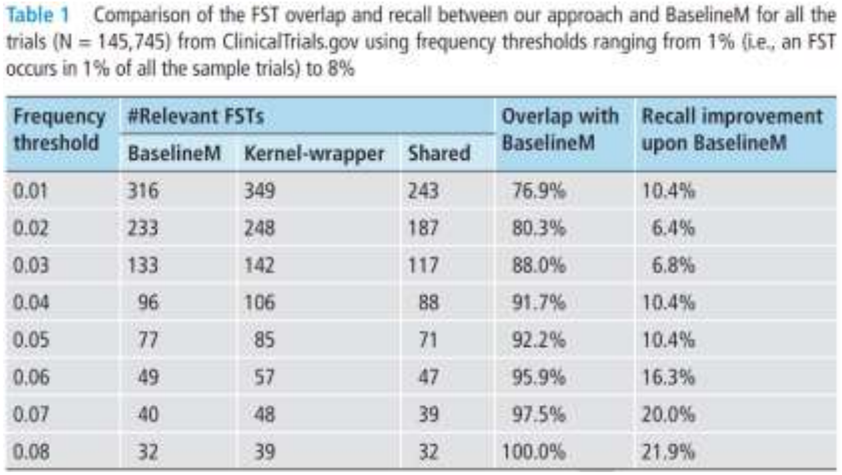
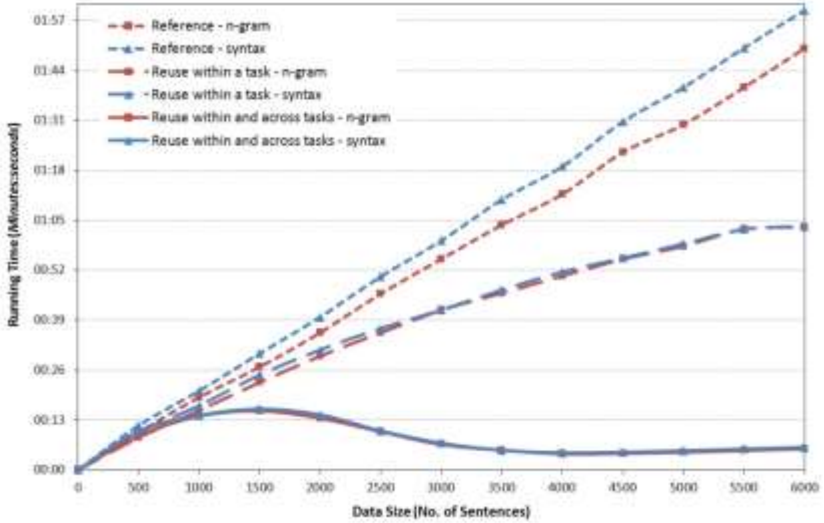


Table 1 Comparison of the FST overlap and recall between our approach and BaselineM for all the trials (N = 145,745) from ClinicalTrials.gov using frequency thresholds ranging from 1% (i.e., an FST occurs in 1% of all the sample trials) to 8%

Frequency threshold	#Relevant FSTs			Overlap with BaselineM	Recall improvement upon BaselineM
	BaselineM	Kernel-wrapper	Shared		
0.01	316	349	243	76.9%	10.4%
0.02	233	248	187	80.3%	6.4%
0.03	133	142	117	88.0%	6.8%
0.04	96	106	88	91.7%	10.4%
0.05	77	85	71	92.2%	10.4%
0.06	49	57	47	95.9%	16.3%
0.07	40	48	39	97.5%	20.0%
0.08	32	39	32	100.0%	21.9%



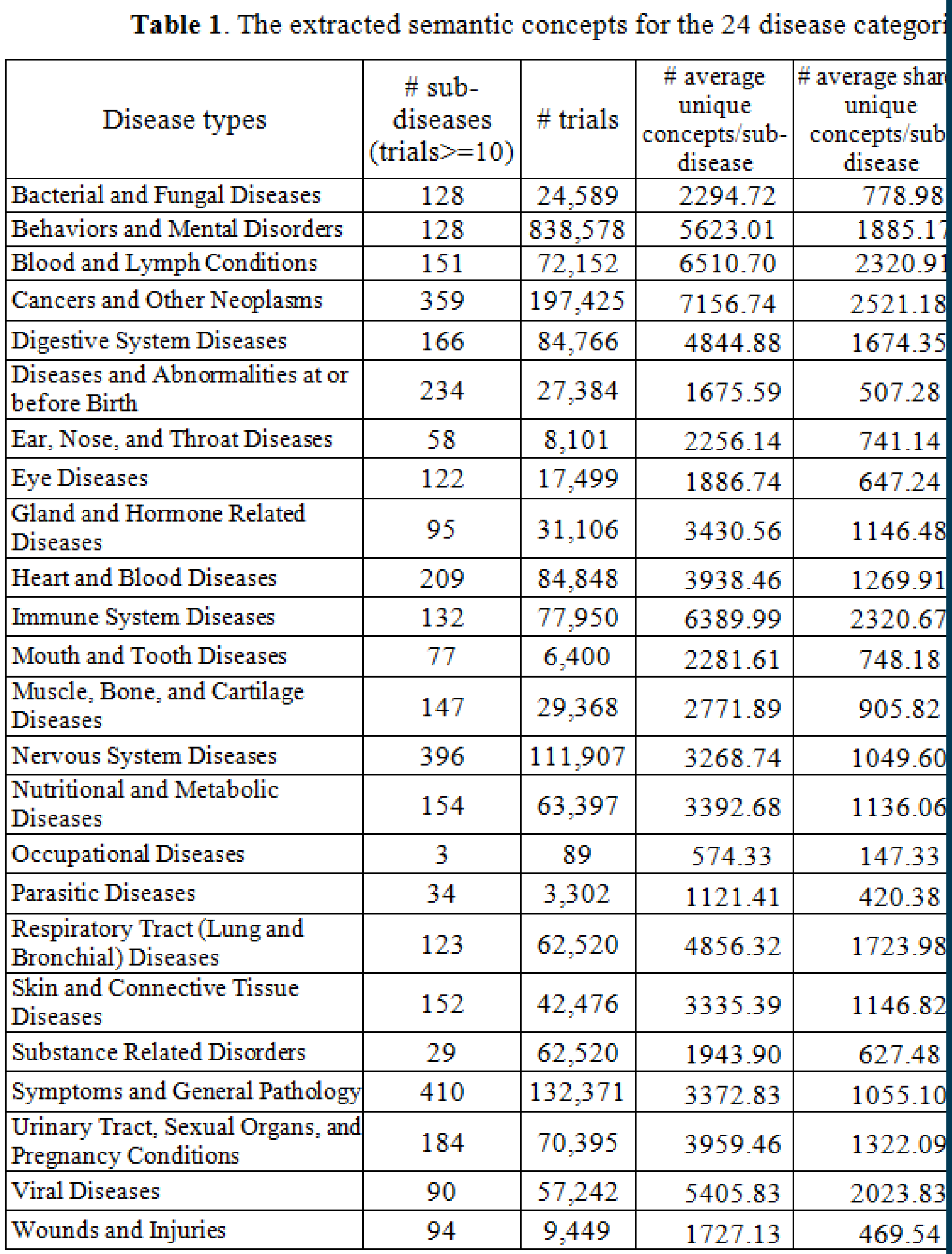
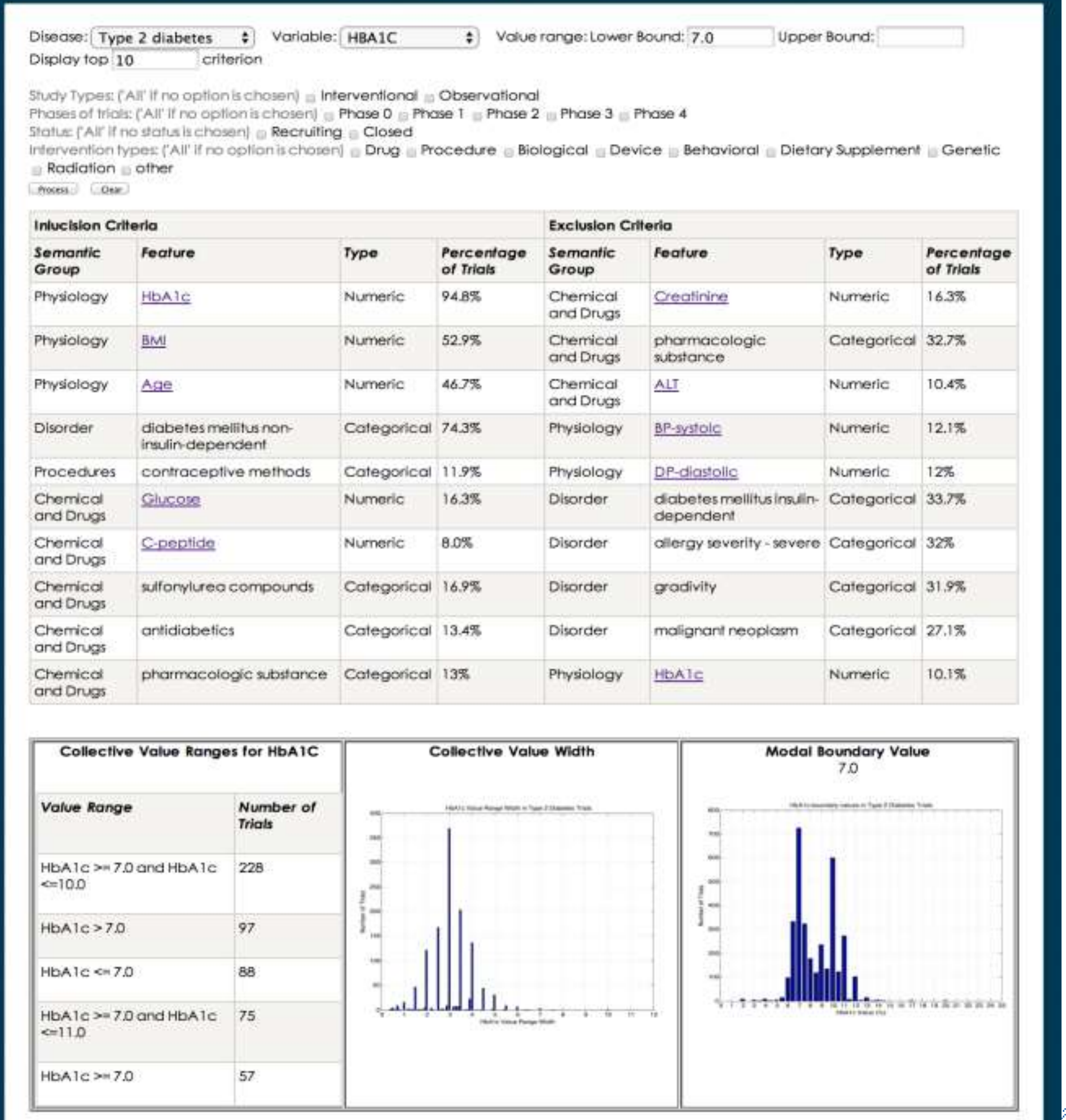


Table 1. The extracted semantic concepts for the 24 disease categori

Disease types	# sub-diseases (trials>=10)	# trials	# average unique concepts/sub-disease	# average shar unique concepts/sub disease
Bacterial and Fungal Diseases	128	24,589	2294.72	778.98
Behaviors and Mental Disorders	128	838,578	5623.01	1885.17
Blood and Lymph Conditions	151	72,152	6510.70	2320.91
Cancers and Other Neoplasms	359	197,425	7156.74	2521.18
Digestive System Diseases	166	84,766	4844.88	1674.35
Diseases and Abnormalities at or before Birth	234	27,384	1675.59	507.28
Ear, Nose, and Throat Diseases	58	8,101	2256.14	741.14
Eye Diseases	122	17,499	1886.74	647.24
Gland and Hormone Related Diseases	95	31,106	3430.56	1146.48
Heart and Blood Diseases	209	84,848	3938.46	1269.91
Immune System Diseases	132	77,950	6389.99	2320.67
Mouth and Tooth Diseases	77	6,400	2281.61	748.18
Muscle, Bone, and Cartilage Diseases	147	29,368	2771.89	905.82
Nervous System Diseases	396	111,907	3268.74	1049.60
Nutritional and Metabolic Diseases	154	63,397	3392.68	1136.06
Occupational Diseases	3	89	574.33	147.33
Parasitic Diseases	34	3,302	1121.41	420.38
Respiratory Tract (Lung and Bronchial) Diseases	123	62,520	4856.32	1723.98
Skin and Connective Tissue Diseases	152	42,476	3335.39	1146.82
Substance Related Disorders	29	62,520	1943.90	627.48
Symptoms and General Pathology	410	132,371	3372.83	1055.10
Urinary Tract, Sexual Organs, and Pregnancy Conditions	184	70,395	3959.46	1322.09
Viral Diseases	90	57,242	5405.83	2023.83
Wounds and Injuries	94	9,449	1727.13	469.54

Commonalities in Target Populations in Eligibility Criteria



Disease: Type 2 diabetes Variable: HBA1C Value range: Lower Bound: 7.0 Upper Bound:
Display top 10 criterion

Study Types: ('All' if no option is chosen) □ Interventional □ Observational
Phases of trials: ('All' if no option is chosen) □ Phase 0 □ Phase 1 □ Phase 2 □ Phase 3 □ Phase 4
Status: ('All' if no status is chosen) □ Recruiting □ Closed
Intervention types: ('All' if no option is chosen) □ Drug □ Procedure □ Biological □ Device □ Behavioral □ Dietary Supplement □ Genetic □ Radiation □ other

Inlucision Criteria				Exclusion Criteria			
Semantic Group	*Feature*	*Type*	*Percentage of Trials*	*Semantic Group*	*Feature*	*Type*	*Percentage of Trials*
Physiology	HbA1c	Numeric	94.8%	Chemical and Drugs	Creatinine	Numeric	16.3%
Physiology	BMI	Numeric	52.9%	Chemical and Drugs	pharmacologic substance	Categorical	32.7%
Physiology	Age	Numeric	46.7%	Chemical and Drugs	ALT	Numeric	10.4%
Disorder	diabetes mellitus non-insulin-dependent	Categorical	74.3%	Physiology	BP-systolic	Numeric	12.1%
Procedures	contraceptive methods	Categorical	11.9%	Physiology	DP-diastolic	Numeric	12%
Chemical and Drugs	Glucose	Numeric	16.3%	Disorder	diabetes mellitus insulin-dependent	Categorical	33.7%
Chemical and Drugs	C-peptide	Numeric	8.0%	Disorder	allergy severity - severe	Categorical	32%
Chemical and Drugs	sulfonylurea compounds	Categorical	16.9%	Disorder	gradivity	Categorical	31.9%
Chemical and Drugs	antidiabetics	Categorical	13.4%	Disorder	malignant neoplasm	Categorical	27.1%
Chemical and Drugs	pharmacologic substance	Categorical	13%	Physiology	HbA1c	Numeric	10.1%

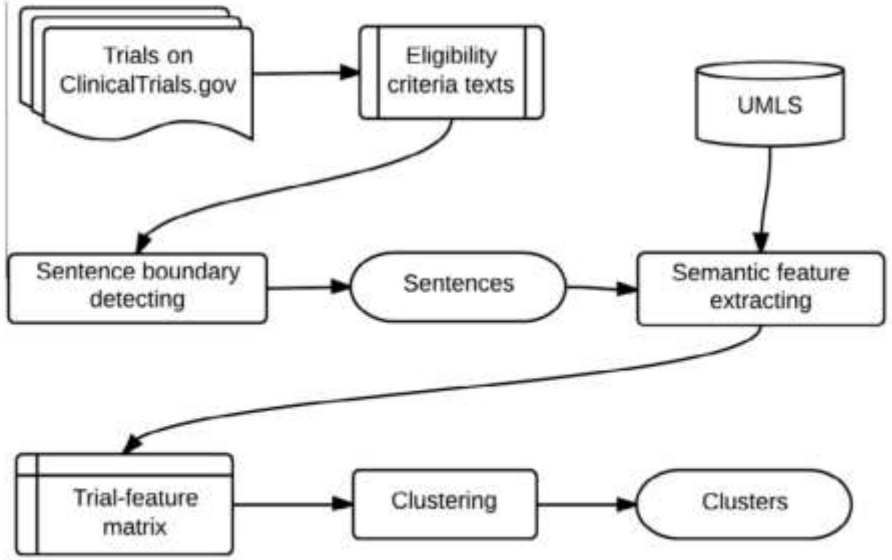
Collective Value Ranges for HbA1C

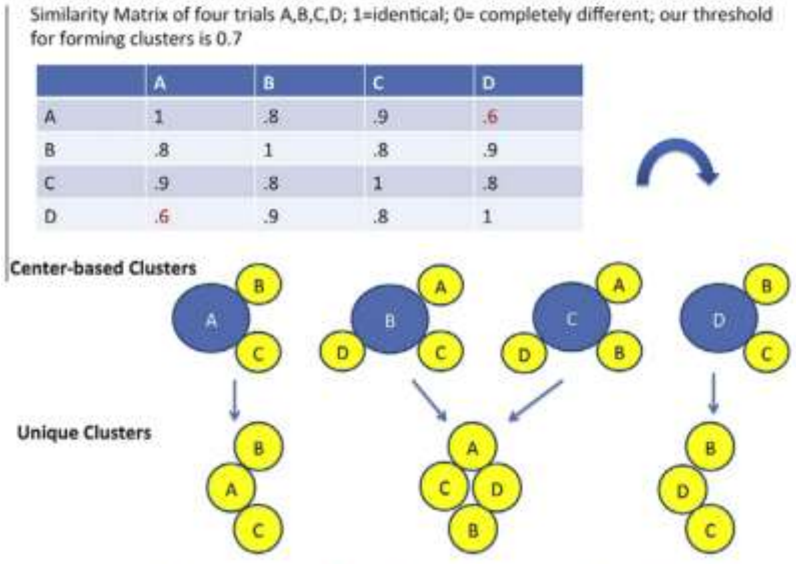
Value Range	*Number of Trials*
HbA1c >= 7.0 and HbA1c <=10.0	228
HbA1c > 7.0	97
HbA1c <= 7.0	88
HbA1c >= 7.0 and HbA1c <=11.0	75
HbA1c >= 7.0	57

Collective Value Width

Modal Boundary Value
7.0

Clinical Trial Clustering


Trials on ClinicalTrials.gov
Eligibility criteria texts
UMLS
Sentence boundary detecting
Sentences
Semantic feature extracting
Trial-feature matrix
Clustering
Clusters


Similarity Matrix of four trials A,B,C,D; 1=identical; 0= completely different; our threshold for forming clusters is 0.7

	A	B	C	D
A	1	.8	.9	.6
B	.8	1	.8	.9
C	.9	.8	1	.8
D	.6	.9	.8	1

Fig. 2. Center-based clusters and unique clusters constructed from four example trials.

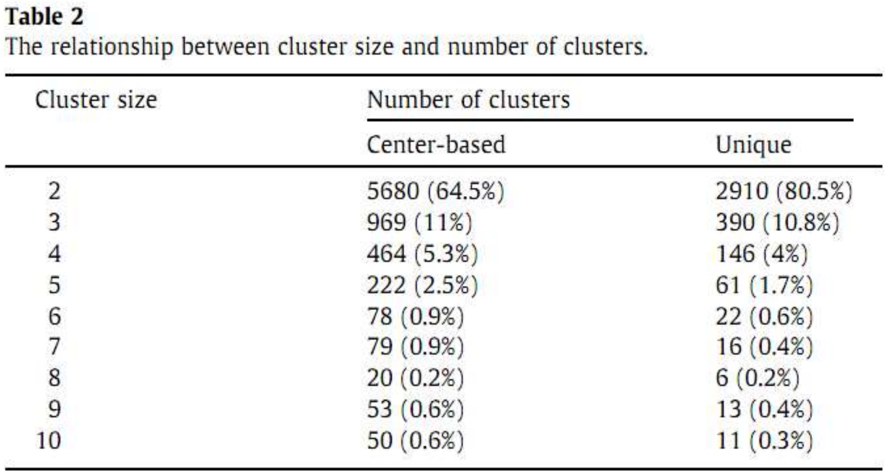


Table 2
The relationship between cluster size and number of clusters.

Cluster size	Number of clusters	
	Center-based	Unique
2	5680 (64.5%)	2910 (80.5%)
3	969 (11%)	390 (10.8%)
4	464 (5.3%)	146 (4%)
5	222 (2.5%)	61 (1.7%)
6	78 (0.9%)	22 (0.6%)
7	79 (0.9%)	16 (0.4%)
8	20 (0.2%)	6 (0.2%)
9	53 (0.6%)	13 (0.4%)
10	50 (0.6%)	11 (0.3%)

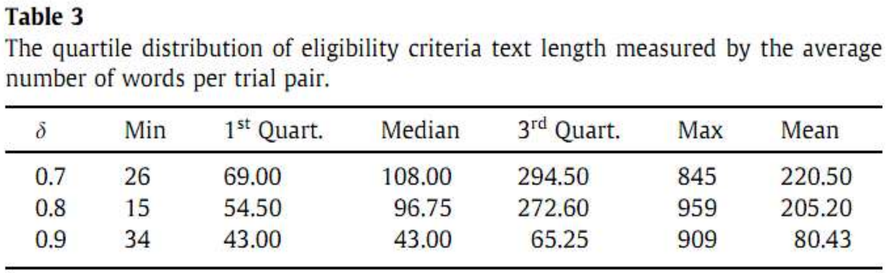


Table 3
The quartile distribution of eligibility criteria text length measured by the average number of words per trial pair.

δ	Min	1st Quart.	Median	3rd Quart.	Max	Mean
0.7	26	69.00	108.00	294.50	845	220.50
0.8	15	54.50	96.75	272.60	959	205.20
0.9	34	43.00	43.00	65.25	909	80.43

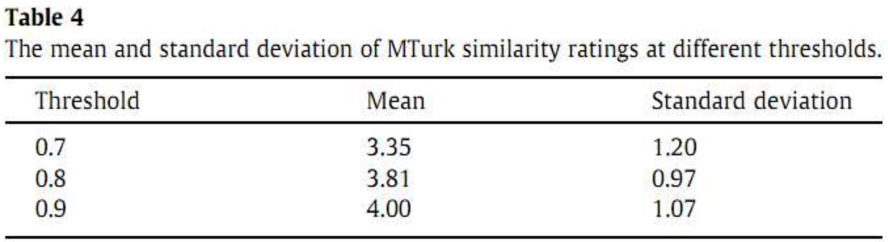
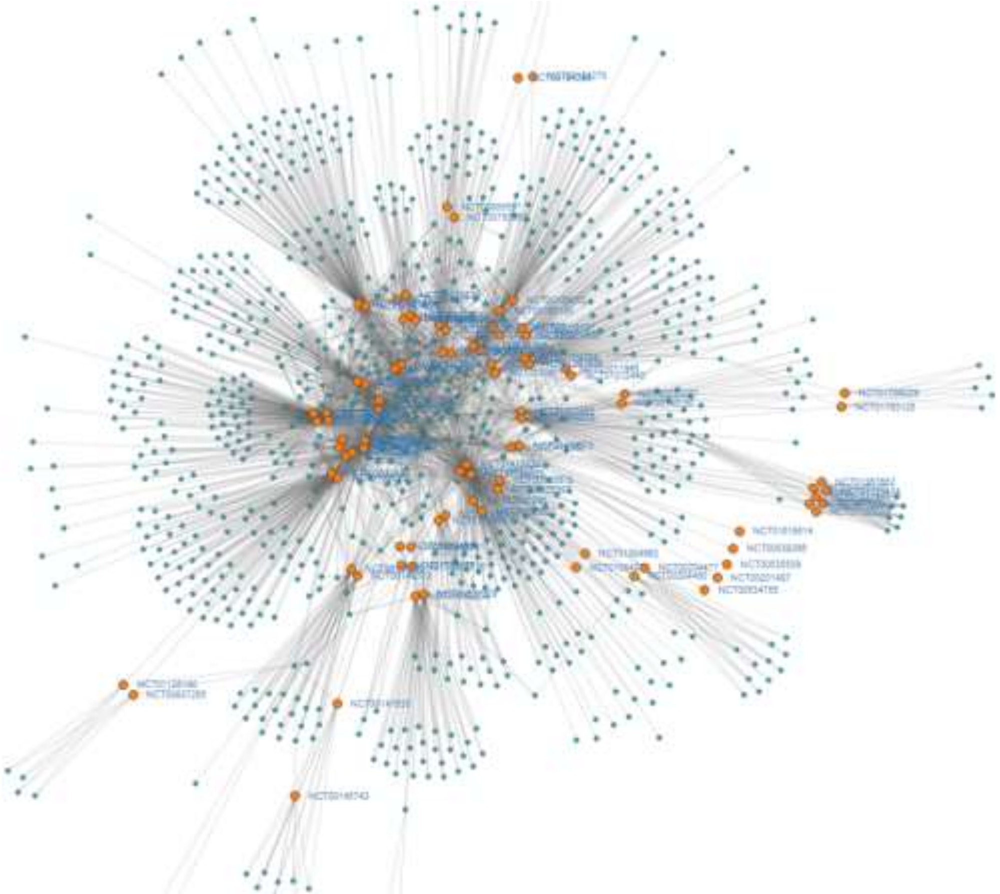


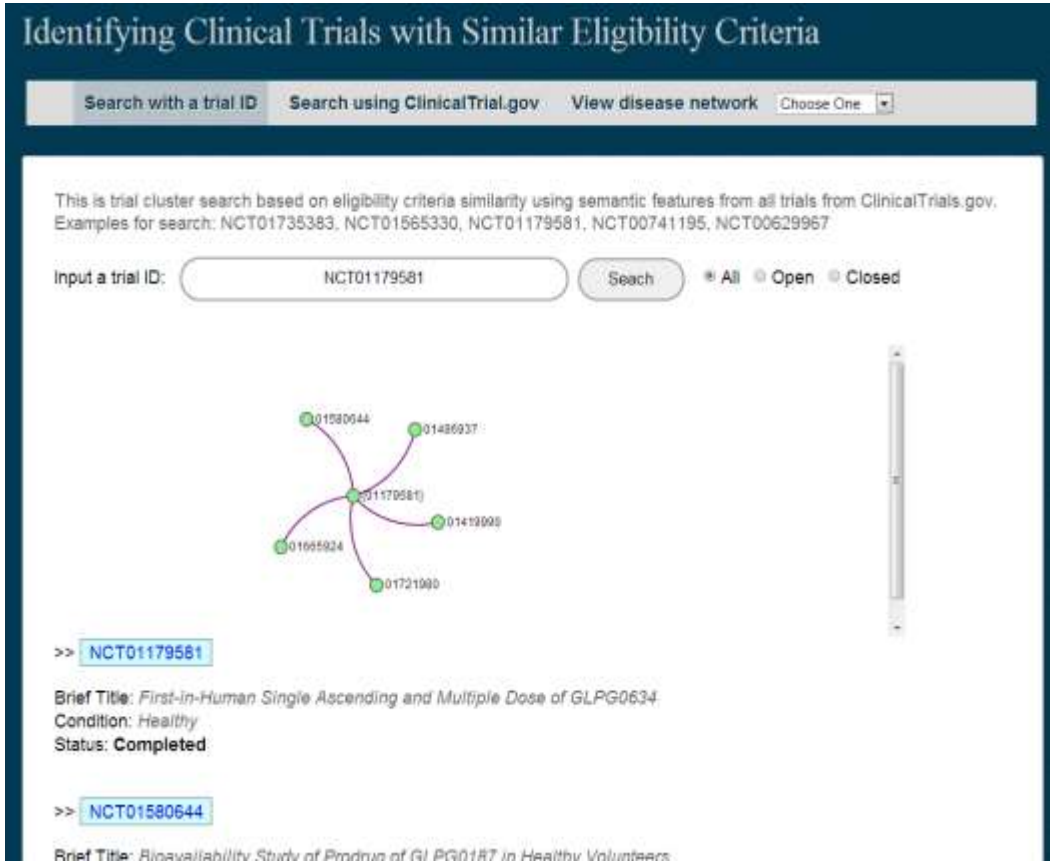
Table 4
The mean and standard deviation of MTurk similarity ratings at different thresholds.

Threshold	Mean	Standard deviation
0.7	3.35	1.20
0.8	3.81	0.97
0.9	4.00	1.07

clustering by similar semantic phenotypes

- Identifying similar semantic phenotypes for 5488 diseases
- **hospitals /researchers**: view trial-phenotypes associations
- **Patients:** a convenient way to retrieve similar trials to attend





Identifying Clinical Trials with Similar Eligibility Criteria

This is trial cluster search based on eligibility criteria similarity using semantic features from all trials from ClinicalTrials.gov.
Examples for search: NCT01735383, NCT01565330, NCT01179581, NCT00741195, NCT00629967

Input a trial ID: NCT01179581 Seach ● All ○ Open ○ Closed

>> NCT01179581

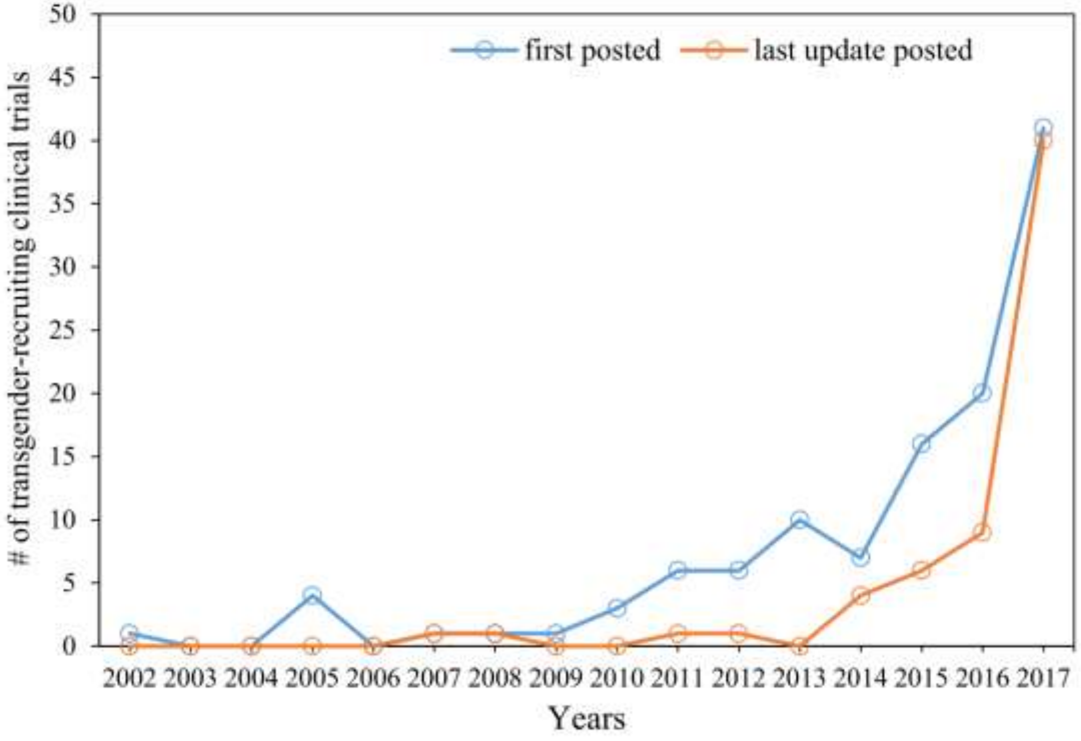
Brief Title: *First-in-Human Single Ascending and Multiple Dose of GLPG0634*
Condition: *Healthy*
Status: **Completed**

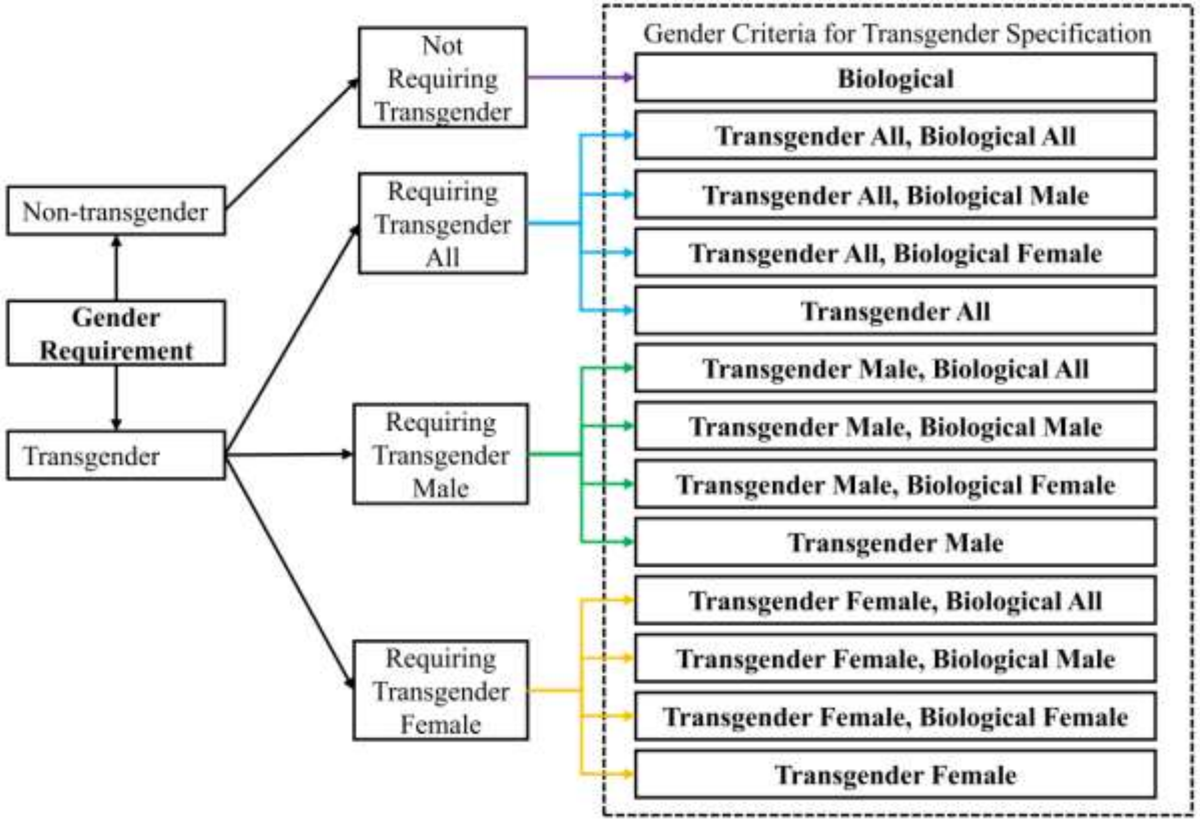
>> NCT01580644

Brief Title: *Bioavailability Study of Prodrug of GLPG0187 in Healthy Volunteers*

Gender Extraction for Enhancing Clinical Trial Recruitment

Motivation





Gender Requirement

Non-transgender

Transgender

Not Requiring Transgender

Requiring Transgender All

Requiring Transgender Male

Requiring Transgender Female

Gender Criteria for Transgender Specification

Biological

Transgender All, Biological All

Transgender All, Biological Male

Transgender All, Biological Female

Transgender All

Transgender Male, Biological All

Transgender Male, Biological Male

Transgender Male, Biological Female

Transgender Male

Transgender Female, Biological All

Transgender Female, Biological Male

Transgender Female, Biological Female

Transgender Female

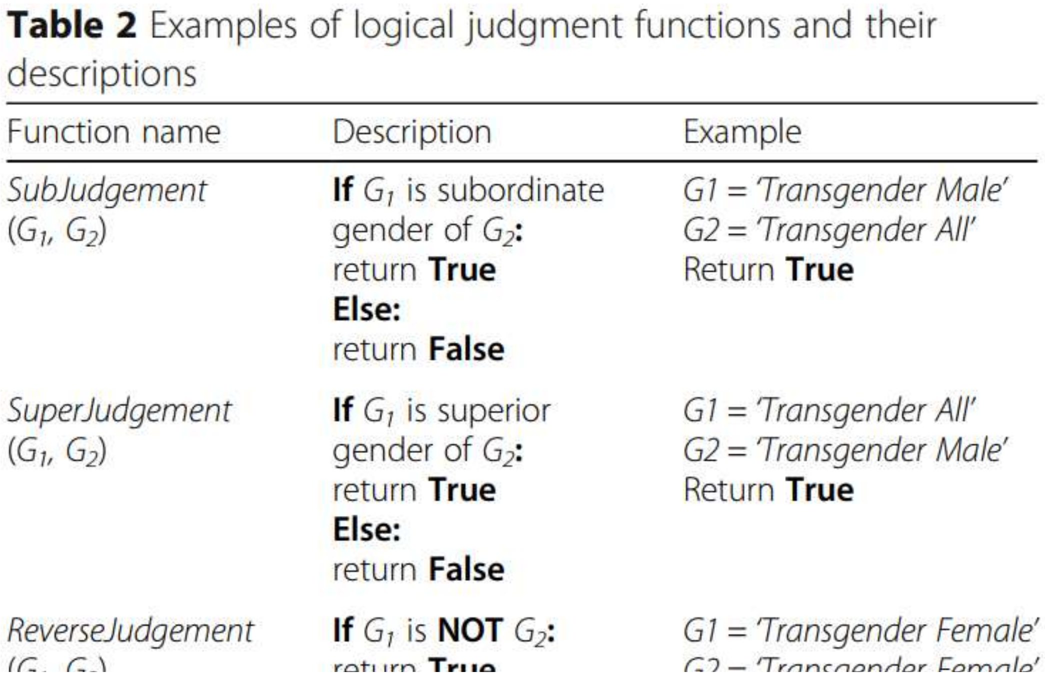


Table 2 Examples of logical judgment functions and their descriptions

Function name	Description	Example

SubJudgement *(G_1, G_2)*	**If** G_1 is subordinate gender of G_2**:**
	return **True**
	Else:
	return **False**
		G1 = 'Transgender Male'
		G2 = 'Transgender All'
		Return **True**

SuperJudgement *(G_1, G_2)*	**If** G_1 is superior gender of G_2**:**
	return **True**
	Else:
	return **False**
		G1 = 'Transgender All'
		G2 = 'Transgender Male'
		Return **True**

ReverseJudgement *(G_1, G_2)*	**If** G_1 is **NOT** G_2**:** return **True**	*G1 = 'Transgender Female'* *G2 = 'Transgender Female'*

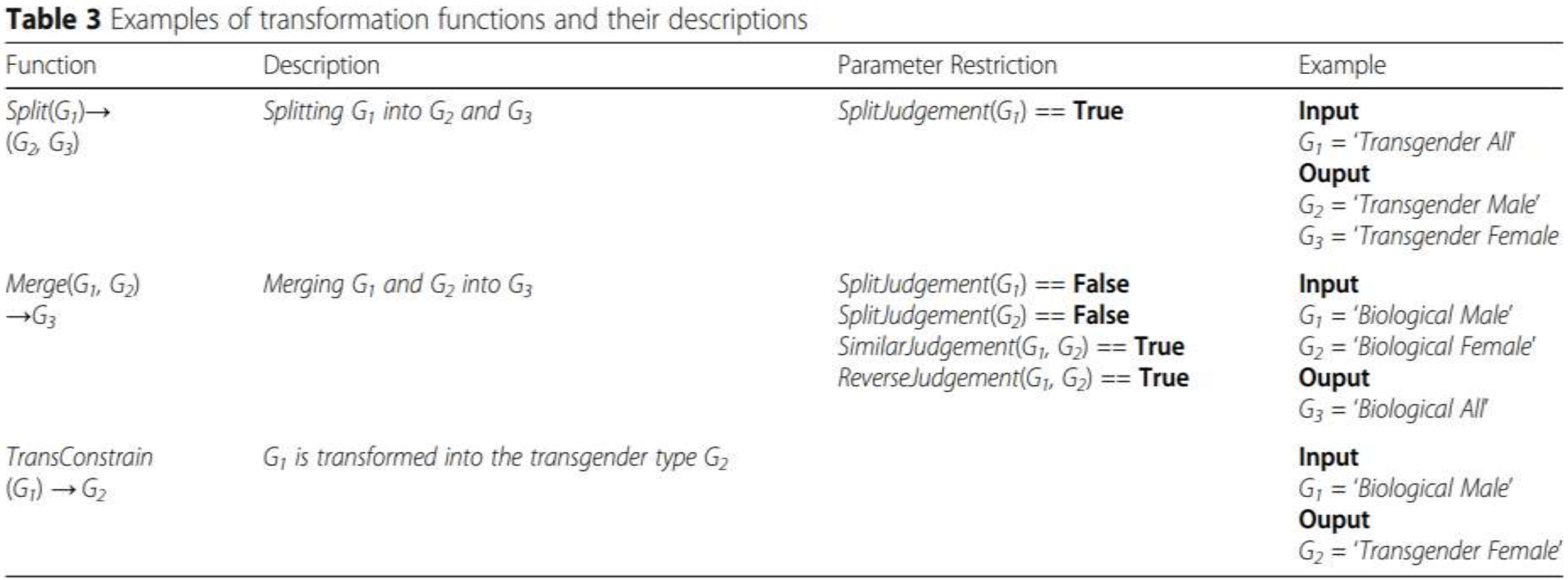
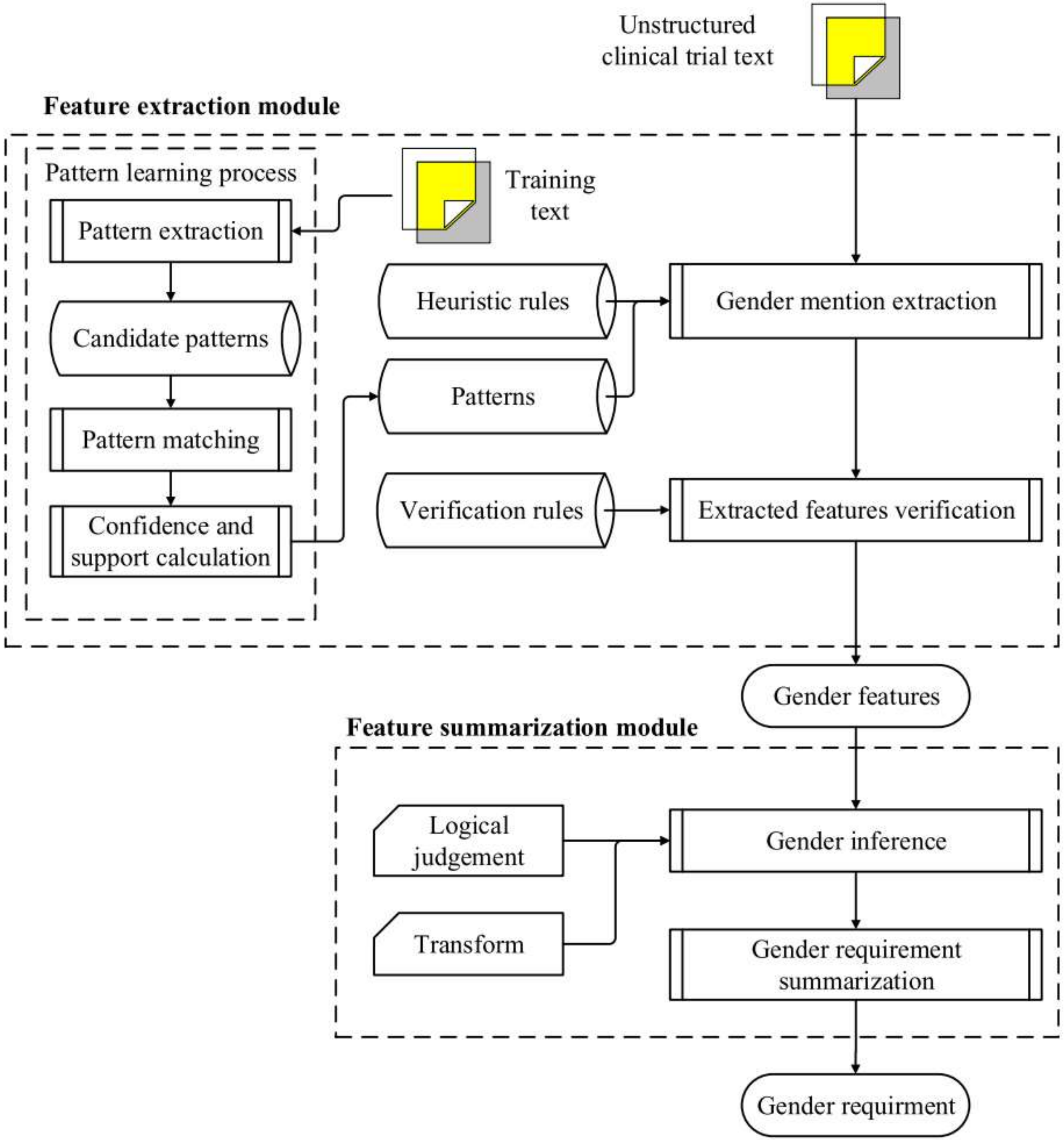
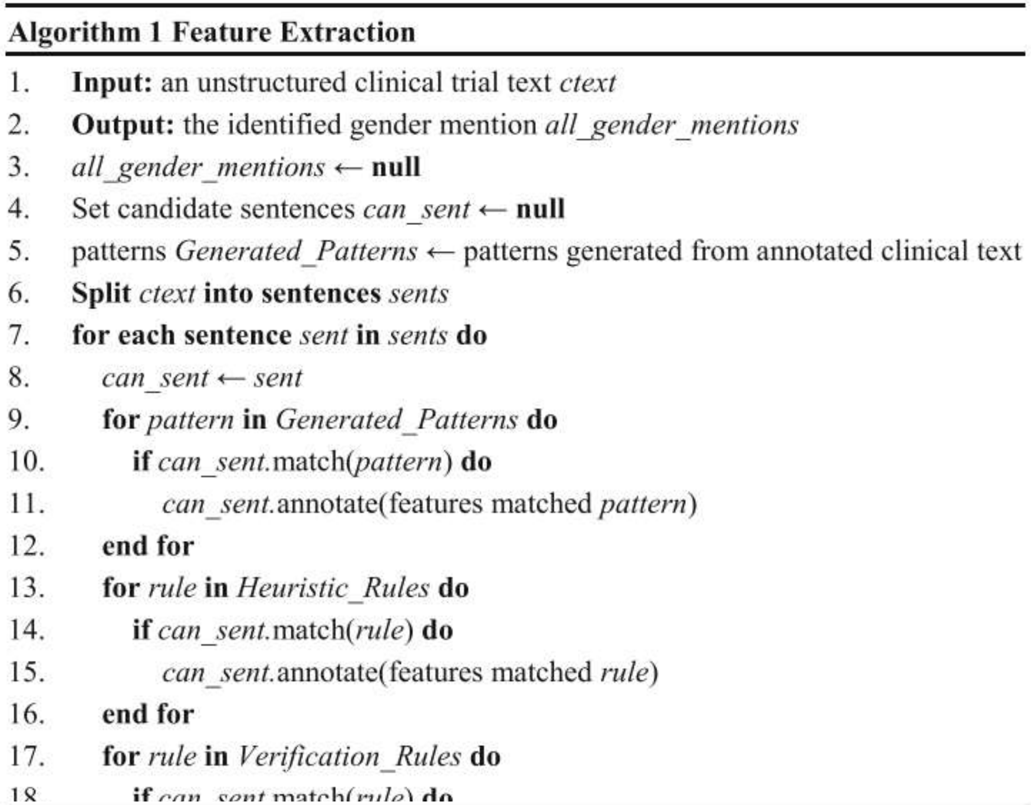


Table 3 Examples of transformation functions and their descriptions

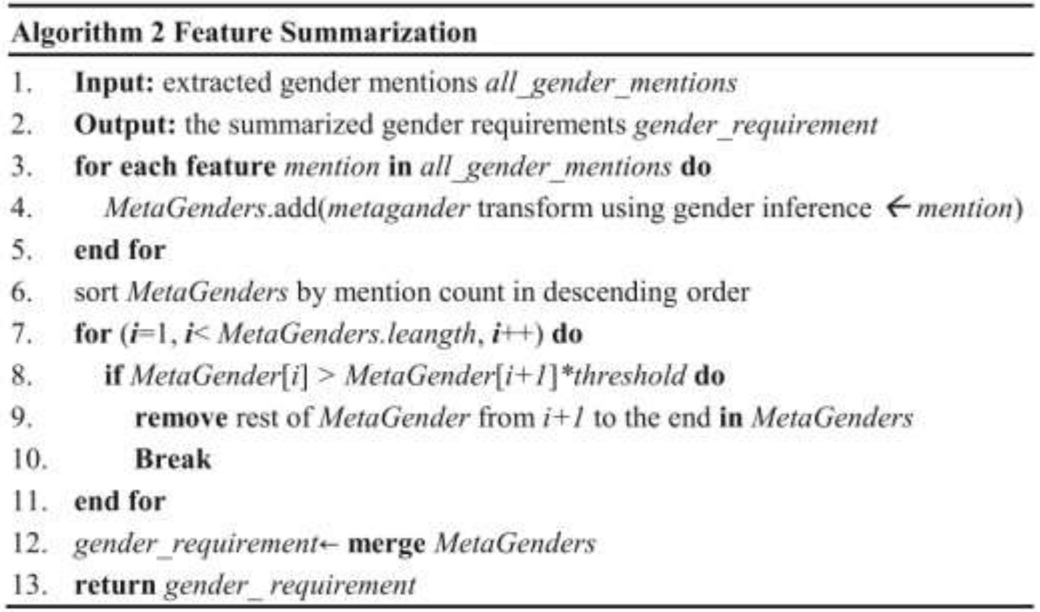
Function	Description	Parameter Restriction	Example
Split(G_1)→ (G_2, G_3)	*Splitting G_1 into G_2 and G_3*	*SplitJudgement(G_1)* == **True**	**Input** *G_1 = 'Transgender All'* **Ouput** *G_2 = 'Transgender Male'* *G_3 = 'Transgender Female*
Merge(G_1, G_2) →G_3	*Merging G_1 and G_2 into G_3*	*SplitJudgement(G_1)* == **False** *SplitJudgement(G_2)* == **False** *SimilarJudgement(G_1, G_2)* == **True** *ReverseJudgement(G_1, G_2)* == **True**	**Input** *G_1 = 'Biological Male'* *G_2 = 'Biological Female'* **Ouput** *G_3 = 'Biological All'*
TransConstrain (G_1) → G_2	*G_1 is transformed into the transgender type G_2*		**Input** *G_1 = 'Biological Male'* **Ouput** *G_2 = 'Transgender Female'*


Unstructured clinical trial text
Feature extraction module
Pattern learning process
Training text
Pattern extraction
Candidate patterns
Pattern matching
Confidence and support calculation
Heuristic rules
Patterns
Verification rules
Gender mention extraction
Extracted features verification
Gender features
Feature summarization module
Logical judgement
Transform
Gender inference
Gender requirement summarization
Gender requirment



Algorithm 1 Feature Extraction

```
Input: an unstructured clinical trial text ctext
Output: the identified gender mention all_gender_mentions
all_gender_mentions ← null
Set candidate sentences can_sent ← null
patterns Generated_Patterns ← patterns generated from annotated clinical text
Split ctext into sentences sents
for each sentence sent in sents do
  can_sent ← sent
  for pattern in Generated_Patterns do
    if can_sent.match(pattern) do
      can_sent.annotate(features matched pattern)
  end for
  for rule in Heuristic_Rules do
    if can_sent.match(rule) do
      can_sent.annotate(features matched rule)
  end for
  for rule in Verification_Rules do
    if can_sent.match(rule) do
```



Algorithm 2 Feature Summarization

```
Input: extracted gender mentions all_gender_mentions
Output: the summarized gender requirements gender_requirement
for each feature mention in all_gender_mentions do
  MetaGenders.add(metagander transform using gender inference ← mention)
end for
sort MetaGenders by mention count in descending order
for (i=1, i< MetaGenders.leangth, i++) do
  if MetaGender[i] > MetaGender[i+1]*threshold do
    remove rest of MetaGender from i+1 to the end in MetaGenders
    Break
end for
gender_requirement← merge MetaGenders
return gender_ requirement
```

Dataset

- 277,012 clinicals trials as dataset

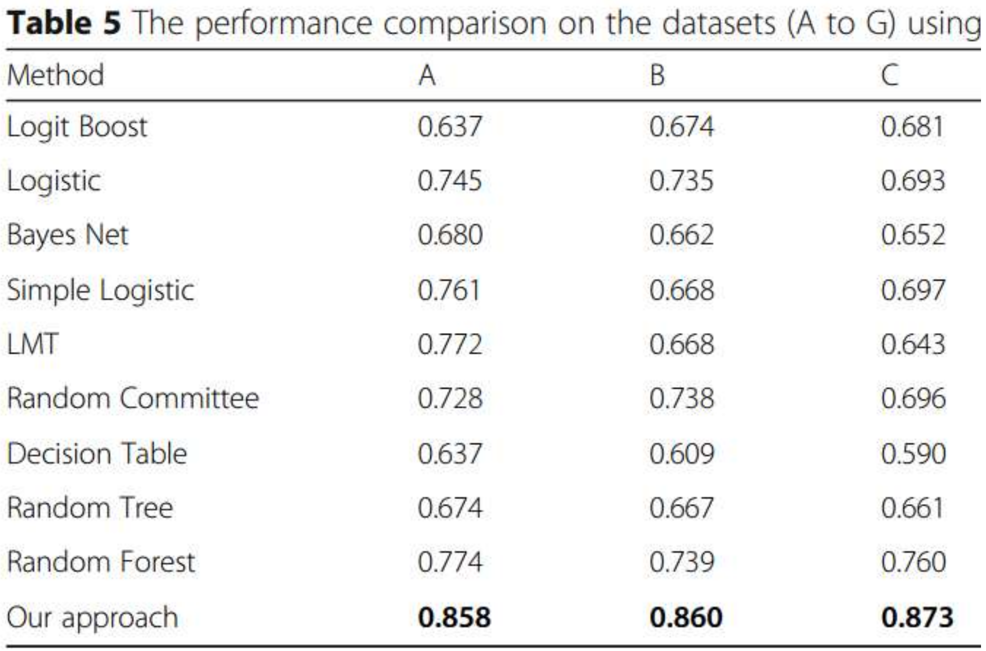
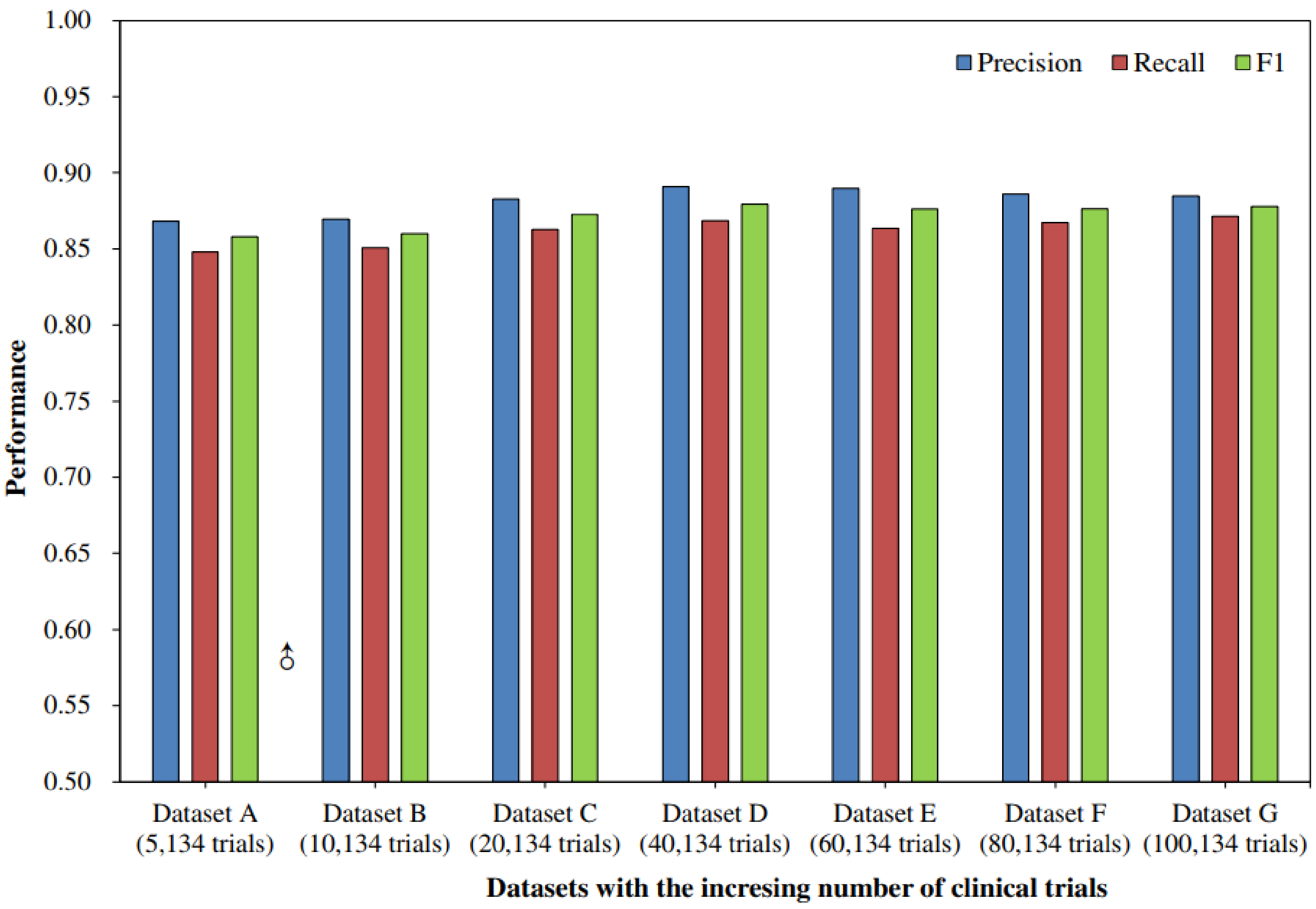
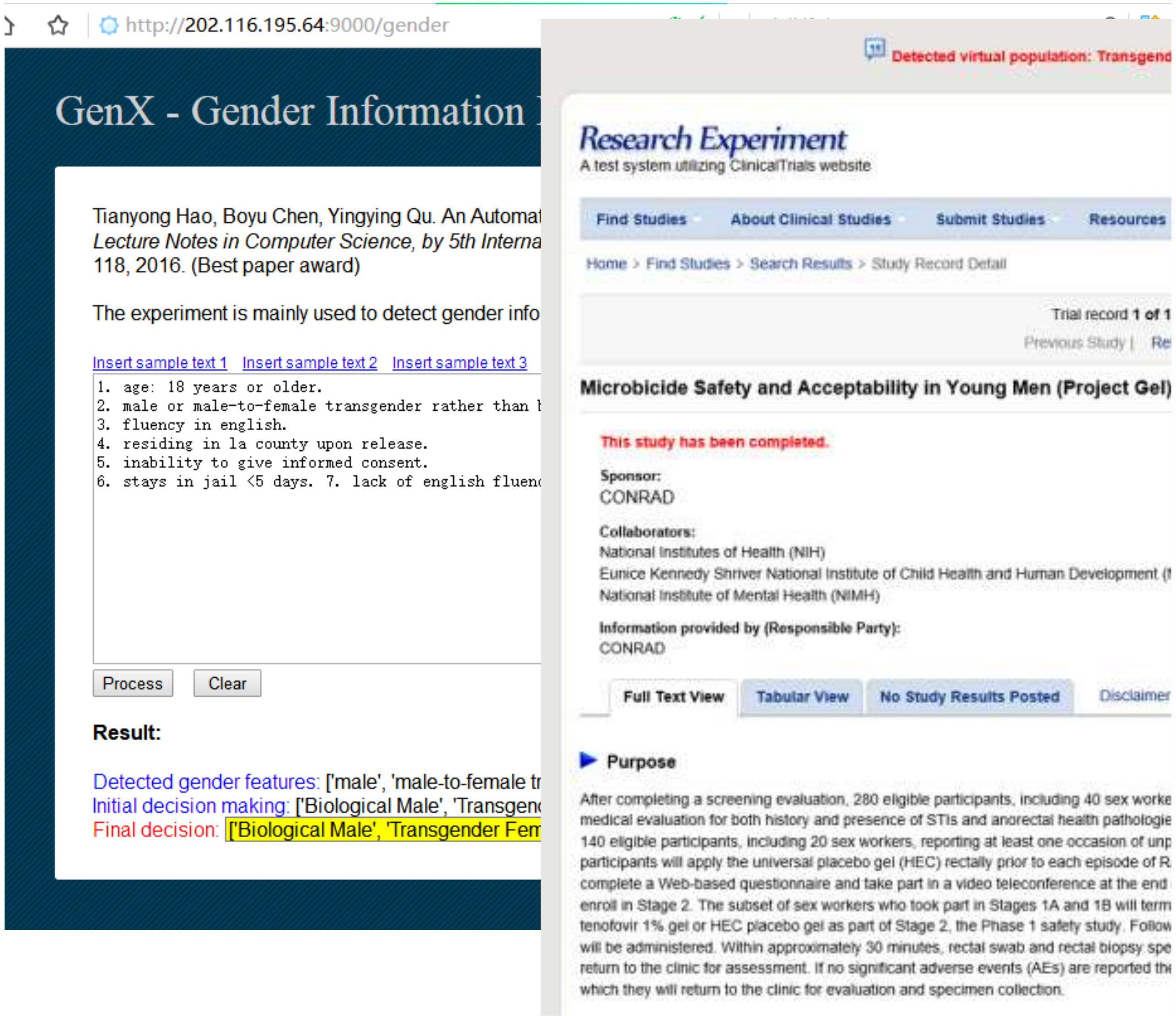


Table 5 The performance comparison on the datasets (A to G) using

Method	A	B	C
Logit Boost	0.637	0.674	0.681
Logistic	0.745	0.735	0.693
Bayes Net	0.680	0.662	0.652
Simple Logistic	0.761	0.668	0.697
LMT	0.772	0.668	0.643
Random Committee	0.728	0.738	0.696
Decision Table	0.637	0.609	0.590
Random Tree	0.674	0.667	0.661
Random Forest	0.774	0.739	0.760
Our approach	**0.858**	**0.860**	**0.873**





http://202.116.195.64:9000/gender

GenX - Gender Information

Tianyong Hao, Boyu Chen, Yingying Qu. An Automat
Lecture Notes in Computer Science, by 5th Interna
118, 2016. (Best paper award)

The experiment is mainly used to detect gender info

```
1. age: 18 years or older.
2. male or male-to-female transgender rather than l
3. fluency in english.
4. residing in la county upon release.
5. inability to give informed consent.
6. stays in jail <5 days. 7. lack of english fluen
```

Result:

Detected gender features: ['male', 'male-to-female tr
Initial decision making: ['Biological Male', 'Transgend
Final decision: ['Biological Male', 'Transgender Fem

Detected virtual population: Transgend

Research Experiment
A test system utilizing ClinicalTrials website

Trial record 1 of 1
Previous Study | Re

Microbicide Safety and Acceptability in Young Men (Project Gel)

This study has been completed.

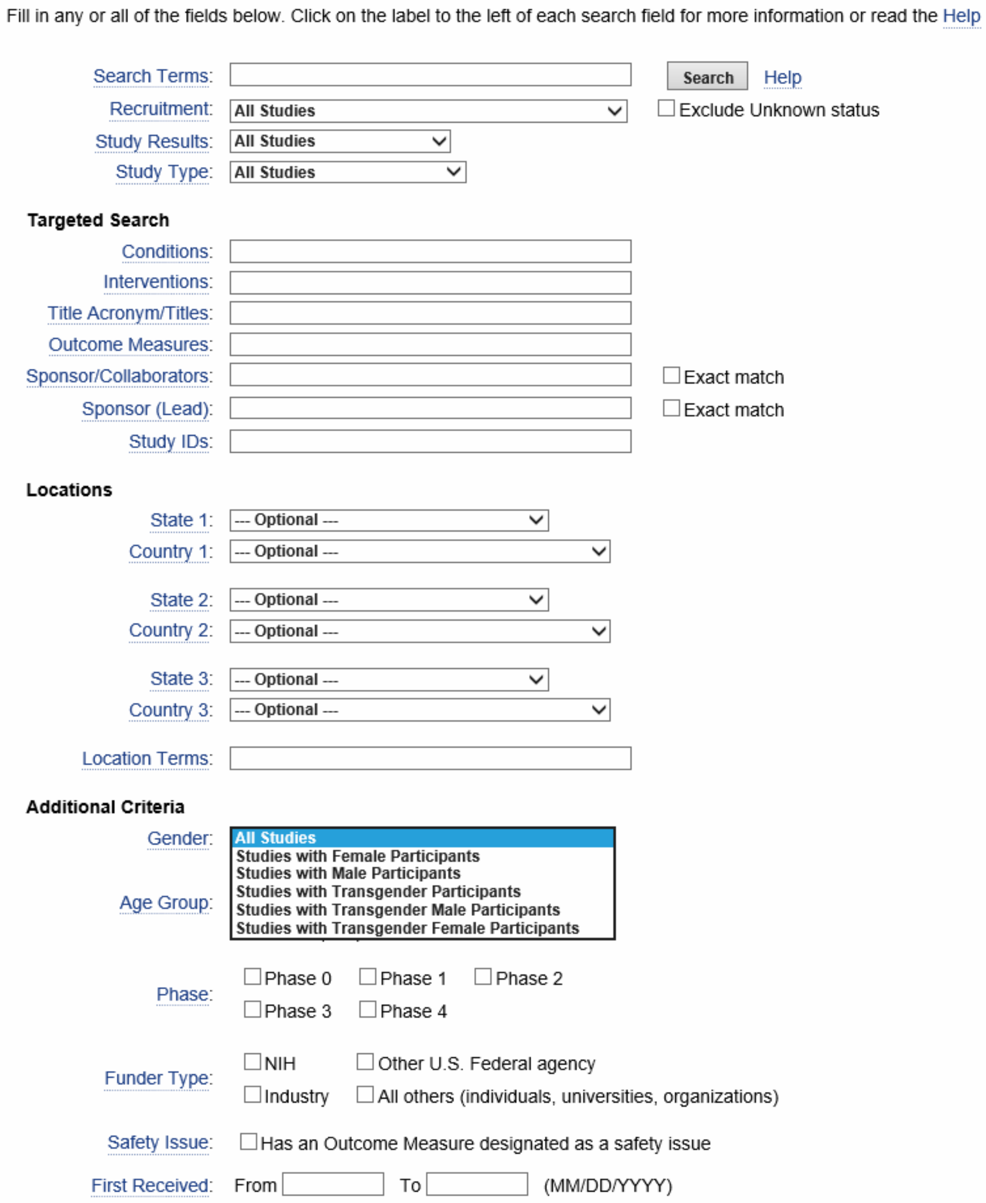
Sponsor:
CONRAD

Collaborators:
National Institutes of Health (NIH)
Eunice Kennedy Shriver National Institute of Child Health and Human Development (
National Institute of Mental Health (NIMH)

Information provided by (Responsible Party):
CONRAD

Purpose

After completing a screening evaluation, 280 eligible participants, including 40 sex worke
medical evaluation for both history and presence of STIs and anorectal health pathologie
140 eligible participants, including 20 sex workers, reporting at least one occasion of unp
participants will apply the universal placebo gel (HEC) rectally prior to each episode of R
complete a Web-based questionnaire and take part in a video teleconference at the end
enroll in Stage 2. The subset of sex workers who took part in Stages 1A and 1B will term
tenofovir 1% gel or HEC placebo gel as part of Stage 2, the Phase 1 safety study. Follow
will be administered. Within approximately 30 minutes, rectal swab and rectal biopsy spe
return to the clinic for assessment. If no significant adverse events (AEs) are reported the
which they will return to the clinic for evaluation and specimen collection.



Fill in any or all of the fields below. Click on the label to the left of each search field for more information or read the Help

Search Terms:

Recruitment: All Studies ☐ Exclude Unknown status

Study Results: All Studies

Study Type: All Studies

Targeted Search

Conditions:

Interventions:

Title Acronym/Titles:

Outcome Measures:

Sponsor/Collaborators: ☐ Exact match

Sponsor (Lead): ☐ Exact match

Study IDs:

Locations

State 1: --- Optional ---

Country 1: --- Optional ---

State 2: --- Optional ---

Country 2: --- Optional ---

State 3: --- Optional ---

Country 3: --- Optional ---

Location Terms:

Additional Criteria

Gender: All Studies
Studies with Female Participants
Studies with Male Participants
Studies with Transgender Participants
Studies with Transgender Male Participants
Studies with Transgender Female Participants

Age Group:

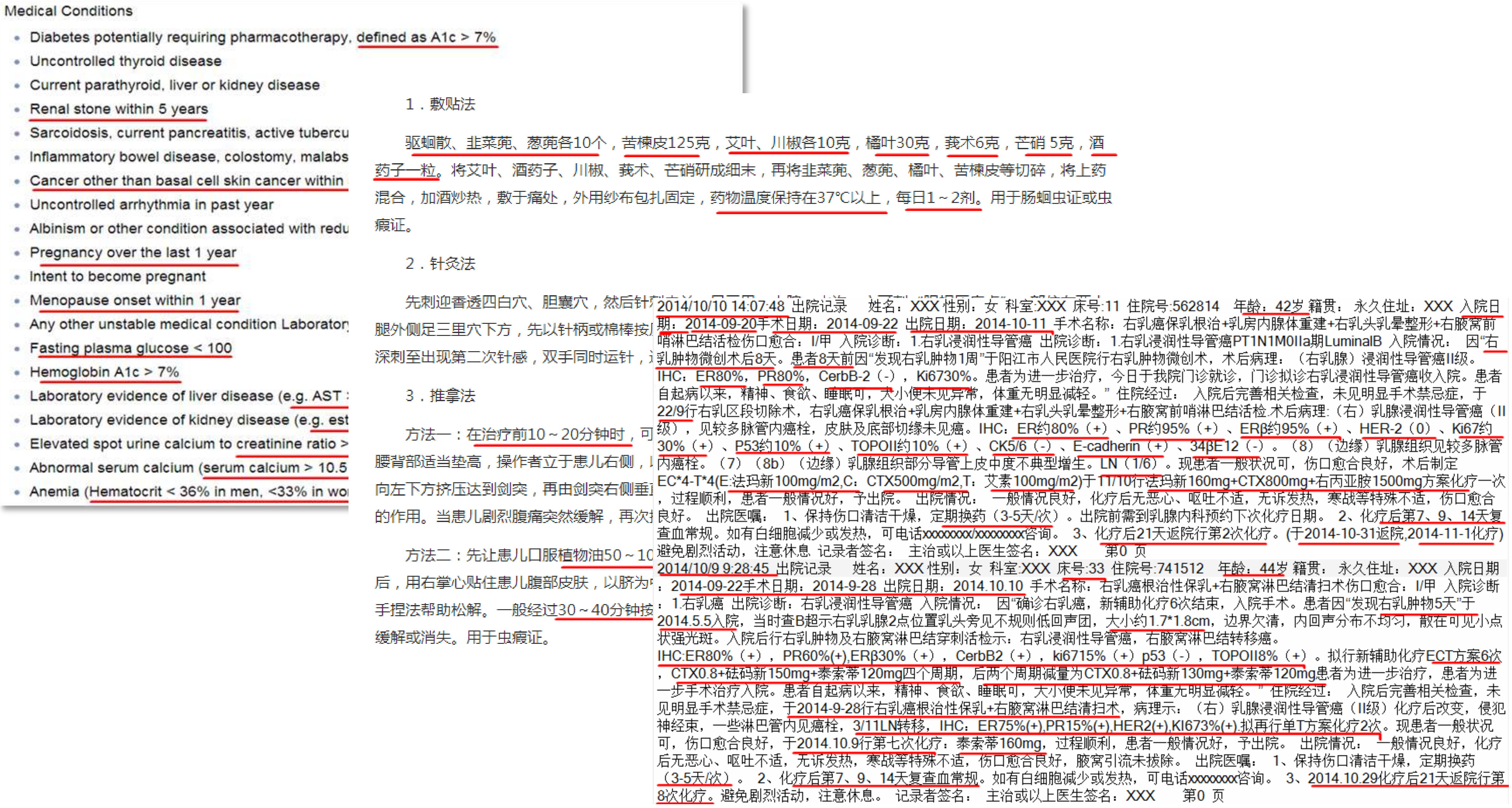
Phase: ☐ Phase 0 ☐ Phase 1 ☐ Phase 2
☐ Phase 3 ☐ Phase 4

Funder Type: ☐ NIH ☐ Other U.S. Federal agency
☐ Industry ☐ All others (individuals, universities, organizations)

Safety Issue: ☐ Has an Outcome Measure designated as a safety issue

First Received: From To (MM/DD/YYYY)

Measurable Quantitative Information Representation and Extraction



Medical Conditions

- Diabetes potentially requiring pharmacotherapy, defined as A1c > 7%
- Uncontrolled thyroid disease
- Current parathyroid, liver or kidney disease
- Renal stone within 5 years
- Sarcoidosis, current pancreatitis, active tubercu
- Inflammatory bowel disease, colostomy, malabs
- Cancer other than basal cell skin cancer within
- Uncontrolled arrhythmia in past year
- Albinism or other condition associated with redu
- Pregnancy over the last 1 year
- Intent to become pregnant
- Menopause onset within 1 year
- Any other unstable medical condition Laborator
- Fasting plasma glucose < 100
- Hemoglobin A1c > 7%
- Laboratory evidence of liver disease (e.g. AST
- Laboratory evidence of kidney disease (e.g. est
- Elevated spot urine calcium to creatinine ratio >
- Abnormal serum calcium (serum calcium > 10.5
- Anemia (Hematocrit < 36% in men, <33% in wo

1．敷贴法

驱蛔散、韭菜蔸、葱蔸各10个，苦楝皮125克，艾叶、川椒各10克，橘叶30克，莪术6克，芒硝 5克，酒药子一粒。将艾叶、酒药子、川椒、莪术、芒硝研成细末，再将韭菜蔸、葱蔸、橘叶、苦楝皮等切碎，将上药混合，加酒炒热，敷于痛处，外用纱布包扎固定，药物温度保持在37℃以上，每日1～2剂。用于肠蛔虫证或虫瘕证。

2．针灸法

先刺迎香透四白穴、胆囊穴，然后针
腿外侧足三里穴下方，先以针柄或棉棒按
深刺至出现第二次针感，双手同时运针，

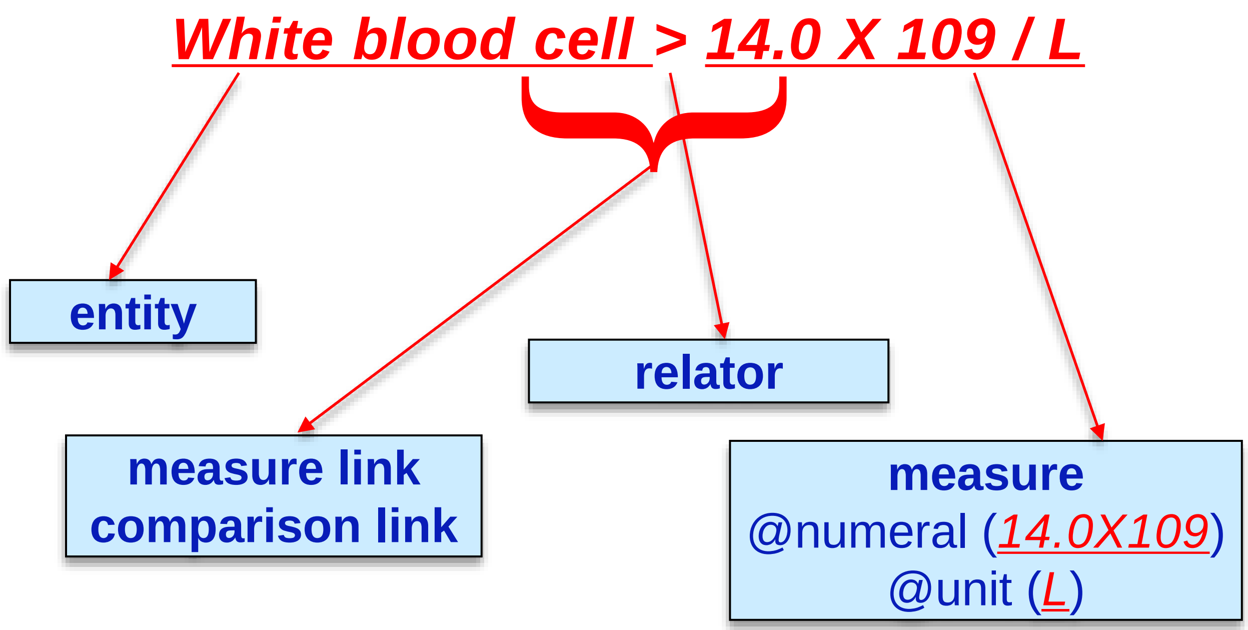
3．推拿法

方法一：在治疗前10～20分钟时，可
腰背部适当垫高，操作者立于患儿右侧，
向左下方挤压达到剑突，再由剑突右侧垂
的作用。当患儿剧烈腹痛突然缓解，再次

方法二：先让患儿口服植物油50～10
后，用右掌心贴住患儿腹部皮肤，以脐为
手捏法帮助松解。一般经过30～40分钟按
缓解或消失。用于虫瘕证。

2014/10/10 14:07:48 出院记录　姓名：XXX 性别：女 科室:XXX 床号:11 住院号:562814　年龄：42岁 籍贯：永久住址：XXX 入院日期：2014-09-20手术日期：2014-09-22 出院日期：2014-10-11 手术名称：右乳癌保乳根治+乳房内腺体重建+右乳头乳晕整形+右腋窝前哨淋巴结活检伤口愈合：I/甲 入院诊断：1.右乳浸润性导管癌 出院诊断：1.右乳浸润性导管癌PT1N1M0IIa期LuminalB 入院情况：因"右乳肿物微创术后8天。患者8天前因"发现右乳肿物1周"于阳江市人民医院行右乳肿物微创术，术后病理：（右乳腺）浸润性导管癌II级。IHC：ER80%，PR80%，CerbB-2（-），Ki6730%。患者为进一步治疗，今日于我院门诊就诊，门诊拟诊右乳浸润性导管癌收入院。患者自起病以来，精神、食欲、睡眠可，大小便未见异常，体重无明显减轻。" 住院经过：入院后完善相关检查，未见明显手术禁忌症，于22/9行右乳区段切除术，右乳癌保乳根治+乳房内腺体重建+右乳头乳晕整形+右腋窝前哨淋巴结活检.术后病理:（右）乳腺浸润性导管癌（II级），见较多脉管内癌栓，皮肤及底部切缘未见癌。IHC：ER约80%（+）、PR约95%（+）、ERβ约95%（+）、HER-2（0）、Ki67约30%（+）、P53约10%（+）、TOPOII约10%（+）、CK5/6（-）、E-cadherin（+）、34βE12（-）。（8）（边缘）乳腺组织见较多脉管内癌栓。（7）（8b）（边缘）乳腺组织部分导管上皮中度不典型增生。LN（1/6）。现患者一般状况可，伤口愈合良好，术后制定EC*4-T*4(E:法玛新100mg/m2,C：CTX500mg/m2,T：艾素100mg/m2)于11/10行法玛新160mg+CTX800mg+右丙亚胺1500mg方案化疗一次，过程顺利，患者一般情况好，予出院。 出院情况：一般情况良好，化疗后无恶心、呕吐不适，无诉发热，寒战等特殊不适，伤口愈合良好。 出院医嘱：1、保持伤口清洁干燥，定期换药（3-5天/次）。出院前需到乳腺内科预约下次化疗日期。 2、化疗后第7、9、14天复查血常规。如有白细胞减少或发热，可电话xxxxxxxx/xxxxxxxx咨询。 3、化疗后21天返院行第2次化疗。(于2014-10-31返院,2014-11-1化疗)避免剧烈活动，注意休息 记录者签名：　主治或以上医生签名：XXX　　第0 页

2014/10/9 9:28:45 出院记录　姓名：XXX 性别：女 科室:XXX 床号:33 住院号:741512　年龄：44岁 籍贯：永久住址：XXX 入院日期：2014-09-22手术日期：2014-9-28 出院日期：2014.10.10 手术名称：右乳癌根治性保乳+右腋窝淋巴结清扫术伤口愈合：I/甲 入院诊断：1.右乳癌 出院诊断：右乳浸润性导管癌 入院情况：因"确诊右乳癌，新辅助化疗6次结束，入院手术。患者因"发现右乳肿物5天"于2014.5.5入院，当时查B超示右乳乳腺2点位置乳头旁见不规则低回声团，大小约1.7*1.8cm，边界欠清，内回声分布不均匀，散在可见小点状强光斑。入院后行右乳肿物及右腋窝淋巴结穿刺活检示：右乳浸润性导管癌，右腋窝淋巴结转移癌。
IHC:ER80%（+），PR60%(+),ERβ30%（+），CerbB2（+），ki6715%（+）p53（-），TOPOII8%（+）。拟行新辅助化疗ECT方案6次，CTX0.8+砝码新150mg+泰素蒂120mg四个周期，后两个周期减量为CTX0.8+砝码新130mg+泰素蒂120mg患者为进一步治疗，患者为进一步手术治疗入院。患者自起病以来，精神、食欲、睡眠可，大小便未见异常，体重无明显减轻。" 住院经过：入院后完善相关检查，未见明显手术禁忌症，于2014-9-28行右乳癌根治性保乳+右腋窝淋巴结清扫术，病理示：（右）乳腺浸润性导管癌（II级）化疗后改变，侵犯神经束，一些淋巴管内见癌栓，3/11LN转移，IHC：ER75%(+),PR15%(+),HER2(+),KI673%(+).拟再行单T方案化疗2次。现患者一般状况可，伤口愈合良好，于2014.10.9行第七次化疗：泰素蒂160mg，过程顺利，患者一般情况好，予出院。 出院情况：一般情况良好，化疗后无恶心、呕吐不适，无诉发热，寒战等特殊不适，伤口愈合良好，腋窝引流未拔除。 出院医嘱：1、保持伤口清洁干燥，定期换药（3-5天/次）。 2、化疗后第7、9、14天复查血常规。如有白细胞减少或发热，可电话xxxxxxxx咨询。 3、2014.10.29化疗后21天返院行第8次化疗。避免剧烈活动，注意休息。 记录者签名：　主治或以上医生签名：XXX　　第0 页



White blood cell > 14.0 X 109 / L

entity

relator

measure link
comparison link

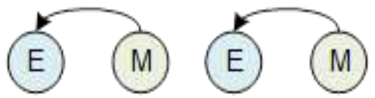
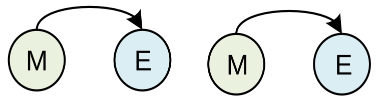
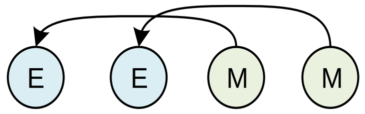
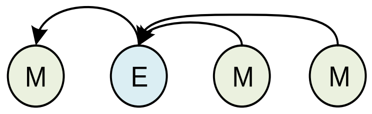
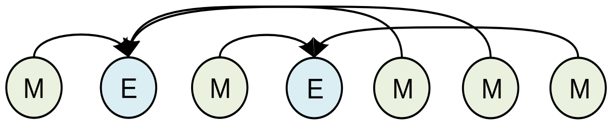
measure
@numeral (14.0X109)
@unit (L)

① “*White blood cell*” describes an **entity**.

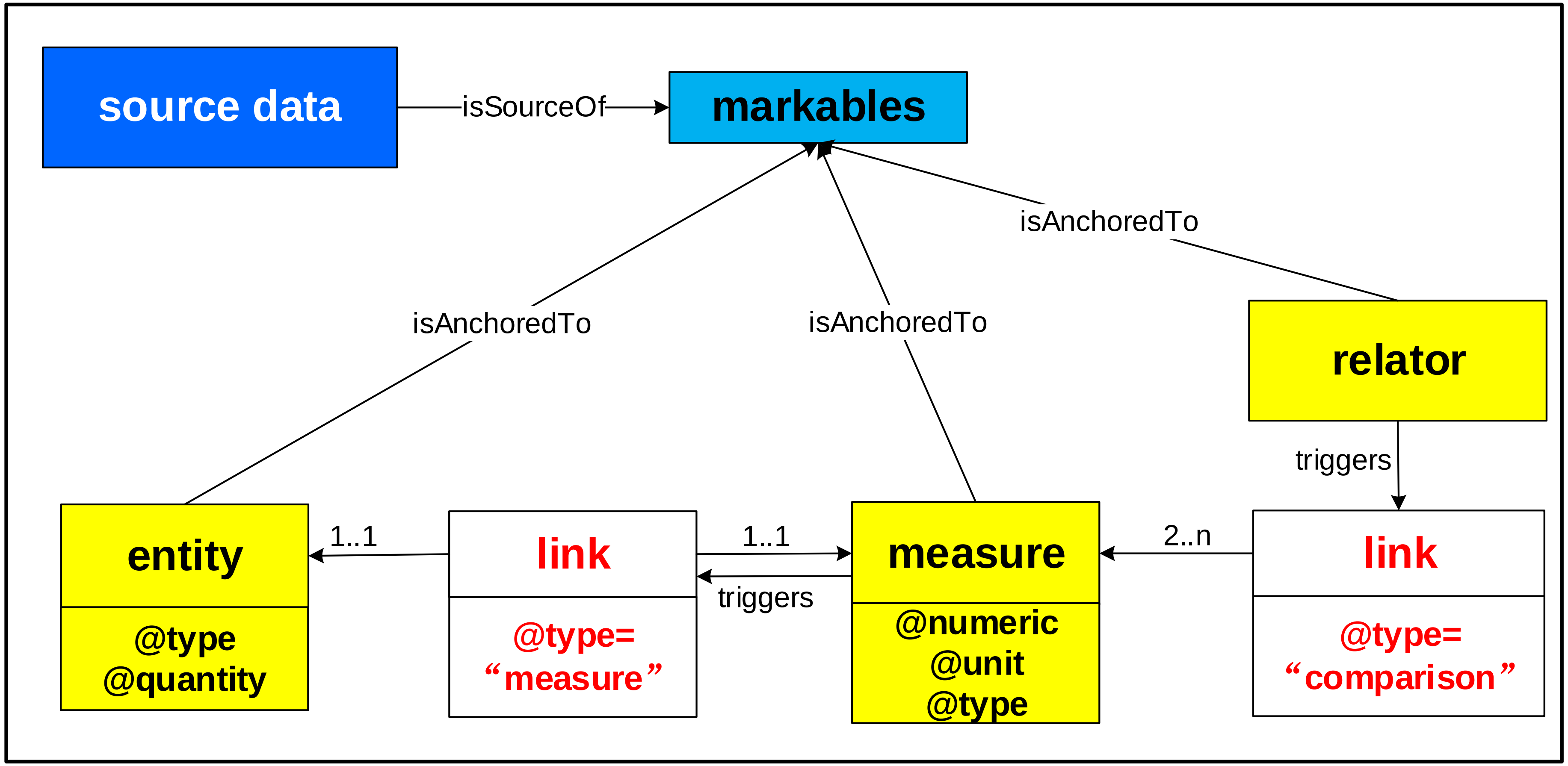
② “*14.0 X 109 / L*” describes a **measure** consisting of two attributes @numeral (“*14.0 X 109*”) and @unit (“*L*”).

③ “>” describes a **relator** relation (“larger than”).

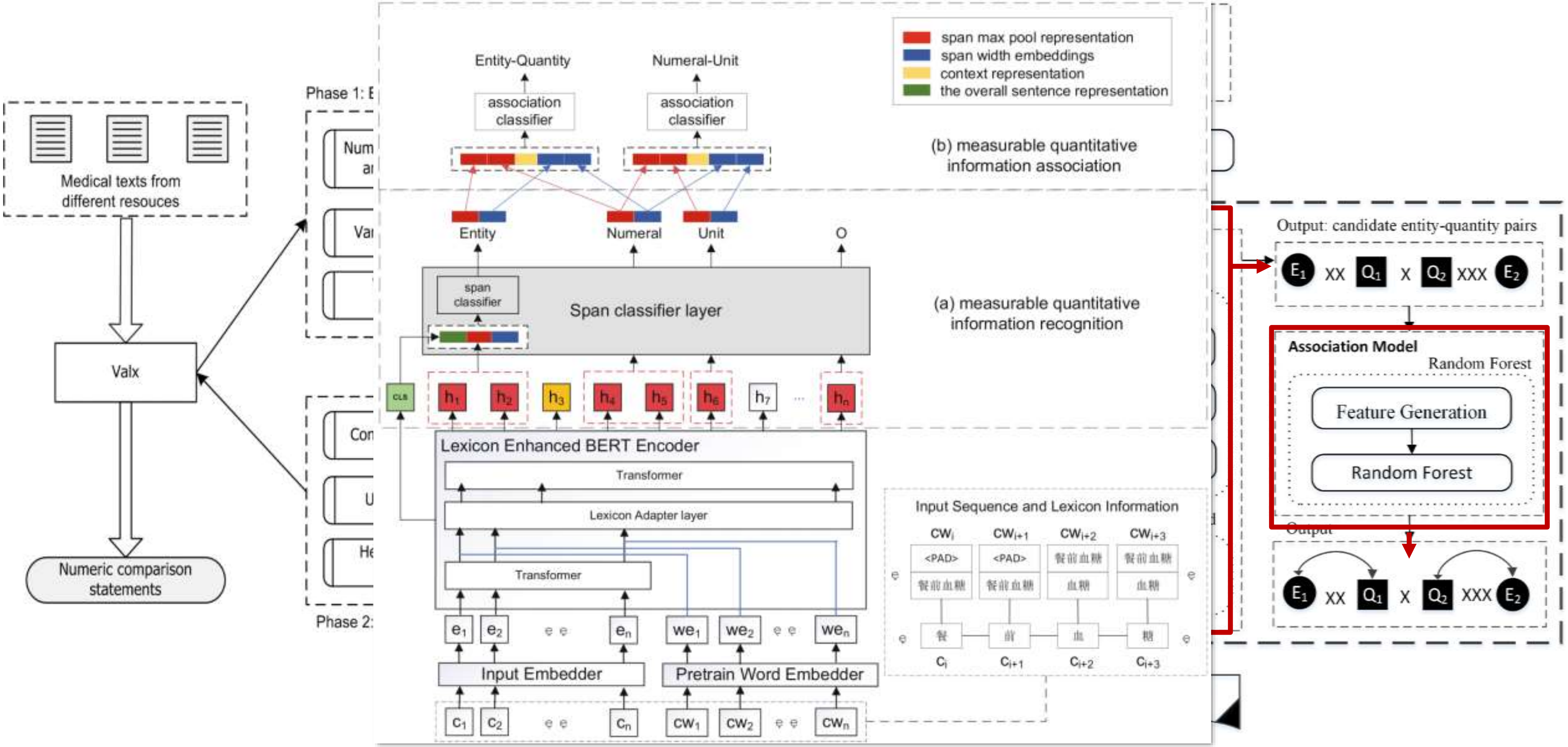
④ A **measure link** and a **comparison link** are triggered by the measure and by the relator, respectively.

Link type	Visual demonstration
Sequential	
Inverted – sequential	
Cross	
Multiple -cross	
Combination	

The metamodel (ISO 24617-11)



Method design



Evaluation design

s disease

c hospital

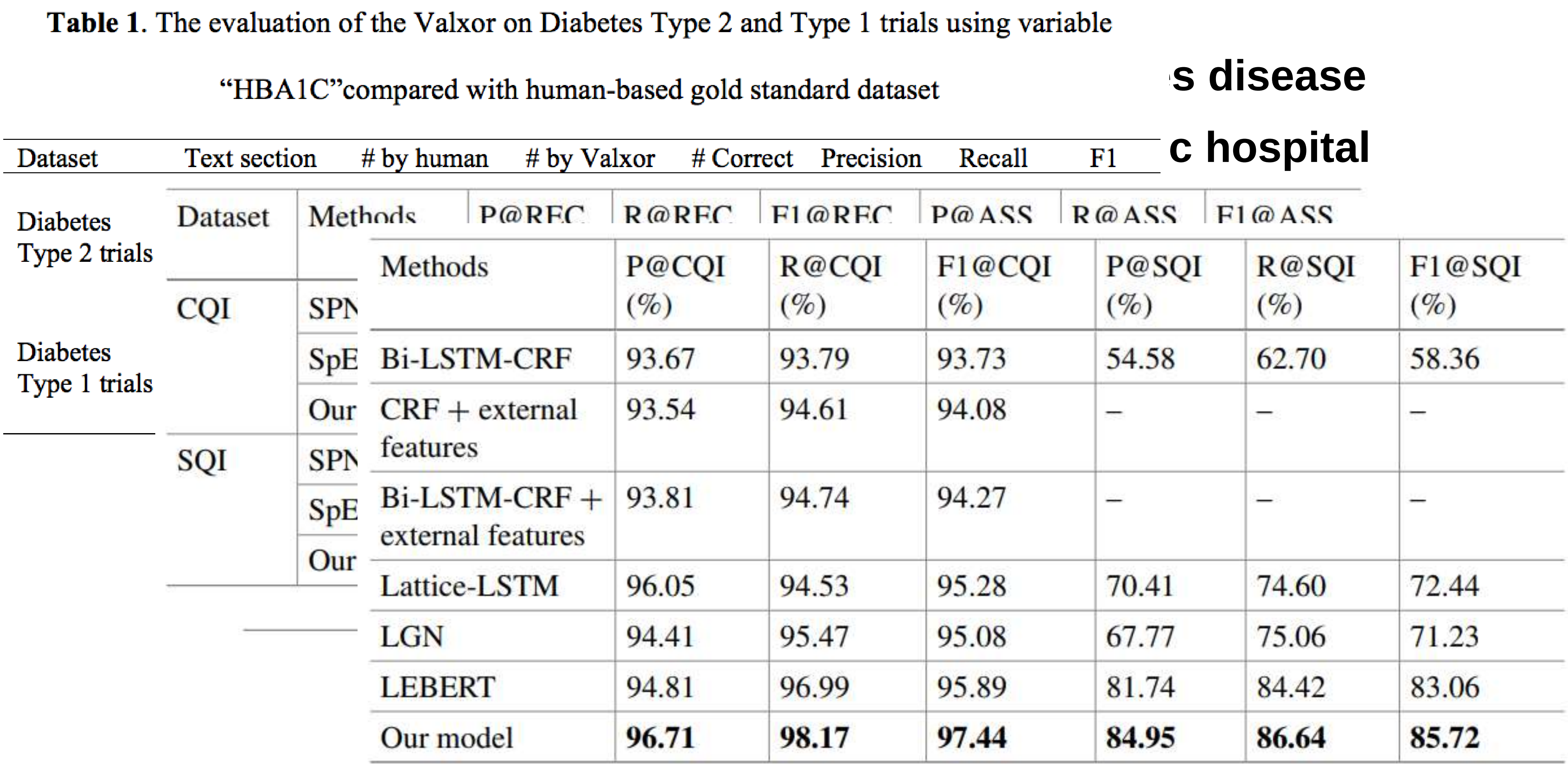


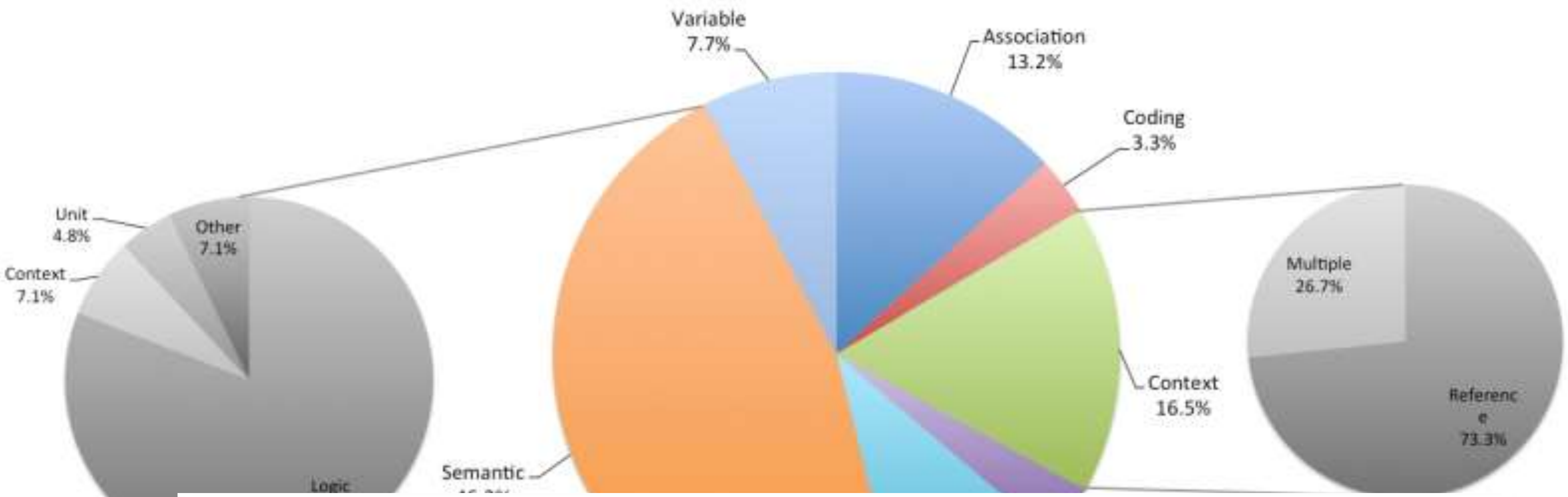
Table 1. The evaluation of the Valxor on Diabetes Type 2 and Type 1 trials using variable "HBA1C"compared with human-based gold standard dataset

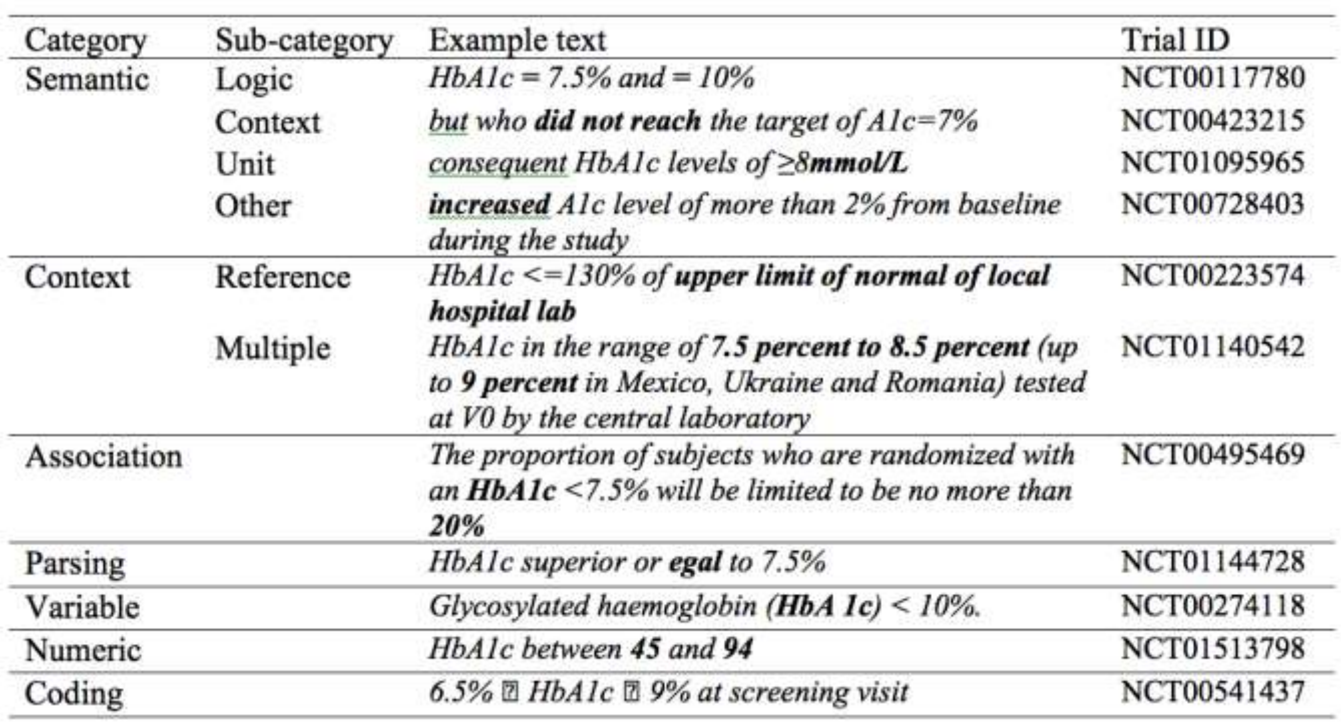
Dataset	Text section	# by human	# by Valxor	# Correct	Precision	Recall	F1
Diabetes Type 2 trials							
Diabetes Type 1 trials							

Dataset	Methods	P@REC	R@REC	F1@REC	P@ASS	R@ASS	F1@ASS
CQI	SPN						
	SpE						
	Our						
SQI	SPN						
	SpE						
	Our						

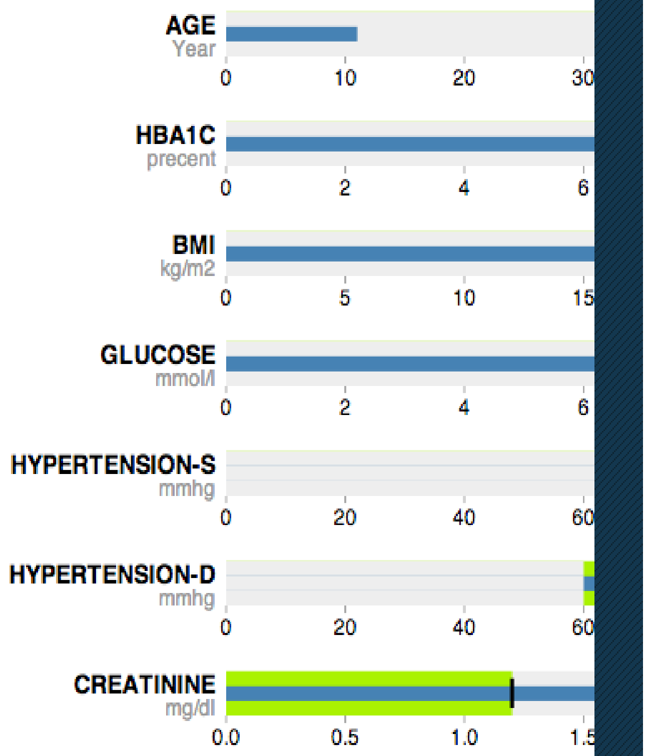
Methods	P@CQI (%)	R@CQI (%)	F1@CQI (%)	P@SQI (%)	R@SQI (%)	F1@SQI (%)
Bi-LSTM-CRF	93.67	93.79	93.73	54.58	62.70	58.36
CRF + external features	93.54	94.61	94.08	–	–	–
Bi-LSTM-CRF + external features	93.81	94.74	94.27	–	–	–
Lattice-LSTM	96.05	94.53	95.28	70.41	74.60	72.44
LGN	94.41	95.47	95.08	67.77	75.06	71.23
LEBERT	94.81	96.99	95.89	81.74	84.42	83.06
Our model	**96.71**	**98.17**	**97.44**	**84.95**	**86.64**	**85.72**

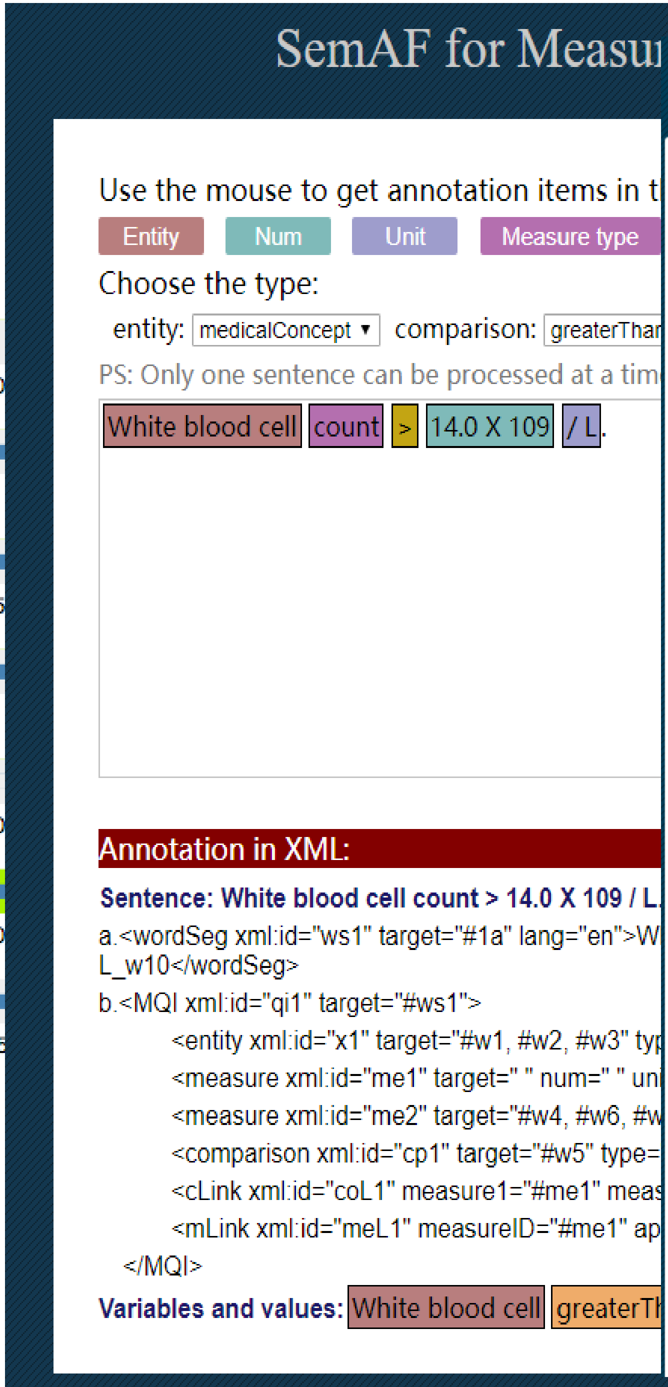
Error analysis





Category	Sub-category	Example text	Trial ID
Semantic	Logic	*HbA1c = 7.5% and = 10%*	NCT00117780
	Context	*but who **did not reach** the target of A1c=7%*	NCT00423215
	Unit	*consequent HbA1c levels of ≥**8mmol/L***	NCT01095965
	Other	***increased** A1c level of more than 2% from baseline during the study*	NCT00728403
Context	Reference	*HbA1c <=130% of **upper limit of normal of local hospital lab***	NCT00223574
	Multiple	*HbA1c in the range of **7.5 percent to 8.5 percent** (up to **9 percent** in Mexico, Ukraine and Romania) tested at V0 by the central laboratory*	NCT01140542
Association		*The proportion of subjects who are randomized with an **HbA1c** <7.5% will be limited to be no more than **20%***	NCT00495469
Parsing		*HbA1c superior or **egal** to 7.5%*	NCT01144728
Variable		*Glycosylated haemoglobin (**HbA 1c**) < 10%.*	NCT00274118
Numeric		*HbA1c between **45** and **94***	NCT01513798
Coding		*6.5% ⍰ HbA1c ⍰ 9% at screening visit*	NCT00541437





SemAF for Measu

Use the mouse to get annotation items in t

Choose the type:

entity: medicalConcept ▾ comparison: greaterThan

PS: Only one sentence can be processed at a tim

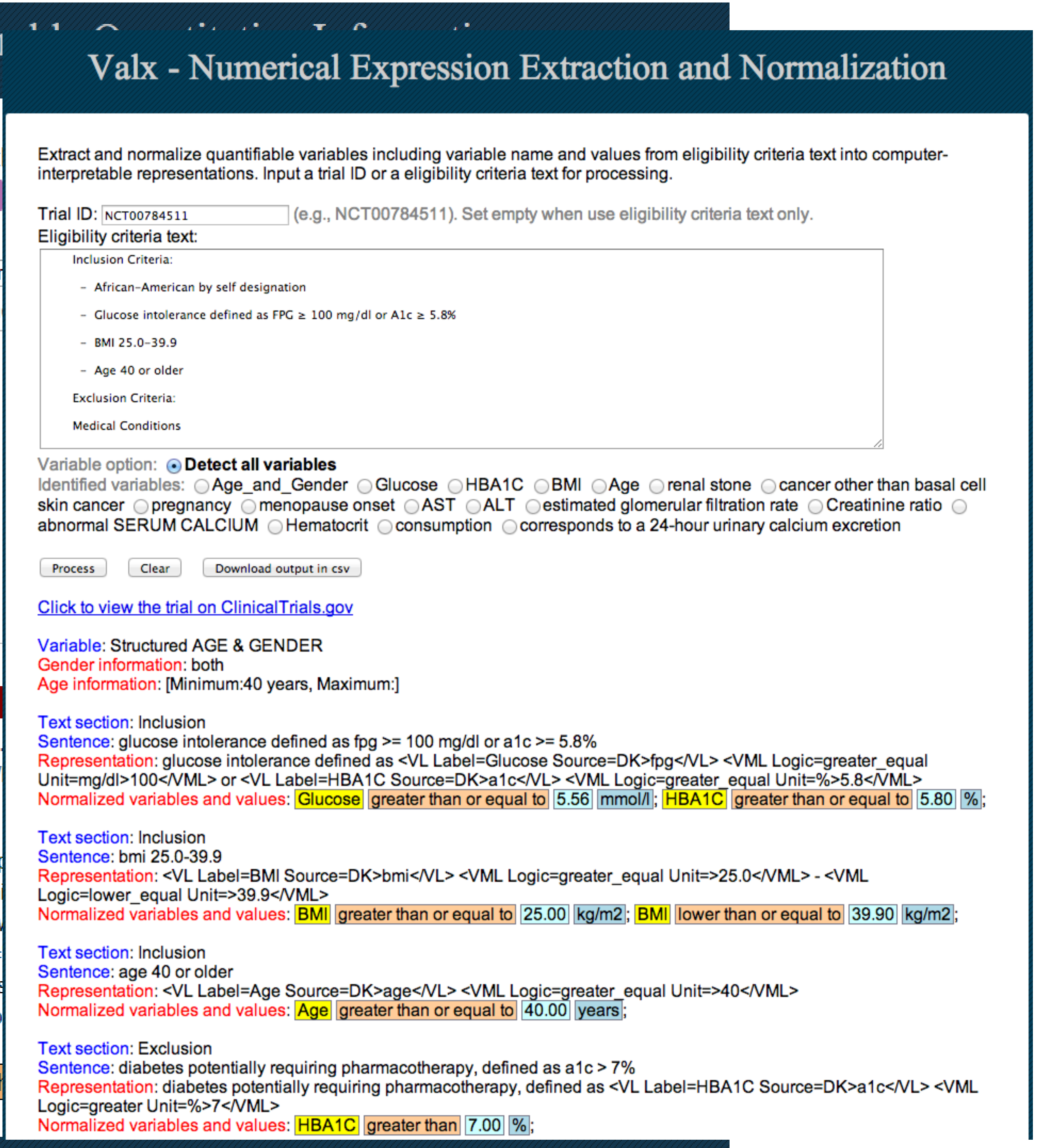
White blood cell count > 14.0 X 109 / L.

Annotation in XML:

Sentence: White blood cell count > 14.0 X 109 / L.

```
a.<wordSeg xml:id="ws1" target="#1a" lang="en">W
L_w10</wordSeg>
b.<MQI xml:id="qi1" target="#ws1">
        <entity xml:id="x1" target="#w1, #w2, #w3" typ
        <measure xml:id="me1" target=" " num=" " uni
        <measure xml:id="me2" target="#w4, #w6, #w
        <comparison xml:id="cp1" target="#w5" type=
        <cLink xml:id="coL1" measure1="#me1" meas
        <mLink xml:id="meL1" measureID="#me1" ap
   </MQI>
```

Variables and values: White blood cell greaterTh



Valx - Numerical Expression Extraction and Normalization

Extract and normalize quantifiable variables including variable name and values from eligibility criteria text into computer-interpretable representations. Input a trial ID or a eligibility criteria text for processing.

Trial ID: NCT00784511 (e.g., NCT00784511). Set empty when use eligibility criteria text only.

Eligibility criteria text:

Inclusion Criteria:

- African-American by self designation
- Glucose intolerance defined as FPG ≥ 100 mg/dl or A1c ≥ 5.8%
- BMI 25.0-39.9
- Age 40 or older

Exclusion Criteria:

Medical Conditions

Variable option: ● **Detect all variables**

Identified variables: ○ Age_and_Gender ○ Glucose ○ HBA1C ○ BMI ○ Age ○ renal stone ○ cancer other than basal cell skin cancer ○ pregnancy ○ menopause onset ○ AST ○ ALT ○ estimated glomerular filtration rate ○ Creatinine ratio ○ abnormal SERUM CALCIUM ○ Hematocrit ○ consumption ○ corresponds to a 24-hour urinary calcium excretion

Click to view the trial on ClinicalTrials.gov

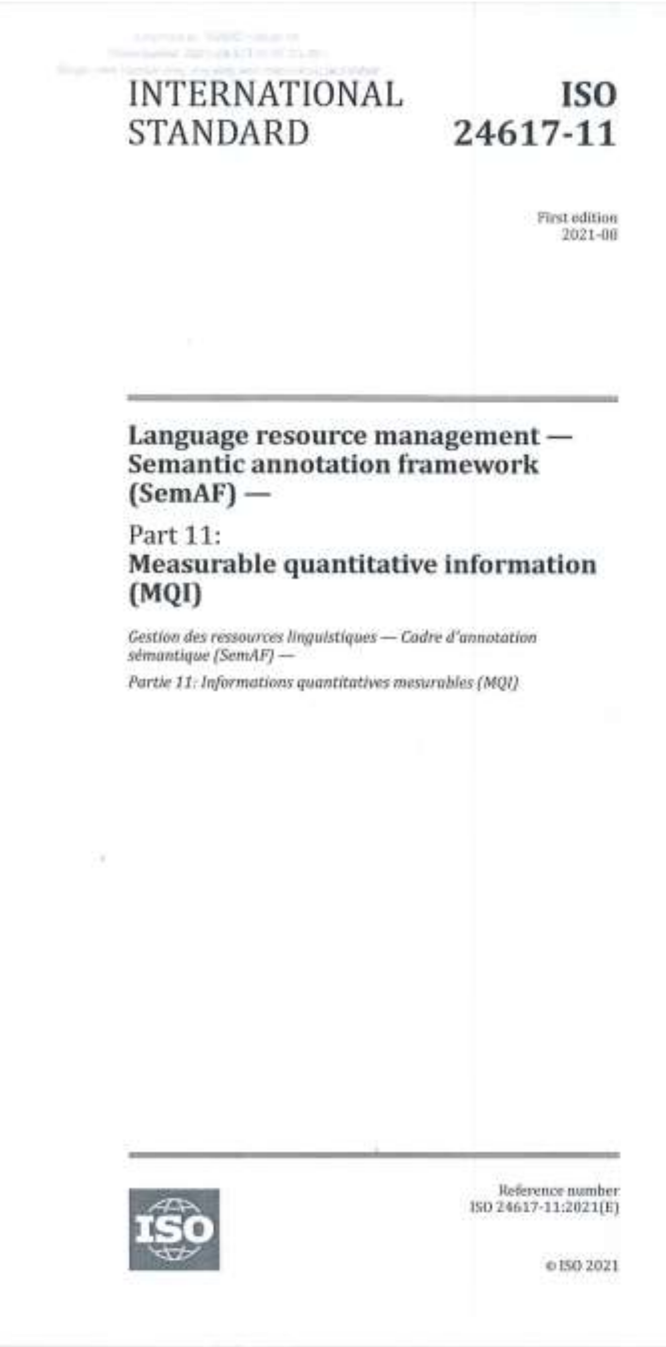
Variable: Structured AGE & GENDER
Gender information: both
Age information: [Minimum:40 years, Maximum:]

Text section: Inclusion
Sentence: glucose intolerance defined as fpg >= 100 mg/dl or a1c >= 5.8%
Representation: glucose intolerance defined as <VL Label=Glucose Source=DK>fpg</VL> <VML Logic=greater_equal Unit=mg/dl>100</VML> or <VL Label=HBA1C Source=DK>a1c</VL> <VML Logic=greater_equal Unit=%>5.8</VML>
Normalized variables and values: Glucose greater than or equal to 5.56 mmol/l; HBA1C greater than or equal to 5.80 %;

Text section: Inclusion
Sentence: bmi 25.0-39.9
Representation: <VL Label=BMI Source=DK>bmi</VL> <VML Logic=greater_equal Unit=>25.0</VML> - <VML Logic=lower_equal Unit=>39.9</VML>
Normalized variables and values: BMI greater than or equal to 25.00 kg/m2; BMI lower than or equal to 39.90 kg/m2;

Text section: Inclusion
Sentence: age 40 or older
Representation: <VL Label=Age Source=DK>age</VL> <VML Logic=greater_equal Unit=>40</VML>
Normalized variables and values: Age greater than or equal to 40.00 years;

Text section: Exclusion
Sentence: diabetes potentially requiring pharmacotherapy, defined as a1c > 7%
Representation: diabetes potentially requiring pharmacotherapy, defined as <VL Label=HBA1C Source=DK>a1c</VL> <VML Logic=greater Unit=%>7</VML>
Normalized variables and values: HBA1C greater than 7.00 %;



INTERNATIONAL STANDARD

ISO 24617-11

First edition
2021-08

Language resource management — Semantic annotation framework (SemAF) —

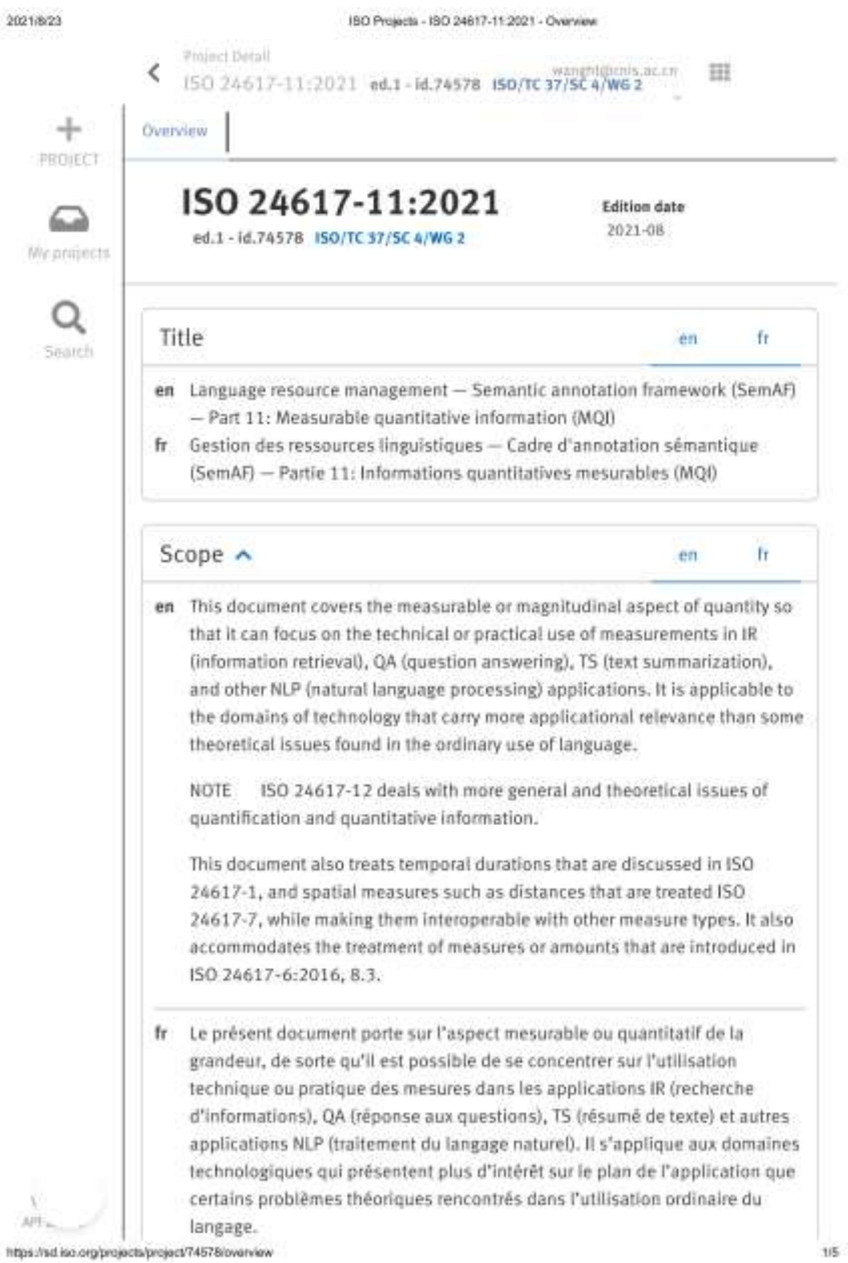
Part 11:
Measurable quantitative information (MQI)

Gestion des ressources linguistiques — Cadre d'annotation sémantique (SemAF) —

Partie 11: Informations quantitatives mesurables (MQI)

Reference number
ISO 24617-11:2021(E)

© ISO 2021



2021/8/23 ISO Projects - ISO 24617-11:2021 - Overview

ISO 24617-11:2021

ed.1 - id.74578 ISO/TC 37/SC 4/WG 2

Edition date
2021-08

Title (en | fr)

en Language resource management — Semantic annotation framework (SemAF) — Part 11: Measurable quantitative information (MQI)

fr Gestion des ressources linguistiques — Cadre d'annotation sémantique (SemAF) — Partie 11: Informations quantitatives mesurables (MQI)

Scope (en | fr)

en This document covers the measurable or magnitudinal aspect of quantity so that it can focus on the technical or practical use of measurements in IR (information retrieval), QA (question answering), TS (text summarization), and other NLP (natural language processing) applications. It is applicable to the domains of technology that carry more applicational relevance than some theoretical issues found in the ordinary use of language.

NOTE ISO 24617-12 deals with more general and theoretical issues of quantification and quantitative information.

This document also treats temporal durations that are discussed in ISO 24617-1, and spatial measures such as distances that are treated ISO 24617-7, while making them interoperable with other measure types. It also accommodates the treatment of measures or amounts that are introduced in ISO 24617-6:2016, 8.3.

fr Le présent document porte sur l'aspect mesurable ou quantitatif de la grandeur, de sorte qu'il est possible de se concentrer sur l'utilisation technique ou pratique des mesures dans les applications IR (recherche d'informations), QA (réponse aux questions), TS (résumé de texte) et autres applications NLP (traitement du langage naturel). Il s'applique aux domaines technologiques qui présentent plus d'intérêt sur le plan de l'application que certains problèmes théoriques rencontrés dans l'utilisation ordinaire du langage.

https://isd.iso.org/projects/project/74578/overview 1/5

Thank you

haoty@126.com

華南師範大學
SOUTH CHINA NORMAL UNIVERSITY

文本分析与挖掘
Text Analytics and Mining Lab

iHAVC 智慧健康与可视化计算
Intelligent Health and Visual Computing