Strategies and Recent Progress Update for Curating Chinese Standard Vocabularies Using

the OHDSI Common Data Model

Cui Tao

University of Texas Health Science Center at Houston

陶萃

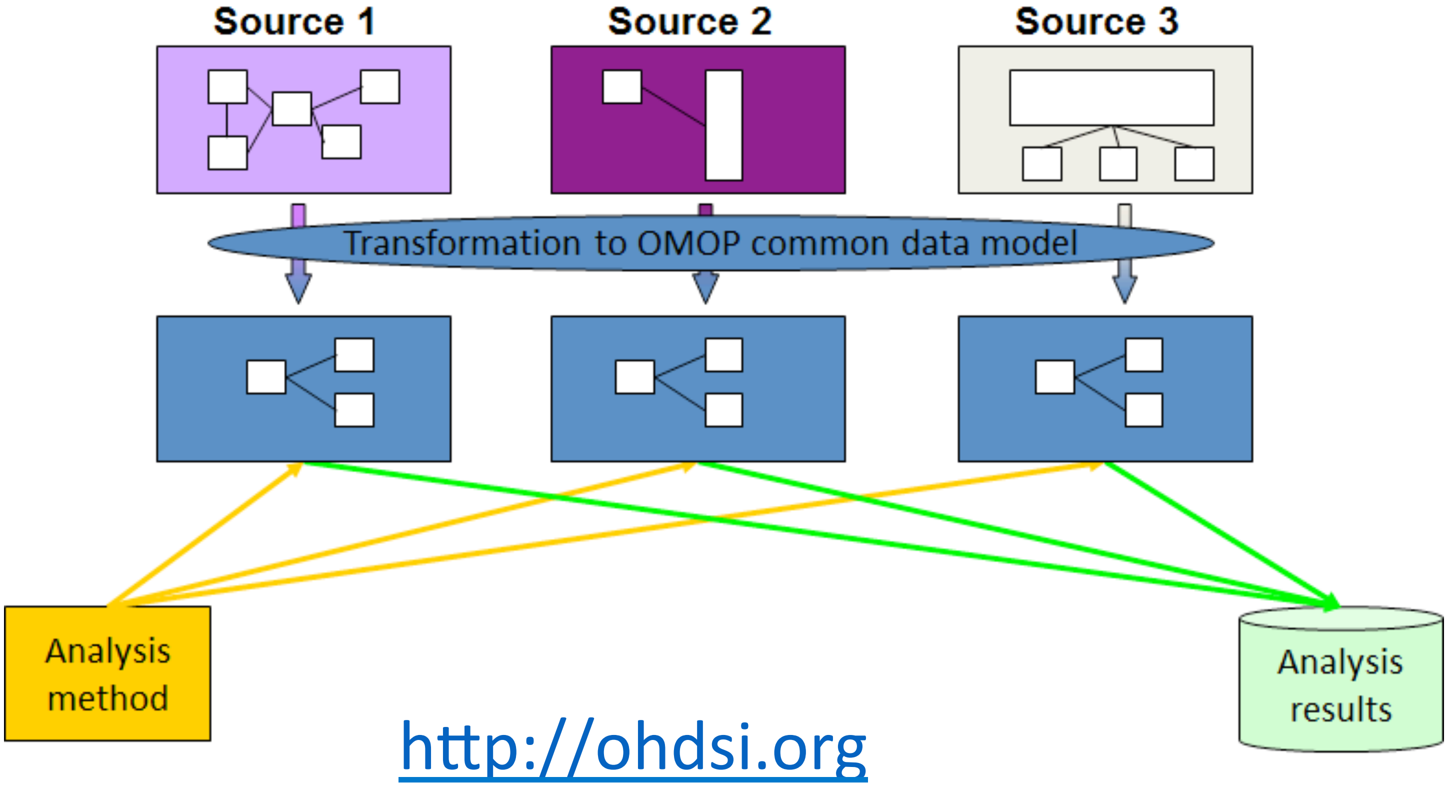
德州大学休斯顿生物医学信息学院

On behalf of the OHDSI China CDM and Vocabulary WG Core Team

Observational Health Data Sciences and Informatics (OHDSI)

- OHDSI has developed tools for transforming, characterizing, and analyzing disparate data sources across the health care delivery spectrum;
- Requires use of a Common Data Model;
- Standard ontologies are set by the consortium;
- A suite of tools allow for data to be mapped and transformed to fit the model.

OHDSI Common Data Model



http://ohdsi.org

OHDSI China Strategy

- In Phase I, we build a set of core standard vocabularies in the domains of conditions (疾病）, procedures （手术）, laboratory tests （临床检验）, and clinical drugs （药品）.
- In Phase 2, create high-quality mappings between local code systems to international standard vocabularies.

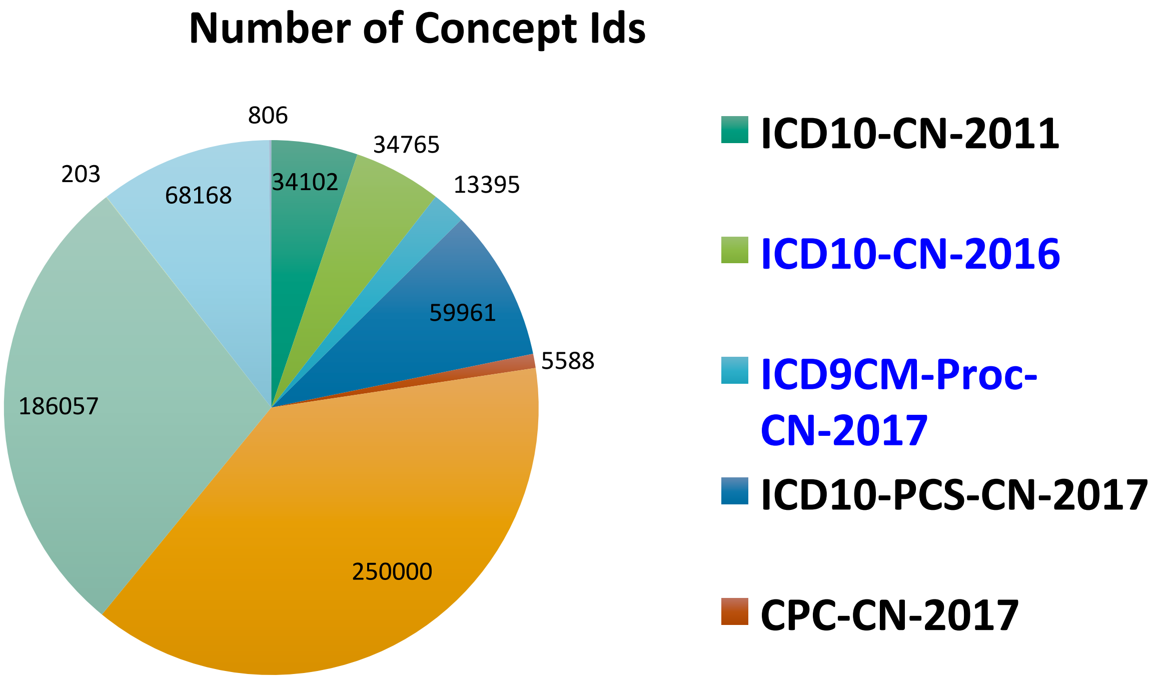
Recent Activities

- HIMSS OHDSI Meetup (March 5, 2018)
- Biocuration 2018 Preconference (April 8, 2018)
- WG Testing through June, 2018
- Collaboration with other WGs using sample or real-world data
- OHDSI China 2018 Annual Symposium (June 30, 2018) – V1 Progress Report
- OHDSI China Webinar July – V1 Progress Report
- OHDSI Collaborator meeting Aug 21,2018

Phase I Core Vocabularies

- Diseases/Conditions 疾病分类标准编码
  - ICD10-CN-2016 (Core)
  - ICD10-CN-2011
- Procedures 手术操作标准编码
  - ICD9CM-Procedures-CN (Core)
  - ICD10PCS-CN
  - Common Procedure Codes
- Drugs 药品标准编码
  - Normalized Chinese Clinical Drugs (NCCD) (Core)
  - NDC-CN
- Laboratory Tests 临床检验标准编码
  - LOINC-CN (Core)
  - Common Laboratory Test Codes-CN

Statistics As of April 8, 2018 (n=653,045)



OHDSI China Phase I Vocabularies for Trial Use （by March 26, 2018）

- Main goal:
  - To assess whether core vocabularies can meet data annotation needs for real-world data.
- Requirements:
  - 1） Testers need to have real world datasets at least in one of four domains.
  - 2） Testers agree to provide all codes, code names, record frequency and mappings to core vocabularies.
  - 3） OHDSI China will provide mapping tools to testers;
  - 4） OHDSI China will analyze and review the mappings, make the improvement plan, and release updated core vocabularies and mappings.
  - 5） OHDSI China will produce journal publications based on the analysis and review.

Template for source data

- Source Datasets*
  - Source data start/end dates$
  - Source data hospital type (general, specialized, others)
  - Total count of distinct patients
  - Total count of outpatient
  - Total count of inpatient

  - If diseases, procedures, drugs, and lab tests are collected from different source datasets, each of the source datasets should be described。
  - $Data between January 1, 2013 and December 31, 2017 is preferred.

Template for Disease Coding Data Collection

- Fields
  - Source data ICD code
  - Source data ICD version
  - Source data ICD code name
  - Source data ICD code frequency
  - Source data clinical diagnosis name （Optional）

Template for Procedure Coding Data Collection

- Fields
  - Source data procedure ICD9 code
  - Source data procedure ICD9 version
  - Source data procedure ICD9 code name
  - Source data clinical procedure name （Optional）
  - Source data procedure code frequency

Template for Drug Coding Data Collection

- Fields
    - Source Data Drug Name
    - Source Data Drug Strength
    - Source Data Drug Dose Form
    - Source Data Drug Brand Name
    - Source Data Drug Manufacturer
    - Source Data Drug Record Frequency
    - Source Data Drug Code 1
    - Source Data Drug Code Type 1
    - Source Data Drug Code 2 (Optional)
    - Source Data Drug Code Type 2 (Optional)

Template for Lab Test Coding Data Collection

- Fields
  - Source Data Test Name
  - Source Data Test Specimen
  - Source Data Test Units
  - Source Data Test Record Frequency
  - Source Data Test Normal Range (Optional)
  - Source Data Test Abnormal Flag (Optional)
  - Source Data Test Code (Optional)
  - Source Data Test Code Type (Optional)

Source Data Collection

	Disease	Drug	Procedure	Lab
Source 1	** 1,998		999	
Source 2	4		** 77,661	
Source 3	83,295	822	4,725	1,127
Source 4	231,537	** 143,515	11,395	8,859
Source 5	15,707	2,555	7,624	4,070
Source 6	1,910	4,304	49	648
Source 7	1,166			
Source 8	66,754			
Source 9	21,531	2,211		8,933
Source 10	399	200	200	

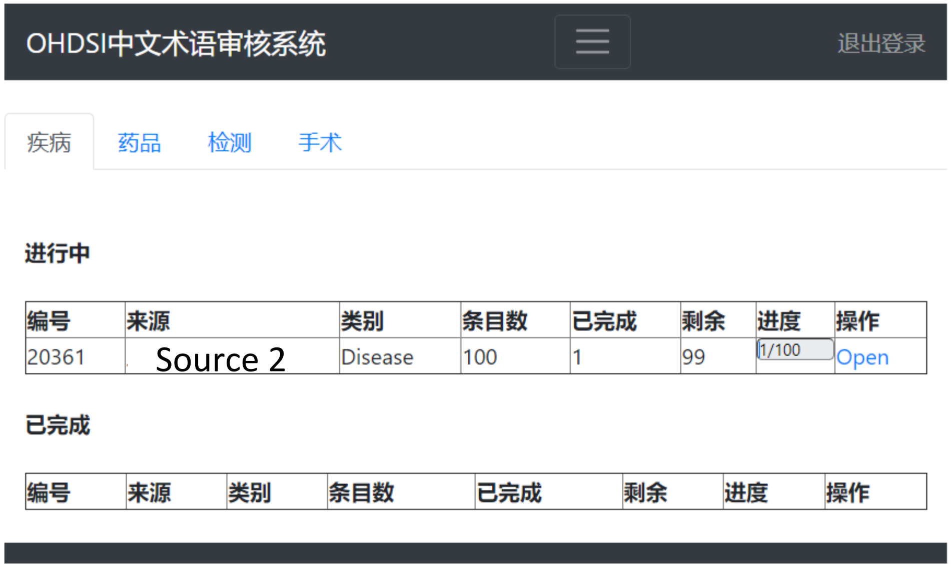
** Contains duplicates

OHDSI Chinese Mapping Evaluation Tool

- Register/Login
- Obtain/Assign review tasks
- Each task includes 100 terms

OHDSI China

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



OHDSI中文术语审核系统

进行中

编号	来源	类别	条目数	已完成	剩余	进度	操作
20361	Source 2	Disease	100	1	99	1/100	Open

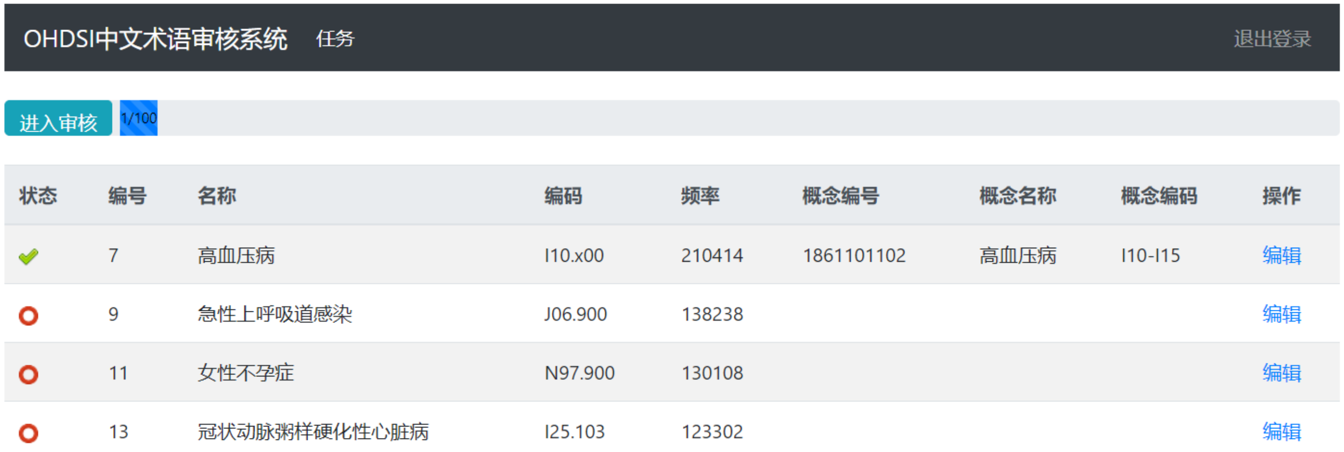
已完成

编号	来源	类别	条目数	已完成	剩余	进度	操作

OHDSI Chinese Mapping Evaluation Tool

One single task

- Show term list
- Show progress
- Enter evaluate
- Edit each term evaluation



OHDSI中文术语审核系统 任务 退出登录

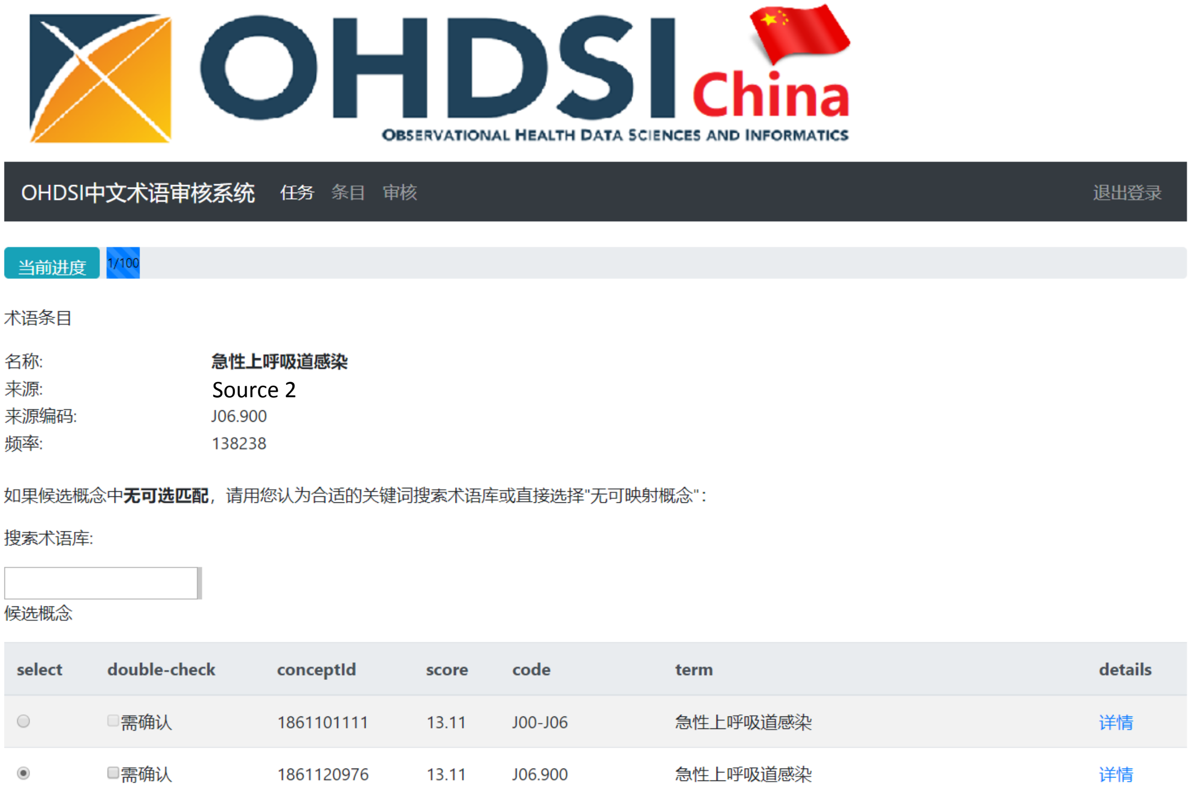
进入审核 1/100

状态	编号	名称	编码	频率	概念编号	概念名称	概念编码	操作
✓	7	高血压病	I10.x00	210414	1861101102	高血压病	I10-I15	编辑
○	9	急性上呼吸道感染	J06.900	138238				编辑
○	11	女性不孕症	N97.900	130108				编辑
○	13	冠状动脉粥样硬化性心脏病	I25.103	123302				编辑

OHDSI Chinese Mapping Evaluation Tool

Term Mapping

- For each source term, automatically search for a match；
- User can define their own keyword for search；
- For those uncertain matches, the annotator can marked them as “pending”



术语条目

名称:	**急性上呼吸道感染**
来源:	Source 2
来源编码:	J06.900
频率:	138238

如果候选概念中**无可选匹配**，请用您认为合适的关键词搜索术语库或直接选择"无可映射概念"：

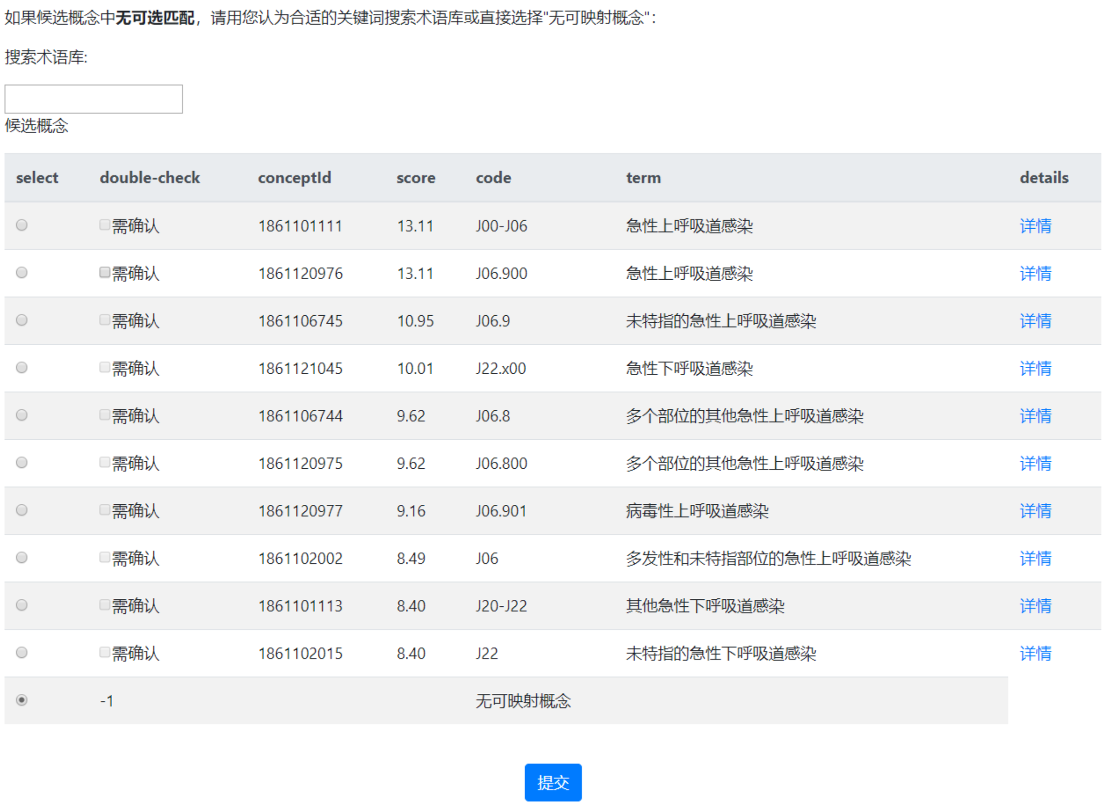
搜索术语库:

候选概念

select	double-check	conceptId	score	code	term	details
○	□需确认	1861101111	13.11	J00-J06	急性上呼吸道感染	详情
●	□需确认	1861120976	13.11	J06.900	急性上呼吸道感染	详情

OHDSI Chinese Mapping Evaluation Tool

If no appropriate match can be found, choose “NO MATCH”



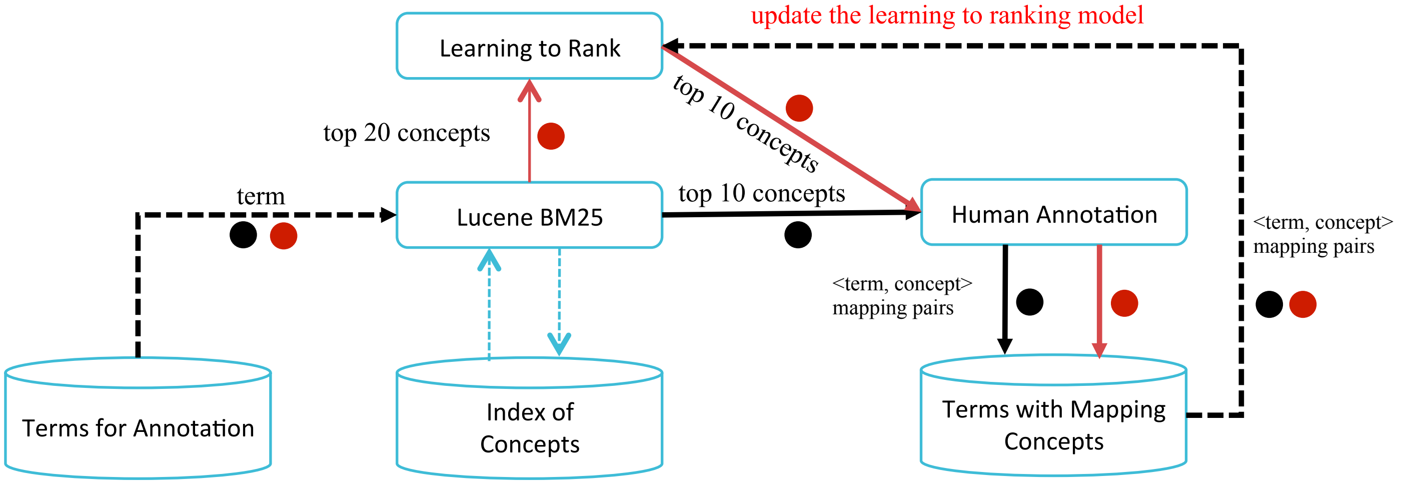
如果候选概念中**无可选匹配**，请用您认为合适的关键词搜索术语库或直接选择"无可映射概念"：

搜索术语库:

候选概念

select	double-check	conceptId	score	code	term	details
○	□需确认	1861101111	13.11	J00-J06	急性上呼吸道感染	详情
○	□需确认	1861120976	13.11	J06.900	急性上呼吸道感染	详情
○	□需确认	1861106745	10.95	J06.9	未特指的急性上呼吸道感染	详情
○	□需确认	1861121045	10.01	J22.x00	急性下呼吸道感染	详情
○	□需确认	1861106744	9.62	J06.8	多个部位的其他急性上呼吸道感染	详情
○	□需确认	1861120975	9.62	J06.800	多个部位的其他急性上呼吸道感染	详情
○	□需确认	1861120977	9.16	J06.901	病毒性上呼吸道感染	详情
○	□需确认	1861102002	8.49	J06	多发性和未特指部位的急性上呼吸道感染	详情
○	□需确认	1861101113	8.40	J20-J22	其他急性下呼吸道感染	详情
○	□需确认	1861102015	8.40	J22	未特指的急性下呼吸道感染	详情
●	-1			无可映射概念		

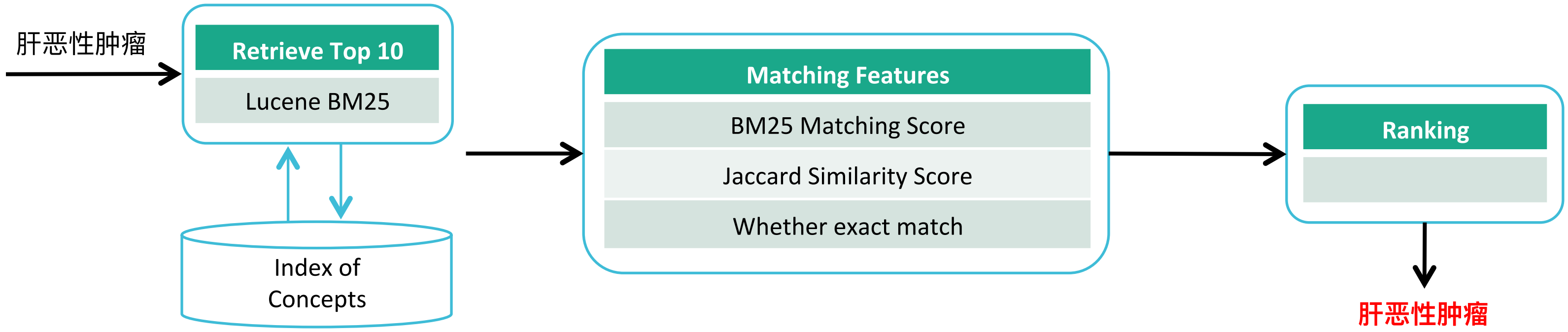
Iterative Annotation



Initial round: ● ● ● ●

Iterative round: ● ● ● ● ●

Learning to Rank (L2R)



肝恶性肿瘤

Retrieve Top 10

Lucene BM25

Index of Concepts

Matching Features

BM25 Matching Score

Jaccard Similarity Score

Whether exact match

Ranking

肝恶性肿瘤

Top 10 Concepts
肝管恶性肿瘤
肝恶性肿瘤
…
恶性肿瘤

Top 10 Concepts	BM25	Jaccard	Exact
肝管恶性肿瘤	10.52	0.83	0
肝恶性肿瘤	10.19	1.0	1
…			
恶性肿瘤	9.41	0.8	0

Top 10 Concepts	score
肝管恶性肿瘤	0.87
肝恶性肿瘤	**0.98**
…	
恶性肿瘤	0.83

First Round Testing

Drug	3759
Disease	2891
Procedure	2248
Lab Test	1277

	Source	First Round
Drug	Source 1	2211
Drug	Source 2	1548
Lab Test	Source 3	1277
Procedure	Source 2	1050
Procedure	Source 4	936
Procedure	Source 3	262
Disease	Source 4 In patient	882
Disease	Source 4 Out patient	786
Disease	Source 2	1223

Progress

		First Round	Evaluated	Progress
Drug	Source 1	2211	642	29.0%
Drug	Source 2	1548	1400	90.4%
Lab Test	Source 3	1277	38	3.0%
Procedure	Source 2	1050	1000	95.2%
Procedure	Source 4	936	839	89.6%
Procedure	Source 3	262	200	76.3%
Disease	Source 4 In patient	882	435	49.3%
Disease	Source 4 Out patient	786	700	89.1%
Disease	Source 2	1223	1101	90.0%

Mapping Result

		Mapped		Pending		No Available Mapping	
Category	**Total**	**Number**	**%**	**Number**	**%**	**Number**	**%**
Disease	2236	2003	89.6%	81	3.6%	151	6.8%
Lab	38	8	21.1%	7	18.4%	23	60.5%
Drug	2042	1005	49.2%	891	43.6%	146	7.1%
Procedure	2039	1362	66.8%	165	8.1%	512	25.1%

Discussion

Disease

Mapping Types	Criteria	Examples
Exact matching	The source term could be exactly mapped to a standard ICD-10-CN concept with the same term name. The source code was referred, but not determined factor.	Source term: 牙科检查 Source code: Z01.200 → Standard ICD concept: 牙科检查 Standard ICD code: Z01.200
		Source term: 上消化道出血 Source code: K92.204 (ICD) → Standard ICD concept: 上消化道出血 Standard ICD code: K92.208
Partial matching	The source term could only be partially mapped to a standard ICD-10-CN concept, the concept semantics may be similar,narrowed or broadened. The source code was referred, but not determined factor.	Source term: 特指皮炎 Source code: L30.800 (ICD) → Standard ICD concept: 皮炎，其他特指的 Standard ICD code: L30.800

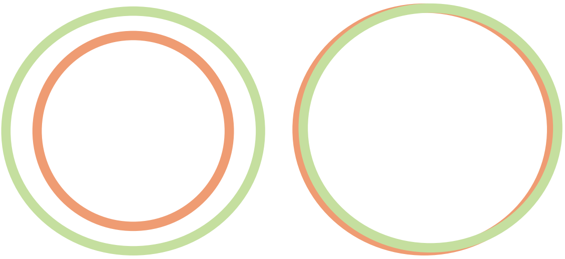
Unmapped Situations	Examples
For some source terms, there are no standard concepts could be mapped. 1.Source code may be same	Source term: 婴儿支气管炎 Source code: J20.902 →Unmapped [Standard ICD concept: 急性气管支气管炎 Standard ICD Code: J20.902]
2.There are similar concepts returned by searching tool, however, with different semantics, it could not be mapped after manual review.	Source term: 乳腺囊肿 Source code: N60.001 → Unmapped [Standard ICD concept: 乳腺脓肿 Standard ICD Code: N61.X03]
3. Very deep level or specific subcategory of diagnosis code， no concepts could be mapped by searching tool	Source term: 早期妊娠状态，13周以下 Source code: Z34.90001 → Unmapped

Procedures

Matching Guidelines of Procedures

Match

- Original term ⊆ Standard term

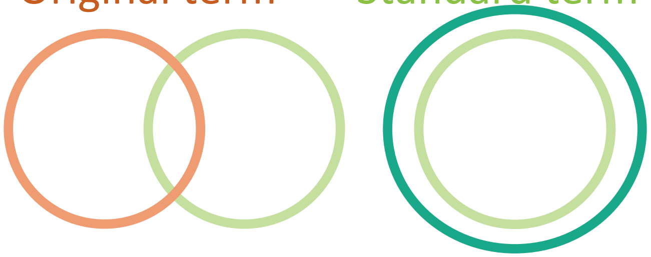


Original term

Standard term

Cannot match

- Original term ⊄ Standard term
- Original term ⊃ Standard term



Original term

Standard term

Detailed Analysis of Mapping Cases

Match

- The same meaning:
  - Original term and standard term have the same words
  - Original term and standard term have different words
- Within (＜) the scope of standard term

Cannot Match

- Partially intersecting
- Broader (＞) than standard terms
- Combination term: including two or more procedures.

Match

- Case 1 same meaning with the same words
  E.g. Original term: 输尿管扩张术 （Code: 59.8 01）
  Mapping to standard term: 输尿管扩张术 （Code: 59.8×01）

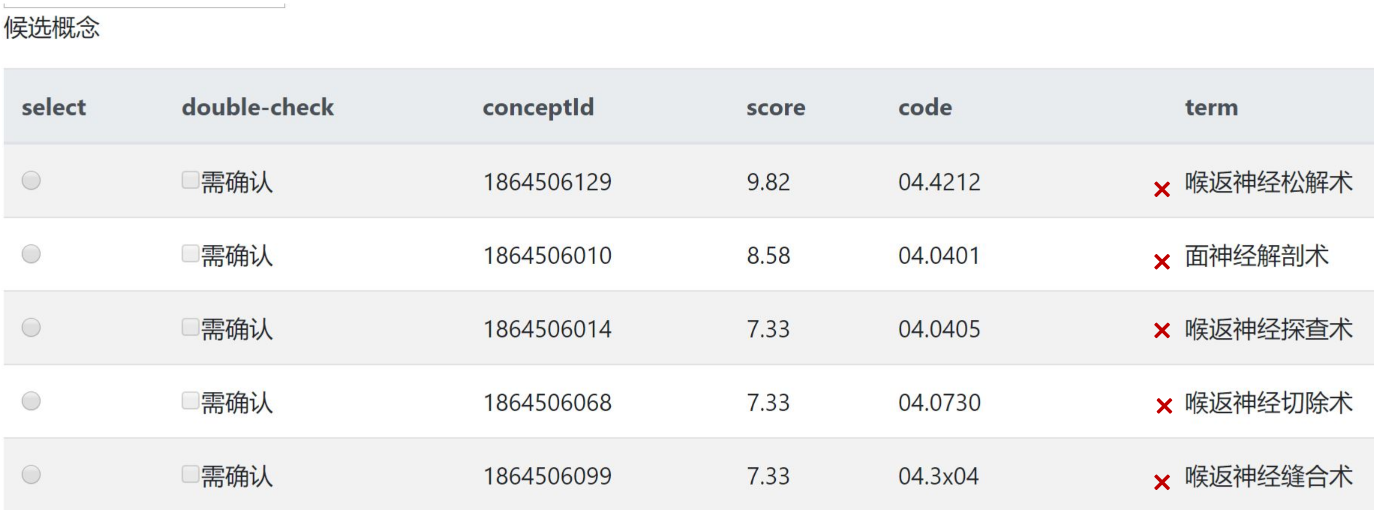
- Case 2 same meaning with different words
  E.g. Original term: 经皮睾丸活检术 （Code: 62.1101）
  Mapping standard term:闭合性(经皮)(针吸)睾丸活组织检查（Code: 62.1100）

- Case 3 within the scope of standard term
  E.g. Original term: 直肠**肿瘤**切除术 （Code: 48.351）
  Mapping standard term:直肠**病损**切除术 （Code: 48.3501） **肿瘤**∈**病损**

Cannot Match

- Case 1 partially intersecting, no exact match

  E.g. Original term: 喉返神经解剖术



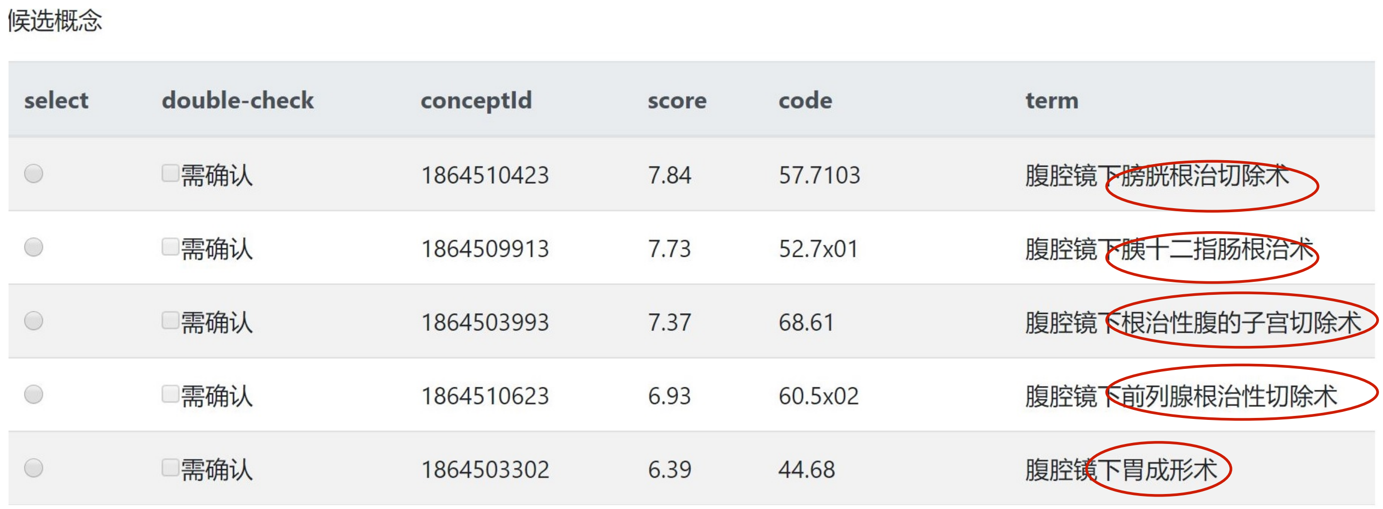
候选概念

select	double-check	conceptId	score	code	term
○	□需确认	1864506129	9.82	04.4212	✗ 喉返神经松解术
○	□需确认	1864506010	8.58	04.0401	✗ 面神经解剖术
○	□需确认	1864506014	7.33	04.0405	✗ 喉返神经探查术
○	□需确认	1864506068	7.33	04.0730	✗ 喉返神经切除术
○	□需确认	1864506099	7.33	04.3x04	✗ 喉返神经缝合术

Cannot Match

- Case 2 broader than the standard terms

E.g. Original term: 腹腔镜下治疗 (What disease was treated?)



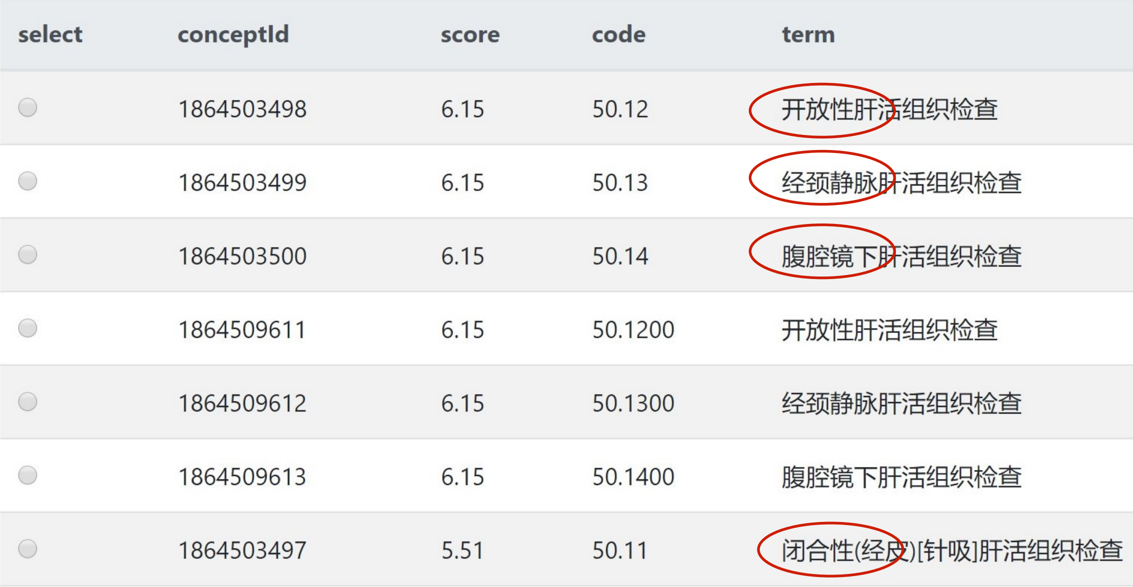
候选概念

select	double-check	conceptId	score	code	term
○	□需确认	1864510423	7.84	57.7103	腹腔镜下膀胱根治切除术
○	□需确认	1864509913	7.73	52.7x01	腹腔镜下胰十二指肠根治术
○	□需确认	1864503993	7.37	68.61	腹腔镜下根治性腹的子宫切除术
○	□需确认	1864510623	6.93	60.5x02	腹腔镜下前列腺根治性切除术
○	□需确认	1864503302	6.39	44.68	腹腔镜下胃成形术

Cannot Match

- Case 2 broader than the standard terms

  E.g. Original term: 肝活组织检查 (How to perform liver biopsy?)

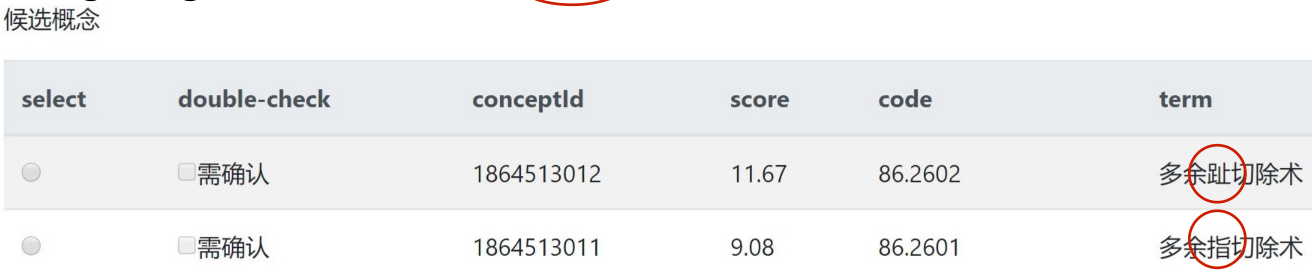


select	conceptId	score	code	term
○	1864503498	6.15	50.12	开放性肝活组织检查
○	1864503499	6.15	50.13	经颈静脉肝活组织检查
○	1864503500	6.15	50.14	腹腔镜下肝活组织检查
○	1864509611	6.15	50.1200	开放性肝活组织检查
○	1864509612	6.15	50.1300	经颈静脉肝活组织检查
○	1864509613	6.15	50.1400	腹腔镜下肝活组织检查
○	1864503497	5.51	50.11	闭合性(经皮)[针吸]肝活组织检查

Cannot Match

- Case 3 combination term, including two or more procedures

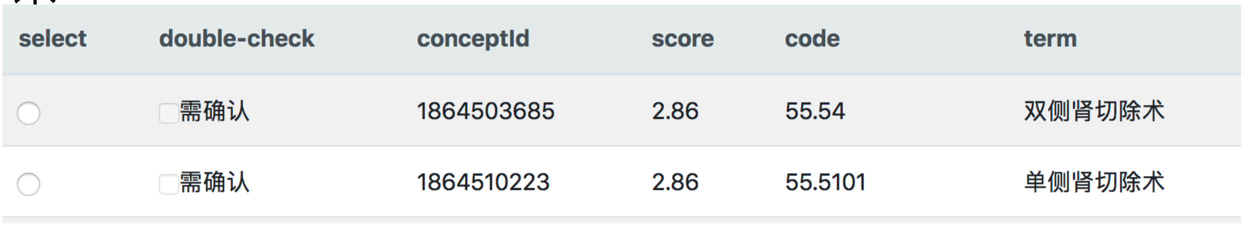
E.g. Original term: 多余指（趾）切除术



候选概念

select	double-check	conceptId	score	code	term
○	□需确认	1864513012	11.67	86.2602	多余趾切除术
○	□需确认	1864513011	9.08	86.2601	多余指切除术

E.g. Original term: 右侧足背囊肿切除+右侧足背脂肪瘤切除+右侧足骰骨病灶清除术

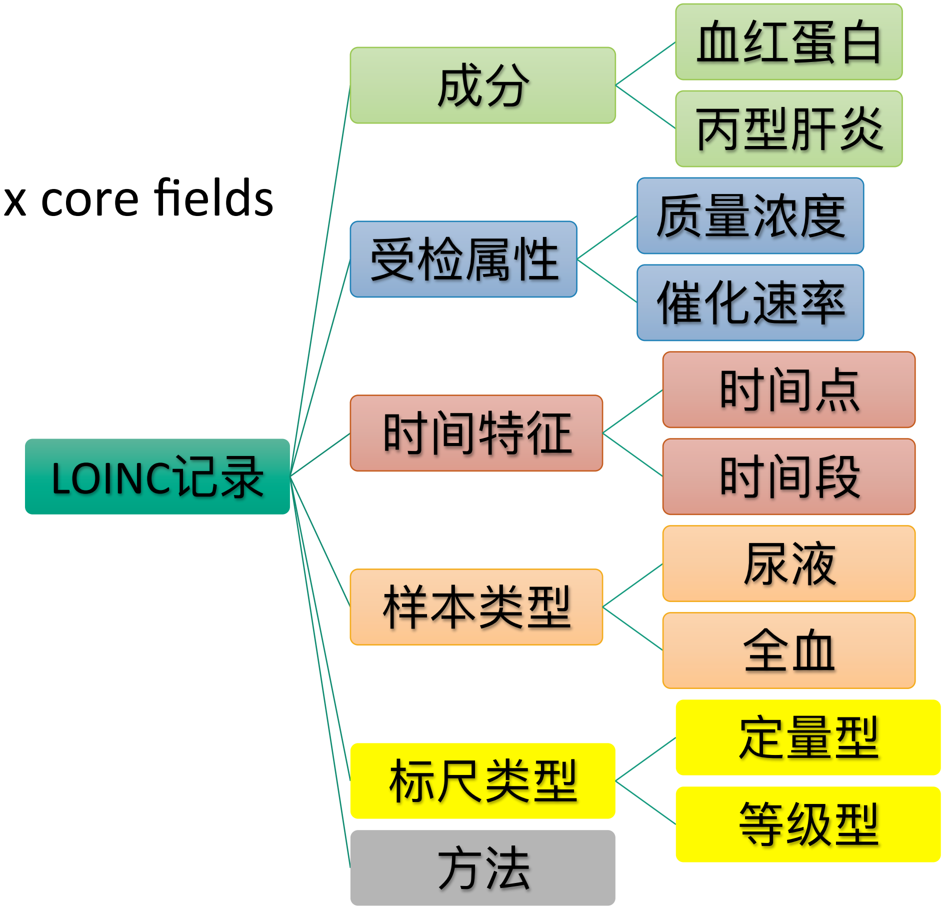


select	double-check	conceptId	score	code	term
○	□需确认	1864503685	2.86	55.54	双侧肾切除术
○	□需确认	1864510223	2.86	55.5101	单侧肾切除术

Lab Test

LONIC code

Each LONIC record has six core fields



- LOINC记录
  - 成分
    - 血红蛋白
    - 丙型肝炎
  - 受检属性
    - 质量浓度
    - 催化速率
  - 时间特征
    - 时间点
    - 时间段
  - 样本类型
    - 尿液
    - 全血
  - 标尺类型
    - 定量型
    - 等级型
  - 方法

LONIC code mapping
Each field in a LOINC record needs to be matched

名称:	3小时镜检白细胞
来源:	Source 3
频率:	447
单位:	/HP
样本:	尿液
更多说明:	0～5

→

LOINC-code	59829_2
成分:	白细胞
受检属性:	计数型速率
时间特征:	3小时
样本类型:	尿沉渣
标尺类型:	定量型
方法:	显微镜检查.光学

Errors in lab test data

Wrong unit

名称:	大便隐血	Fecal occult blood
频率:	2	
单位:	**阴性**	negative
样本:	粪便	Stool
更多说明:		

Errors in lab test data

Mis-match of the component and the sample

名称:	**尿透明度**	名称:	凝血酶原时间
频率:	13	频率:	1
单位:		单位:	sec
样本:	**粪便**	样本:	**尿液**
更多说明:		更多说明:	10.0～15.0

Missing key information of lab test

名称:	**单核细胞数**
单位:	
样本:	血液
更多说明:	

名称:	白蛋白
单位:	
样本:	
更多说明:	

名称:	垂体泌乳素
单位:	mIU/L
样本:	
更多说明:	

No matching code in LOINC

名称:	**全血低切相对指数**
频率:	3040
单位:	
样本:	血液
更多说明:	8.11～14.21

名称:	**肥达氏反应-A**
频率:	12
单位:	
样本:	血液
更多说明:	阴性

名称:	**巨大不成熟细胞比值**
频率:	1
单位:	
样本:	引流液

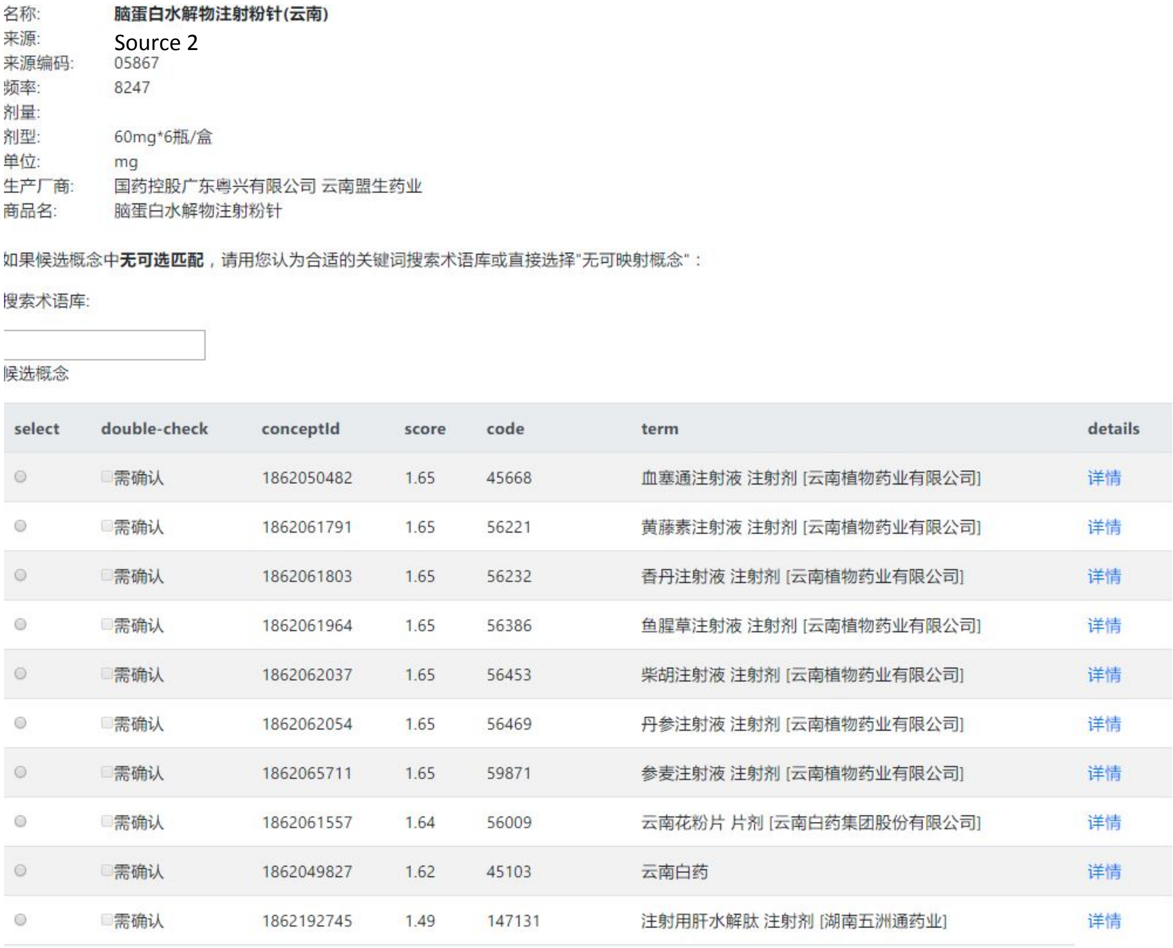
Drug

二、隐射工具本身没有匹配到正确的概念，自己搜索：

脑蛋白水解物注射粉针

（下一页是搜索结果）

There is no match can be found
By the tool. The annotator used key-word based search



名称:	**脑蛋白水解物注射粉针(云南)**
来源:	Source 2
来源编码:	05867
频率:	8247
剂量:	
剂型:	60mg*6瓶/盒
单位:	mg
生产厂商:	国药控股广东粤兴有限公司 云南盟生药业
商品名:	脑蛋白水解物注射粉针

如果候选概念中**无可选匹配**，请用您认为合适的关键词搜索术语库或直接选择"无可映射概念"：

搜索术语库:

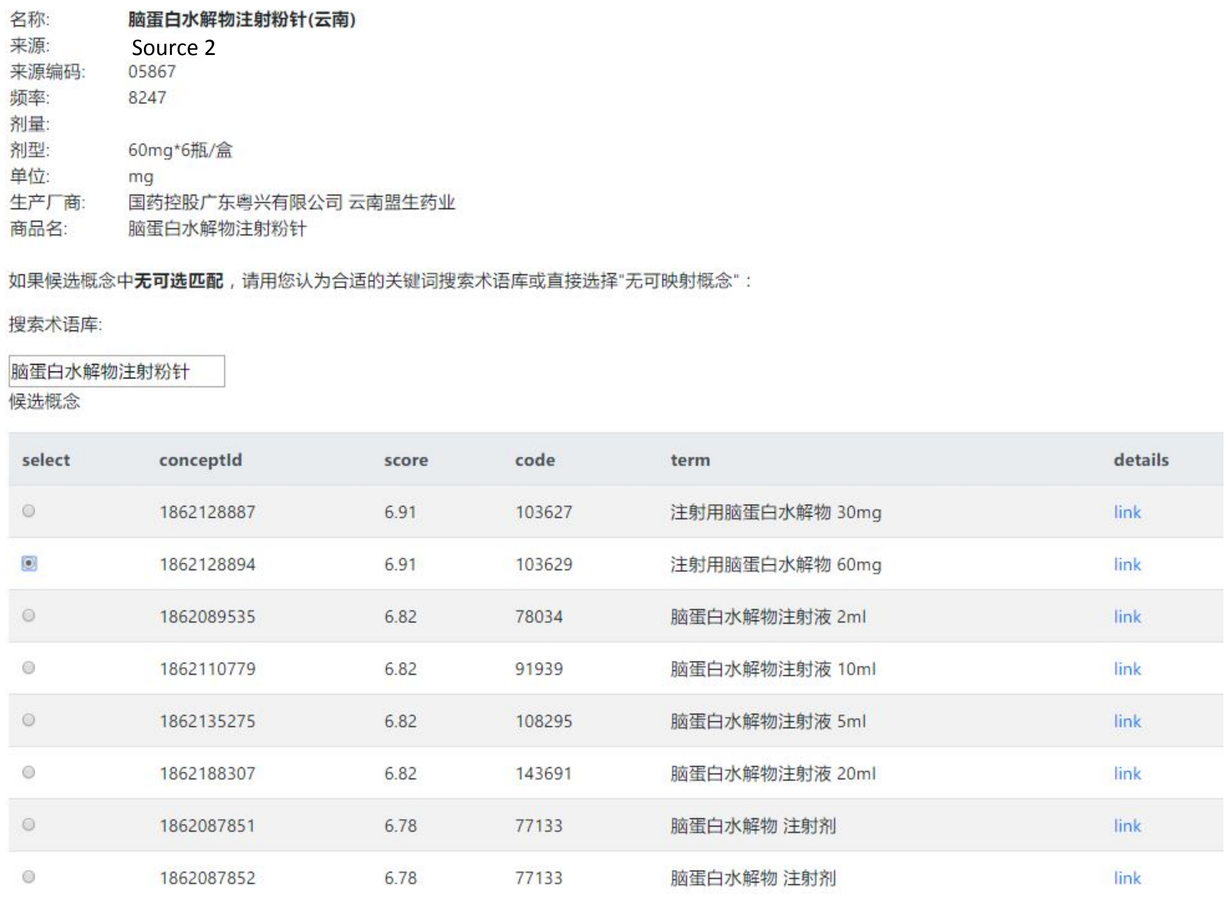
候选概念

select	double-check	conceptId	score	code	term	details
○	□需确认	1862050482	1.65	45668	血塞通注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862061791	1.65	56221	黄藤素注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862061803	1.65	56232	香丹注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862061964	1.65	56386	鱼腥草注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862062037	1.65	56453	柴胡注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862062054	1.65	56469	丹参注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862065711	1.65	59871	参麦注射液 注射剂 [云南植物药业有限公司]	详情
○	□需确认	1862061557	1.64	56009	云南花粉片 片剂 [云南白药集团股份有限公司]	详情
○	□需确认	1862049827	1.62	45103	云南白药	详情
○	□需确认	1862192745	1.49	147131	注射用肝水解肽 注射剂 [湖南五洲通药业]	详情

搜索结果：

结果中没有一个完全匹配的概念，因此选定了一个最接近的

Found a match



名称:	**脑蛋白水解物注射粉针(云南)**
来源:	Source 2
来源编码:	05867
频率:	8247
剂量:	
剂型:	60mg*6瓶/盒
单位:	mg
生产厂商:	国药控股广东粤兴有限公司 云南盟生药业
商品名:	脑蛋白水解物注射粉针

如果候选概念中**无可选匹配**，请用您认为合适的关键词搜索术语库或直接选择"无可映射概念"：

搜索术语库:

脑蛋白水解物注射粉针

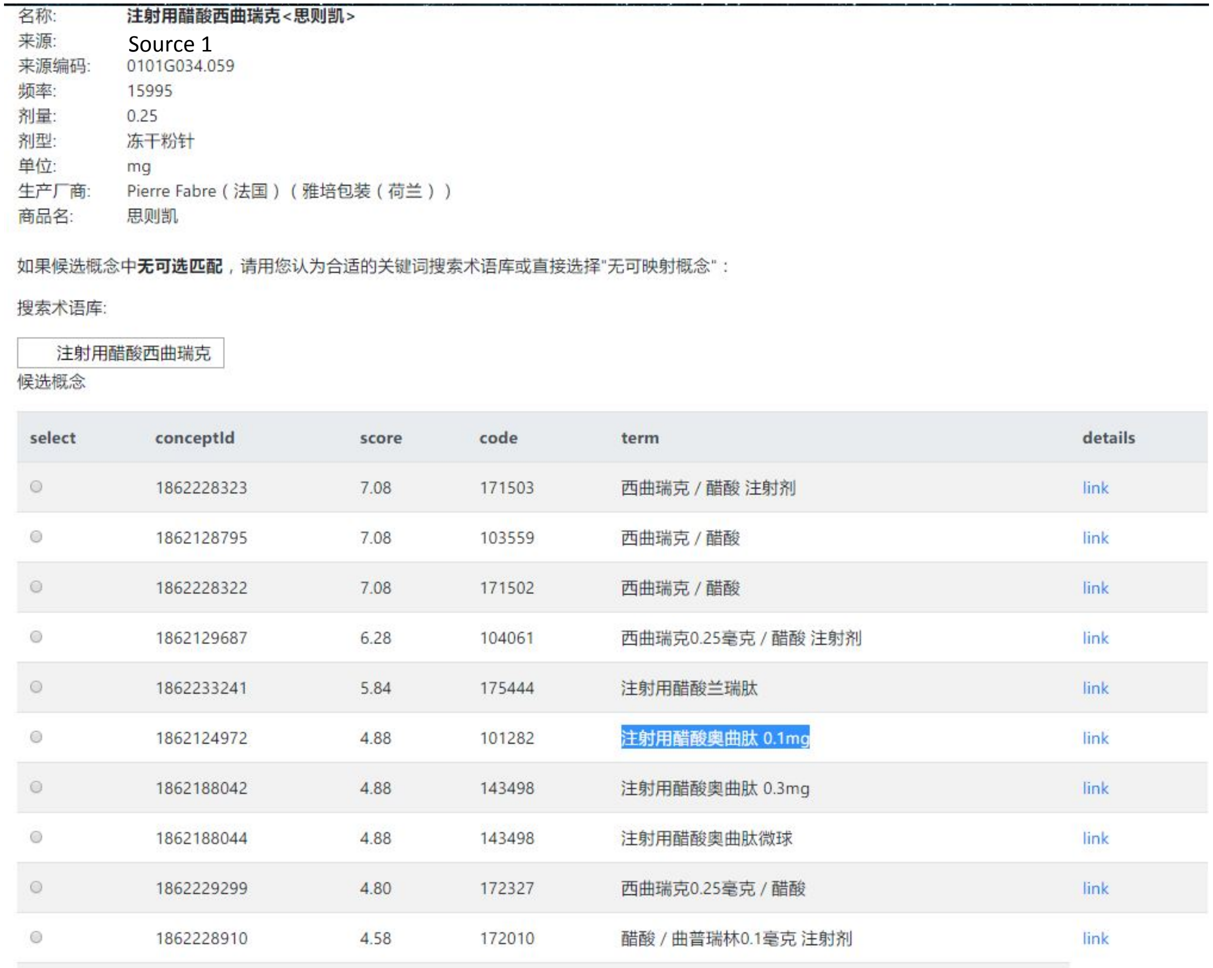
候选概念

select	conceptId	score	code	term	details
○	1862128887	6.91	103627	注射用脑蛋白水解物 30mg	link
●	1862128894	6.91	103629	注射用脑蛋白水解物 60mg	link
○	1862089535	6.82	78034	脑蛋白水解物注射液 2ml	link
○	1862110779	6.82	91939	脑蛋白水解物注射液 10ml	link
○	1862135275	6.82	108295	脑蛋白水解物注射液 5ml	link
○	1862188307	6.82	143691	脑蛋白水解物注射液 20ml	link
○	1862087851	6.78	77133	脑蛋白水解物 注射剂	link
○	1862087852	6.78	77133	脑蛋白水解物 注射剂	link

三、完全没有映射的情况

提供的映射工具没有映射到，自己搜索也没有映射到。

Tried both the tool and manual search, no match



名称:	**注射用醋酸西曲瑞克<思则凯>**
来源:	Source 1
来源编码:	0101G034.059
频率:	15995
剂量:	0.25
剂型:	冻干粉针
单位:	mg
生产厂商:	Pierre Fabre（法国）（雅培包装（荷兰））
商品名:	思则凯

如果候选概念中**无可选匹配**，请用您认为合适的关键词搜索术语库或直接选择"无可映射概念"：

搜索术语库:

注射用醋酸西曲瑞克

候选概念

select	conceptId	score	code	term	details
○	1862228323	7.08	171503	西曲瑞克 / 醋酸 注射剂	link
○	1862128795	7.08	103559	西曲瑞克 / 醋酸	link
○	1862228322	7.08	171502	西曲瑞克 / 醋酸	link
○	1862129687	6.28	104061	西曲瑞克0.25毫克 / 醋酸 注射剂	link
○	1862233241	5.84	175444	注射用醋酸兰瑞肽	link
○	1862124972	4.88	101282	注射用醋酸奥曲肽 0.1mg	link
○	1862188042	4.88	143498	注射用醋酸奥曲肽 0.3mg	link
○	1862188044	4.88	143498	注射用醋酸奥曲肽微球	link
○	1862229299	4.80	172327	西曲瑞克0.25毫克 / 醋酸	link
○	1862228910	4.58	172010	醋酸 / 曲普瑞林0.1毫克 注射剂	link

Questions for Discussion

- Different granularities
- The source codes are only used as references.
  - Customized codes
  - Using existing codes for other purposes
- Consistent mapping guidelines

Next Steps & Deliverables

- Publications
- Mapping service
- Expanded vocabularies

Thank you!