

2003年9月30日
CBI 講義

用語とターミノロジーの情報学

小野木雄三

東京大学大学院医学系研究科

クリニカルバイオインフォマティクス研究ユニット(CBI)

臨床情報工学部門



内容

- なぜ用語が重要なのか
- 統制用語の紹介
 - ICD10
 - MeSH
 - SNOMED-CT
 - UMLS
 - 日本語医学用語
- テキスト検索の基礎

病名・死因統計

文献分類&検索

臨床用語

シソーラスの統合



病名だけを考えてみても...

- ビュルガー病、バージャー氏病、閉塞性血栓血管炎
- 前頭部打撲、前額部打撲、ひたい打撲、眉上部打撲
- 肝臓癌、肝細胞癌、肝癌、肝がん、HCC...
- 子宮頸部と子宮頸部...
- エコノミークラス症候群、肺血栓塞栓症
- 脳溢血、めまい
- 腹痛、かぜ

大江和彦「医療情報標準化における病名コードの意義とその普及」より



何をもって病名とするのか？

目的に応じた分類方法が存在する

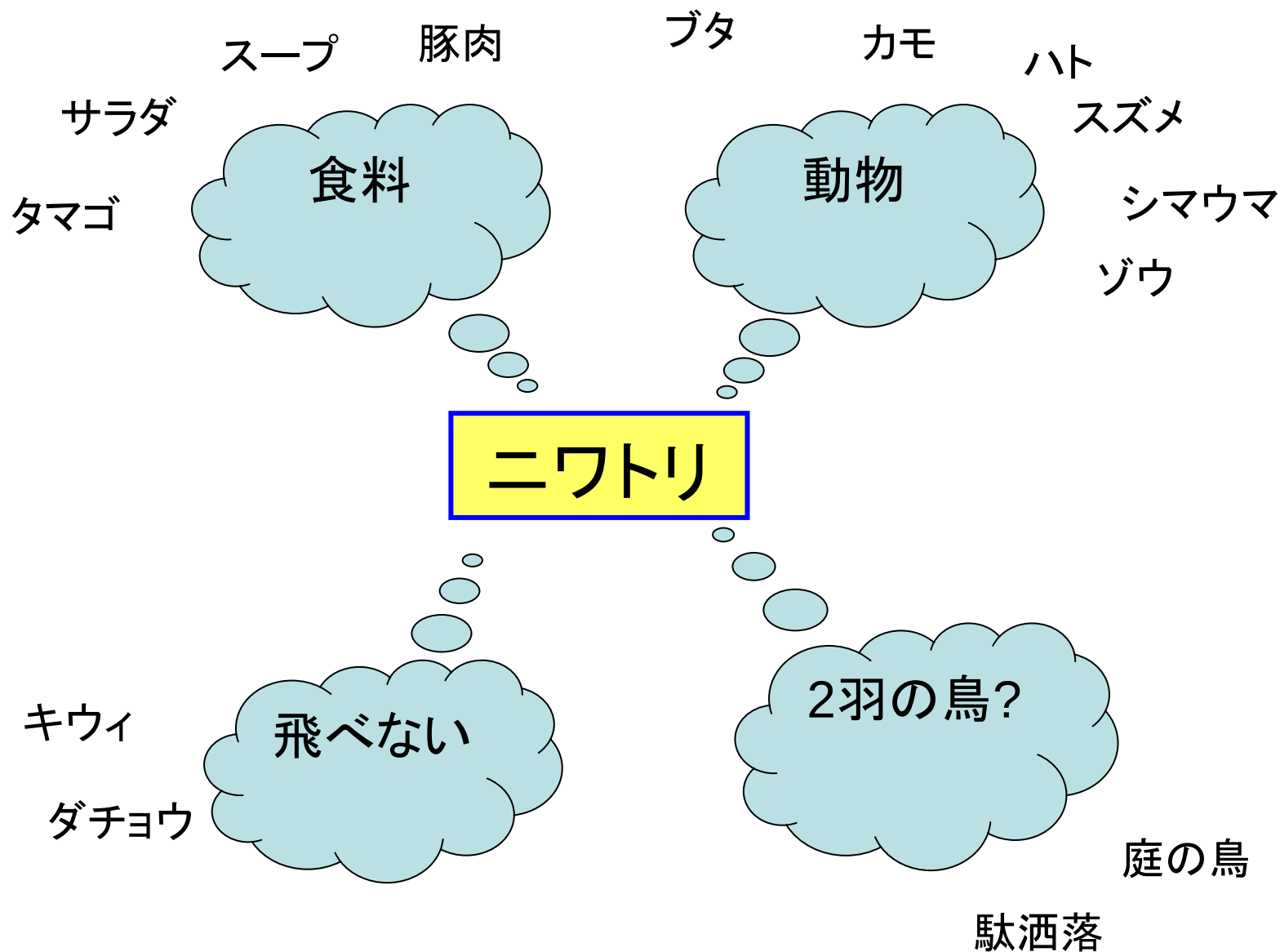
- 治療効果や予後の違いに基づく分類
- 自覚症状の違いによる分類
- 形態的な差異に基づく分類
- 原因、機序(成因)に基づく分類
- 検査結果やその特徴に基づく分類



表記の問題

- 胆嚢/胆のう、 胃癌/胃がん
- 二葉/双葉、 頸部/頸部
- 行う/行なう、 受け付け/受付け/受付
- バイオリン/ヴァイオリン
- パーセント/%、 1月/一月/睦月
- 9月/9月
- 肺がん / lung tumor / Lungen Krebs
- MS / multiple sclerosis / mental status / ...





著作権処理の都合で、
この場所に挿入されていた
『Relativity, lithograph, 1953』の図を
省略させていただきます。

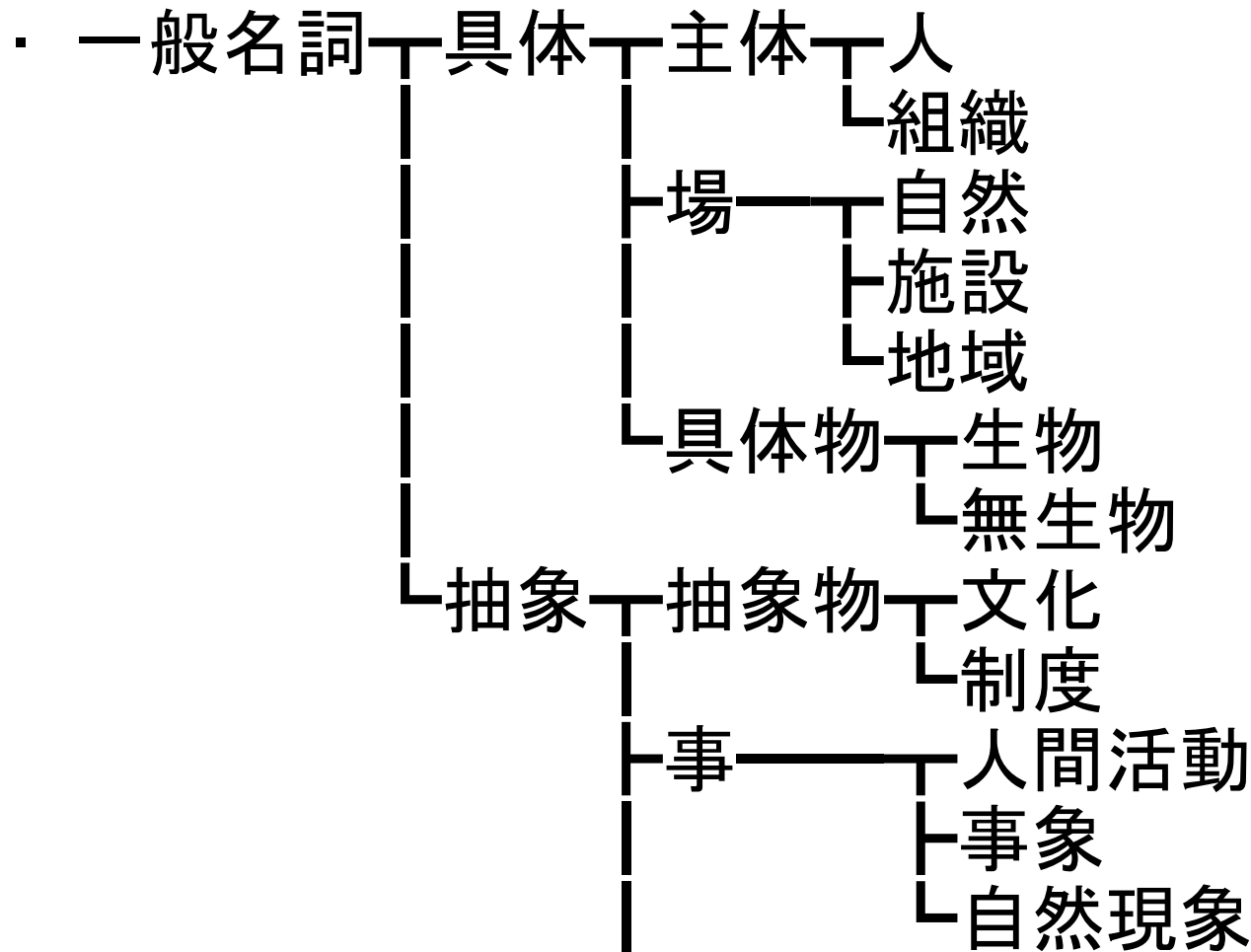


「かける」

が	を	に	英語
主体	時計・絵画・服	壁・釘・枝	hang
主体	橋	土地・川・谷	build
主体	時間・金銭	労力・購入対象	spend
人	腰	椅子	sit down on
人・機械	数	数・価格	multiply
主体	機械	—	start
人	眼鏡	—	wear
主体	鍵・錠	住居・車・金庫	lock
主体	音楽・再生機	—	play
主体	調味料・薬品	人・料理・火	pour on
主体	布	人・身体・家具	spread on



意味属性体系(例)



オントロジー

- 対象世界を記述するために必要なオブジェクト、事象、関係、性質などを概念として表現したものであり、知識工学における対象世界の知識ベースである。
- 言語に依存しない概念と概念間の関係を定義して、上位・下位関係に従って階層的に概念間の制約を記述し、選択制限として意味解析に用いることができる。



診療録の電子化

- 診療情報の共有・転送
 - 人間が解釈して共通理解
 - コンピュータでも共通理解
- 情報の再利用
 - 蓄積した情報から統計的処理・知識発見...
- 人間には一目瞭然 / 機械には理解不能
 - 取得したい情報を明確に特定できるようにしておく必要がある。



情報交換と情報記述

- 属性
- その値
- (時間・対象者・記入者...)

- 属性間の関係を記述：オントロジー



標準化事業

- ICD10対応電子カルテ用標準病名マスタ
- 標準手術・処置マスタ
- 標準医薬品マスタ
- 標準検査項目マスタ
- 標準医療材料データベース
- 症状・診察所見、生理機能検査名・所見、
画像検査名・所見、看護用語・行為、歯科領域
- 電子保存された診療録情報の交換のための
データ項目セット(J-MIX)



データ項目セット(J-MIX)

- | | |
|---------------|---------------|
| 1. 患者基本情報 | 9. 分娩情報 |
| 2. 健康保険・福祉情報 | 10. 診察記録情報 |
| 3. 保険適用傷病名歴情報 | 11. 診療情報交換情報 |
| 4. 診療管理用情報 | 12. 指示実施記録情報 |
| 5. 診療録管理情報 | 13. 診療説明・同意情報 |
| 6. 生活背景情報 | 14. 診療要約情報 |
| 7. 医学的背景情報 | 15. 入院要約情報 |
| 8. 母体情報 | 16. 死亡時記録情報 |



用語の重要性：まとめ

- 概念の分類は対象とする分野に依存する。
- ひとつの概念には様々な表記がありうる。
概念と表記の分離、or 代表語の設定が必要。
- 情報交換では交換対象の指定、および
その値を表現する表記・書式の指定が必要。



統制用語の紹介

- ICD10 病名・死因統計
- MeSH 文献分類&検索
- SNOMED-CT 臨床用語
- UMLS シソーラスの統合



ICD10

- WHOによる国際疾病分類
主として保健行政統計のための分類コード集
- International Classification of Disease and Related Health Problems
- 1992 ICD9の全面改訂版として発行
- 単軸分類、21のカテゴリーから成る分類病名
- 例) J20.7 エコーウイルスによる急性気管支炎
- 疾病統計を目的としているので臨床向けには使いにくい（「その他のウイルス性疾患」など）



感染症及び寄生虫症	A00~B99
新生物	C00~D49
血液および造血器の疾患ならびに免疫機構の障害	D50~D89
内分泌、栄養および代謝疾患	E00~E90
精神および行動の障害	F00~F99
神経系の疾患	G00~G99
眼および付属器の疾患	H00~H59
耳および乳様突起の疾患	H60~H99
循環器系の疾患	I00~I99
呼吸器系の疾患	J00~J99
消化器系の疾患	K00~K99
皮膚および皮下組織の疾患	L00~L99
筋骨格系および結合組織の疾患	M00~M99
尿路性器系の疾患	N00~N99
妊娠、分娩および産褥	O00~O99
周産期に発生した病態	P00~P99
先天奇形、変形および染色体異常	Q00~Q99
症状、徴候および異常臨床所見・異常検査所見で...	R00~R99
損傷、中毒およびその他の外因の影響	S00~T99
傷病および死亡の外因	V00~Y99
健康状態に影響をおよぼす要因および保健サービスの利用	Z00~Z99



ICD10対応電子カルテ標準病名マスター

- ICD10に対応した日本語標準病名
- レセプト電算傷病名マスターに対応

<http://medis.or.jp/> → 「ダウンロード」を選択

病名検索ソフト「病名くん」(<http://www.dis.h.u-tokyo.ac.jp/byoumei>)
標準病名マスター・傷病名マスターの病名・修飾語を検索。

マスター病名検索・階層分類 からも参照可能。
ICD10の分類別に、マスター病名を検索・参照。



MeSH[®]

- Medical Subject Headings
- 米国国立医学図書館NLMで作成された文献インデックスおよび文献検索用の医学用語シソーラス
- 50万語彙、23万概念
- <http://www.nlm.nih.gov/mesh/meshhome.html>



MeSH[®]

南北戦争の頃 Dr. John Shaw Billings

1879 Index Medicus (author & subject)

1951 Standard list of Subject Headings

1958 the Index Mechanization Project (複数topicsの導入)

1960 MEDLARS 開始、1964に完成 (郵便による文献配送)
(the Medical Literature Analysis and Retrieval System)

1963 MeSH database完成、毎年更新、階層構造と多階層

Tuberculosis, Hepatic

Infectious Diseases

→ Tuberculosis

Digestive System Diseases

→ Liver Diseases

1968 real-time (SUNY, Typewriter Exchange)

1971 MEDLINE (MEDLARS Online)



初期のMEDLINE®

- 単数(Eye, Hand, Foot)と複数(Fingers, Toes)の区別
- 名詞一形容詞(Diet, Reducing 体重減量食)と形容詞一名詞(Diabetic Diet 糖尿病食)の区別
- Subheadingの指定(Hypertension/Diet Therapy)
- ANDの指定(Hypertension AND Diabetic Diet)
- 論文誌の指定(New Engl J Med, JAMA)



MEDLINE Database

- インデクシング: 論文を分析してMeSHを8-10個付ける。うち2-4個がmajor focus、(Index Medicus)に印刷される。
- Headingsにはsubheadingsが付く。
例えばアルドメットの副作用に関する論文であれば、
Methyldopa / Adverse Effects となる。
- 他の識別子としてpublication typesとチェックタグを付す。
Historical Article, Review, Letter to the Editor, Guideline, etc.
Human, Animal, Male, Female, etc.



☐ Hematologic and Lymphatic Diseases	0	☐	☐	
+ ☐ Congenital, Hereditary, and Neonatal Diseases and Abnormalities	61	☐	☐	
+ ☐ Skin and Connective Tissue Diseases	0	☐	☐	
- ☐ Nutritional and Metabolic Diseases	0	☐	☐	
- ☐ Metabolic Diseases	808	☐	☐	
☐ Achlorhydria	59	☐	☐	
+ ☐ Acid-Base Imbalance	239	☐	☐	
+ ☐ Amyloidosis	2583	☐	☐	
+ ☐ Brain Diseases, Metabolic	211	☐	☐	
+ ☐ Calcium Metabolism Disorders	80	☐	☐	
+ ☐ Diabetes Insipidus	542	☐	☐	
- ☒ Diabetes Mellitus	14335	☐	☐	
+ ☐ Diabetes, Gestational	1211	☐	☐	
☐ Diabetes Mellitus, Experimental	5181	☐	☐	
+ ☐ Diabetes Mellitus, Insulin-Dependent	13049	☐	☐	
☐ Diabetes Mellitus, Lipoatrophic	55	☐	☐	
☐ Diabetes Mellitus, Non-Insulin-Dependent	16561	☐	☐	
+ ☐ Diabetic Angiopathies	2982	☐	☐	
+ ☐ Diabetic Coma	75	☐	☐	
☐ Diabetic Ketoacidosis	688	☐	☐	
☐ Diabetic Nephropathies	4087	☐	☐	
+ ☐ Diabetic Neuropathies	2352	☐	☐	
☐ Obesity in Diabetes	1157	☐	☐	
☐ Prediabetic State	211	☐	☐	
+ ☐ Pregnancy in Diabetics	1255	☐	☐	
☐ Glucose Intolerance	1427	☐	☐	
+ ☐ Glycosuria	300	☐	☐	
☐ Hyperammonemia	158	☐	☐	

Subheadings for: **Diabetes Mellitus**

☐ **Include All Subheadings (14335)**

-- or choose one or more of these subheadings --

- | | |
|---|---|
|  ☐ /bl - Blood (1668) |  ☐ /mi - Microbiology (39) |
|  ☐ /cf - Cerebrospinal Fluid (1) |  ☐ /mo - Mortality (349) |
|  ☐ /ci - Chemically Induced (217) |  ☐ /nu - Nursing (213) |
|  ☐ /cl - Classification (208) |  ☐ /ps - Parasitology (2) |
|  ☐ /co - Complications (5218) |  ☐ /pa - Pathology (296) |
|  ☐ /cn - Congenital (20) |  ☐ /pp - Physiopathology (1546) |
|  ☐ /di - Diagnosis (1110) |  ☐ /pc - Prevention & Control (815) |
|  ☐ /dh - Diet Therapy (201) |  ☐ /px - Psychology (450) |
|  ☐ /dt - Drug Therapy (1235) |  ☐ /ra - Radiography (15) |
|  ☐ /ec - Economics (351) |  ☐ /ri - Radionuclide Imaging (20) |
|  ☐ /em - Embryology (6) |  ☐ /rt - Radiotherapy (1) |
|  ☐ /en - Enzymology (129) |  ☐ /rh - Rehabilitation (192) |
|  ☐ /ep - Epidemiology (2020) |  ☐ /su - Surgery (152) |
|  ☐ /eh - Ethnology (285) |  ☐ /th - Therapy (1838) |
|  ☐ /et - Etiology (749) |  ☐ /us - Ultrasonography (28) |
|  ☐ /ge - Genetics (719) |  ☐ /ur - Urine (129) |
|  ☐ /hi - History (99) |  ☐ /ve - Veterinary (138) |
|  ☐ /im - Immunology (198) |  ☐ /vi - Virology (20) |
|  ☐ /me - Metabolism (1034) | |



SNOMED[®]-CT

- SNOMED = Systematized Nomenclature of Medicine
- 米国臨床病理医協会が作成・管理
CAP: the College of American Pathologists
- 1965年SNOP Systematized Nomenclature of Pathology
- 1974年に医学全般の用語を含めSNOMEDという名称に
- 2000 SNOMED-RT (Reference Terminology)発表
- これに Clinical Terms Version 3 (ReadCode)
(英国保健省の監修した診療用病名シソーラス)を統合して
SNOMED-CT (Clinical Terms)が完成
87万語彙、34万概念という世界最大規模の医学用語集
(MeSHは50万語彙、23万概念)



SNOMED-CTの語彙と概念

- SNOMEDには、3要素がある。
 - 概念(Concept)
 - 記述子(Description)
 - 関係 (Relationship)
- 「記述子」は「概念」の表現形。
 - 同義語はここに入る。
- 1つの概念は複数の記述子を持つ。
 - 1つの記述子が複数の概念を持つこともある。



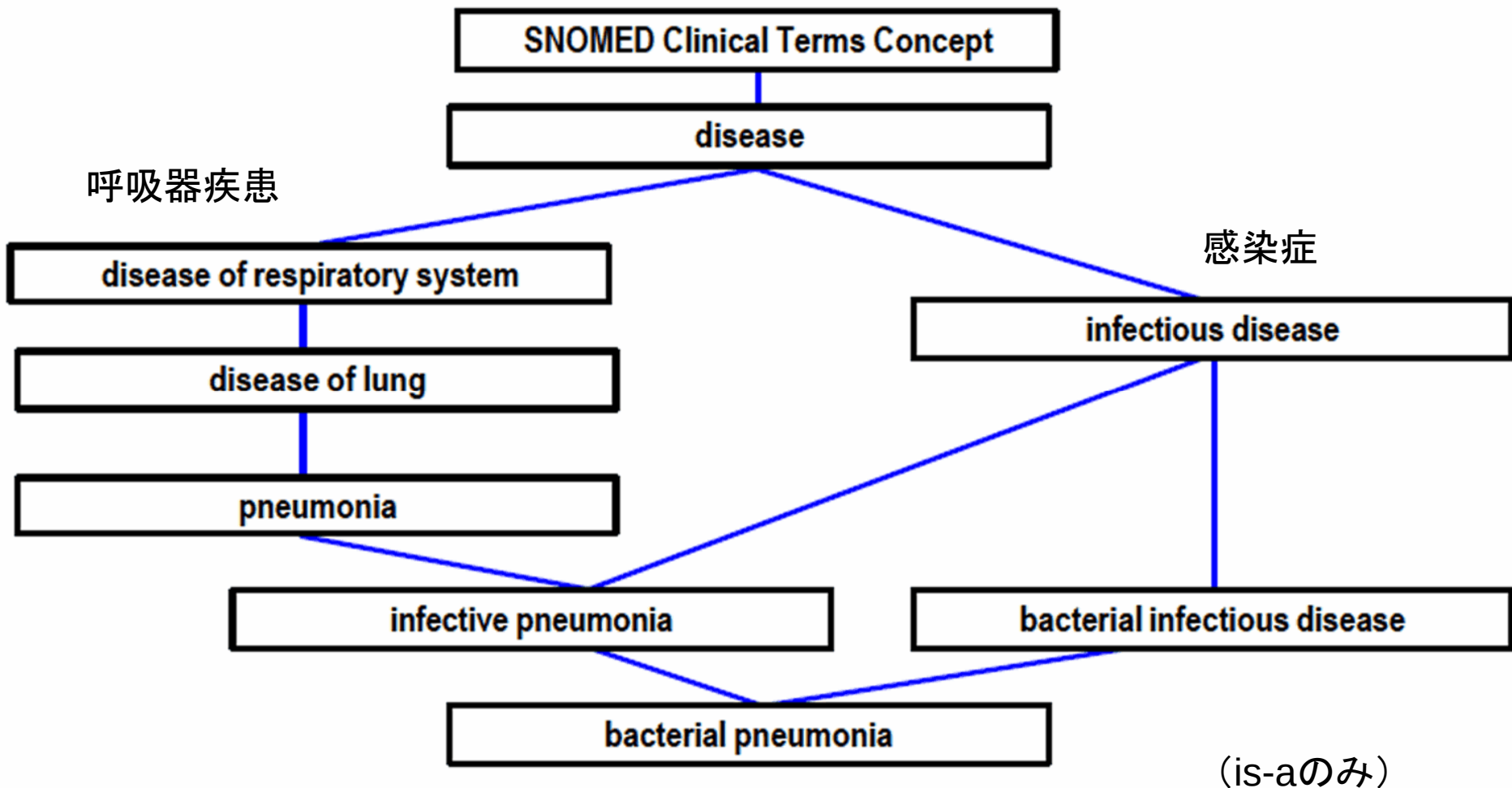
SNOMED-CT概念間の関係

- 概念は、他の概念と「関係」する。
例えばIS-Aという関係は、「～はある種の～(is a kind of)」という意味。
「C型肝炎IS-A感染症」、「C型肝炎IS-A肝臓疾患」など階層構造を作る。
IS-A以外にも様々な関係があり、合計42種類が定義されている。
 部位(FINDING-SITE)、
 原因(CAUSATIVE-AGENT)など
- 「関係」は記述子を結びつけるのではなく、概念と概念を結びつける。
 SNOMED-CTでは、概念志向(Concept Oriented)用語集である。
- 全ての概念は「SNOMED CT Concept」の下位概念になっている。
 直下には、19のカテゴリー
 所見 Finding、疾患 Disease、処置 Procedure」など



Supertypes of bacterial pneumonia	
bacterial pneumonia <i>is a</i> bacterial infectious disease	
	bacterial infectious disease <i>is a</i> infectious disease
	infectious disease <i>is a</i> disease
	disease <i>is a</i> SNOMED Clinical Terms Concept
bacterial pneumonia <i>is a</i> infective pneumonia	
	infective pneumonia <i>is a</i> pneumonia
	pneumonia <i>is a</i> disease of lung
	disease of lung <i>is a</i> disease of respiratory system
	disease of respiratory system <i>is a</i> disease
	disease <i>is a</i> SNOMED Clinical Terms Concept
	infective pneumonia <i>is a</i> infectious disease
	infectious disease <i>is a</i> disease
	disease <i>is a</i> SNOMED Clinical Terms Concept





SNOMED-CT で定義された「関係」

Associated morphology

Is a

Has specimen

Part of

Subject of information

Causative agent

Associated finding

Component

Onset

Severity

Occurrence

Episodicity

Revision status

Access

Approach

Method

Priority

Course

Using

Laterality

Finding site

Direct device

Direct morphology

Direct substance

Has focus

Has intent

Procedure site

Has definitional manifestation

Temporally follows

Indirect morphology

Indirect device

Has interpretation

Interprets

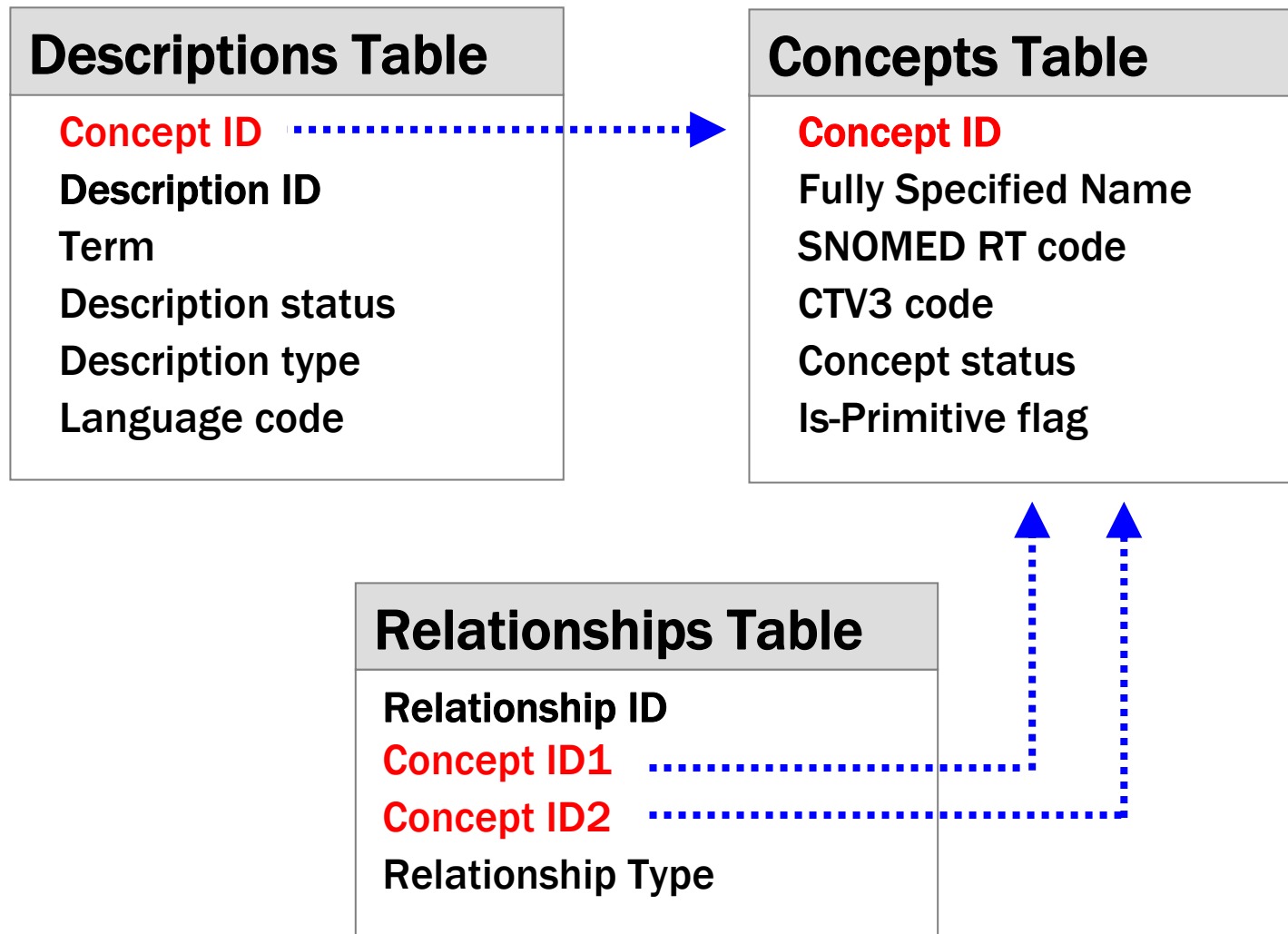
Associated etiologic finding

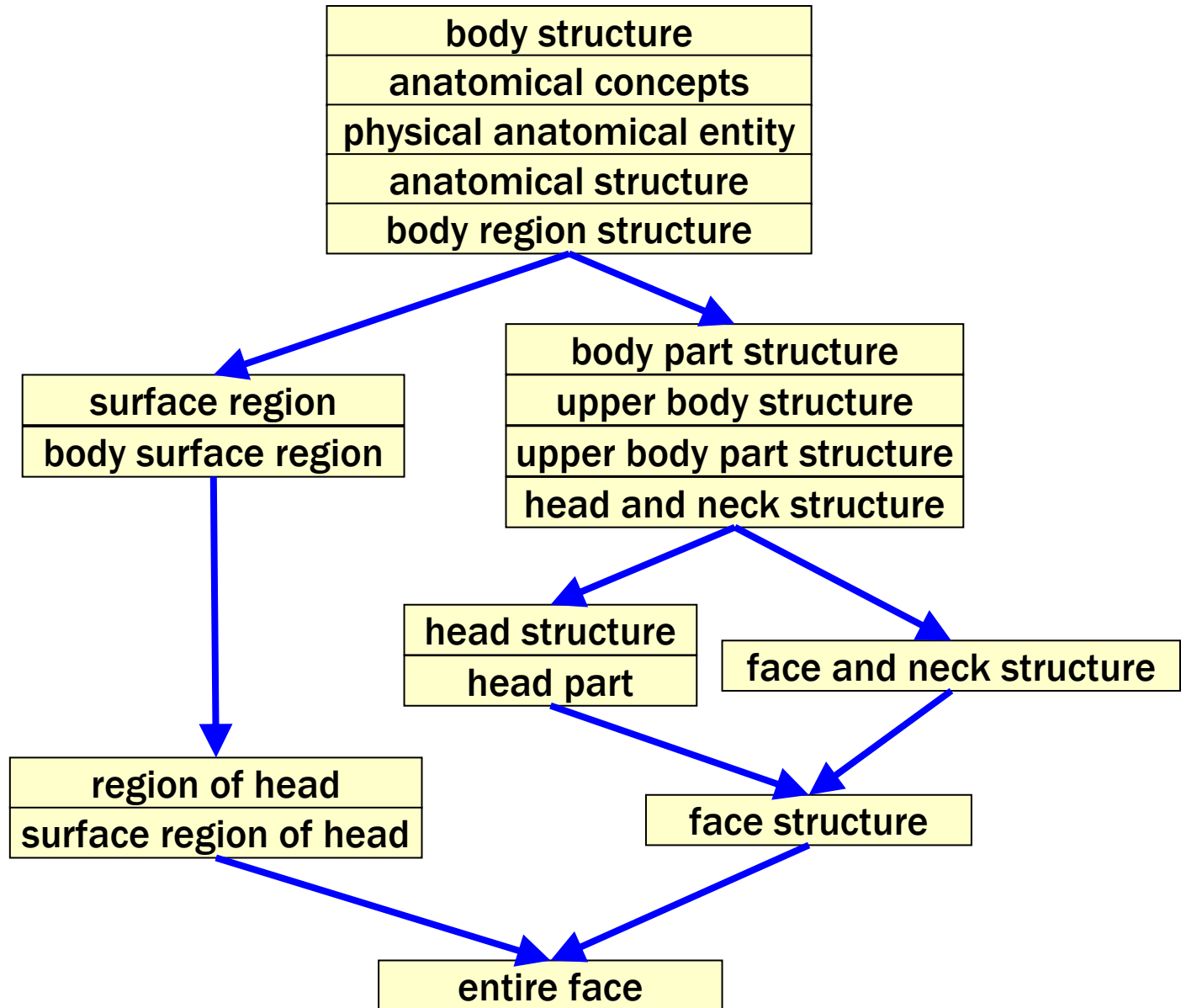
Access instrument

Recipient category

Pathological process







UMLS[®]

- Unified Medical Language System[®]
 - 目標は、生物医学に関連する様々な文書から機械が読める情報を検索・統合できるようにすること。
 - 用語集(典拠)ごとに異なる語彙と定義の多様性、およびそれらの関連性が大きな障壁。
 - 1986年に始まったNLMのproject
- Metathesaurus[®]: 語彙を概念に分類(200万>80万)
- Semantic Network: 一般的な意味関係(188種)
- SPECIALIST lexicon: 統語情報
- 応用: NLM gateway, PubMed[®], Indexing Initiative
- <http://www.nlm.nih.gov/research/umls/umlsmain.html>



Metathesaurus[®]: 言語と語彙数

BAQ	695
DAN	723
DUT	36,491
ENG	1,753,789
FIN	21,086
FRE	36,556
GER	67,987
HEB	485

HUN	718
ITA	23,602
NOR	722
POR	45,711
RUS	42,346
SPA	51,469
SWE	723

(15か国)



MSH	502,729	<u>MeSH</u>
RCD99	347,568	<u>Clinical Terms v3 (Read Codes)</u>
SNMI98	164,179	<u>SNOMED</u> v3.5, 1998
NCBI2001	136,466	National Center for Biotechnology Information Taxonomy
LNC205	79,522	<u>Logical Observation Identifiers, Names and Codes</u> v2.05
UWDA155	79,463	University of Washington Digital Anatomist v1.55
MDR50	73,401	Medical Dictionary for Regulatory Activities Terminology
DMD2002	48,064	German Translation of <u>MeSH</u>
SNM2	44,274	<u>SNOMED</u> update, 1982
RUS2002	42,354	Russian Translation of <u>MeSH</u>
BRMP2002	42,094	Portuguese Translation of <u>MeSH</u>
BRMS2002	40,499	Spanish Translation of <u>MeSH</u>
MMSL01	38,789	Multum MediSource Lexicon
DUT2001	36,391	Dutch Translation of <u>MeSH</u>
...	...	...
NAN99	169	Classification of Nursing Diagnoses
ULT93	84	Ultrasound Structured Attribute Reporting



Source	語彙数	概念数
MSH	502,729	227,733
RCD99	347,568	190,024
SNMI98	164,179	113,658
NCBI2001	136,466	107,586
LNC205	79,522	42,778
UWDA155	79,463	53,412
MDR50	73,401	38,864
DMD2002	48,064	27,300
SNM2	44,274	35,434
RUS2002	42,354	20,564
BRMP2002	42,094	27,288
BRMS2002	40,499	27,398
MMSL01	38,789	34,367
DUT2001	36,391	17,787
...	...	...
NAN99	169	169
ULT93	84	84

総語彙数 2,083,103

総概念数 870,853

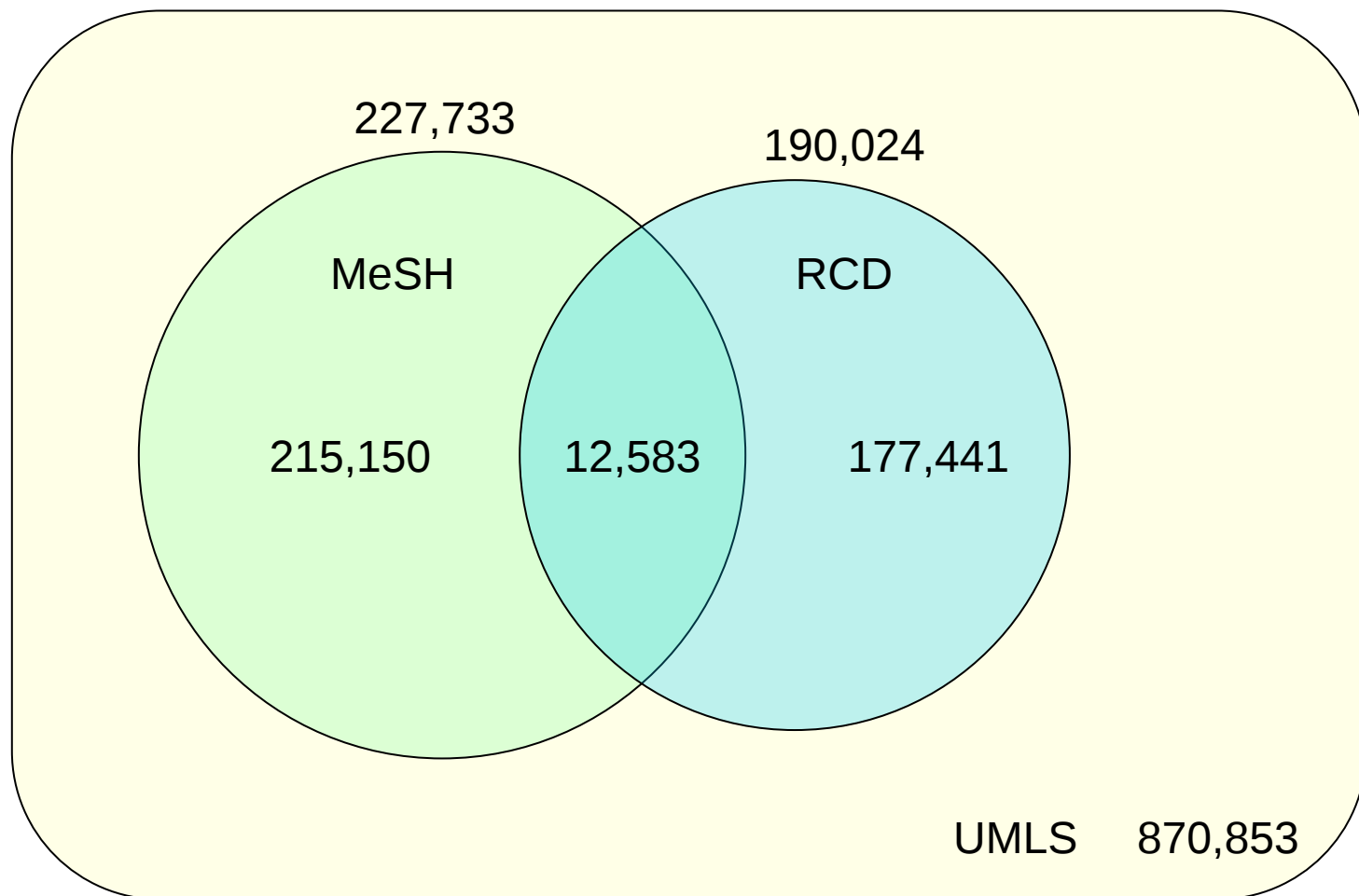


LOINC[®]

- Logical Observation Identifiers, Name and Codes
- 検査項目を分析物・成分、結果属性、時間、部位・材料、スケール(定性・定量)、検査方法によって分類する多軸分類。
- 見出し語はSNOMEDと統一されている。
- 心電図・超音波などの生理検査結果を記述するために、身体所見などの項目を収載したClinical LOINCもある。



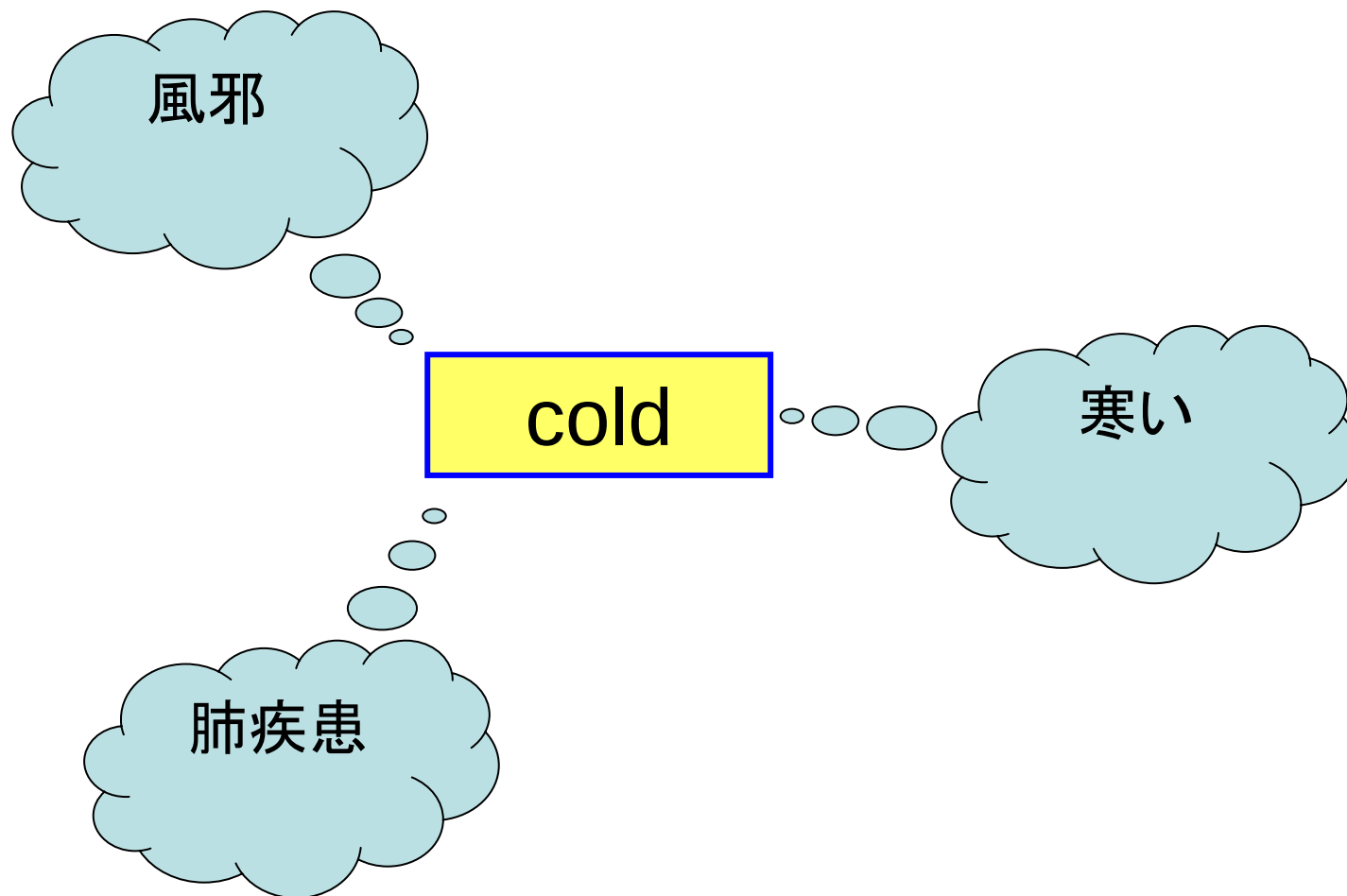
概念の重なり具合



Metathesaurus CUI-LUI-SUI

Concept (<u>CUI</u>)	Terms (<u>LUIs</u>)	Strings (<u>SUIs</u>)
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	L0004238 Atrial Fibrillation (preferred) Atrial Fibrillations	S0016668 Atrial Fibrillation (preferred)
		S0016669 Atrial Fibrillations
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	S0016899 Auricular Fibrillation (preferred)
		S0016900 (plural variant) Auricular Fibrillations





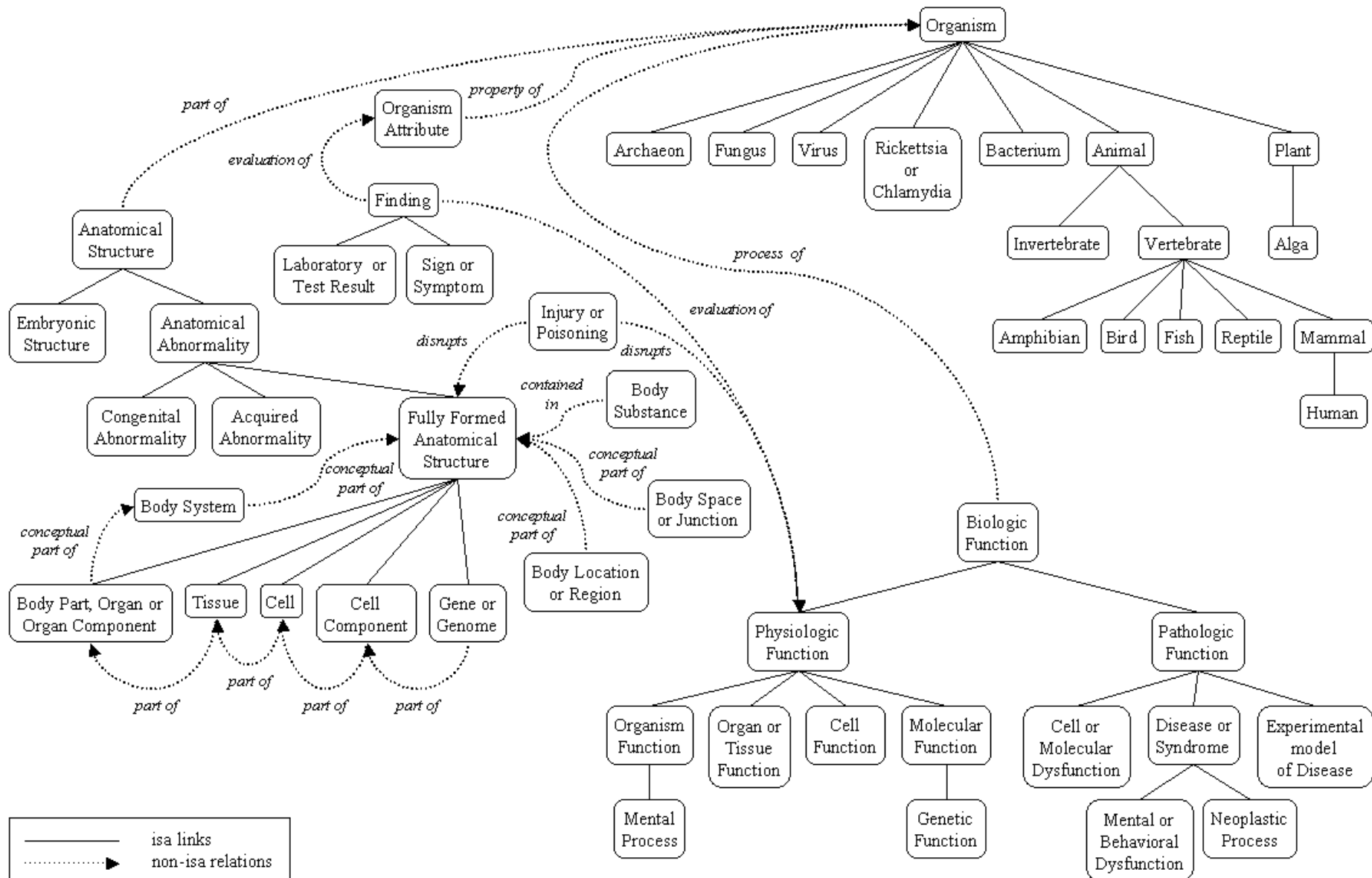
Chronic Obstructive Lung Disease



Concepts (CUIs)	Terms (LUIs)	Strings (SUIs)
C0009264 cold temperature (低温度・冷氣)	L0215040 cold temperature	S0288775 cold temperature
	L0009264 Cold <1> Cold	S0007170 Cold <1>
		S0026353 Cold
C0009443 Common Cold (感冒・風邪)	L0009443 Common Cold	S0026747 Common Cold
	L0009264 Cold <2> Cold	S0007171 Cold <2>
		S0026353 Cold
C0024117 Chronic Obstructive Airway Disease (慢性閉塞性肺疾患)	L0498186 Chronic Obstructive Airway Disease	S0837575 Chronic Obstructive Airway Disease
	L0008703 Chronic Obstructive Lung Disease	S0837576 Chronic Obstructive Lung Disease
	L0009264 COLD <3> COLD	S0829315 COLD <3>
		S0474508 COLD



Semantic Networks

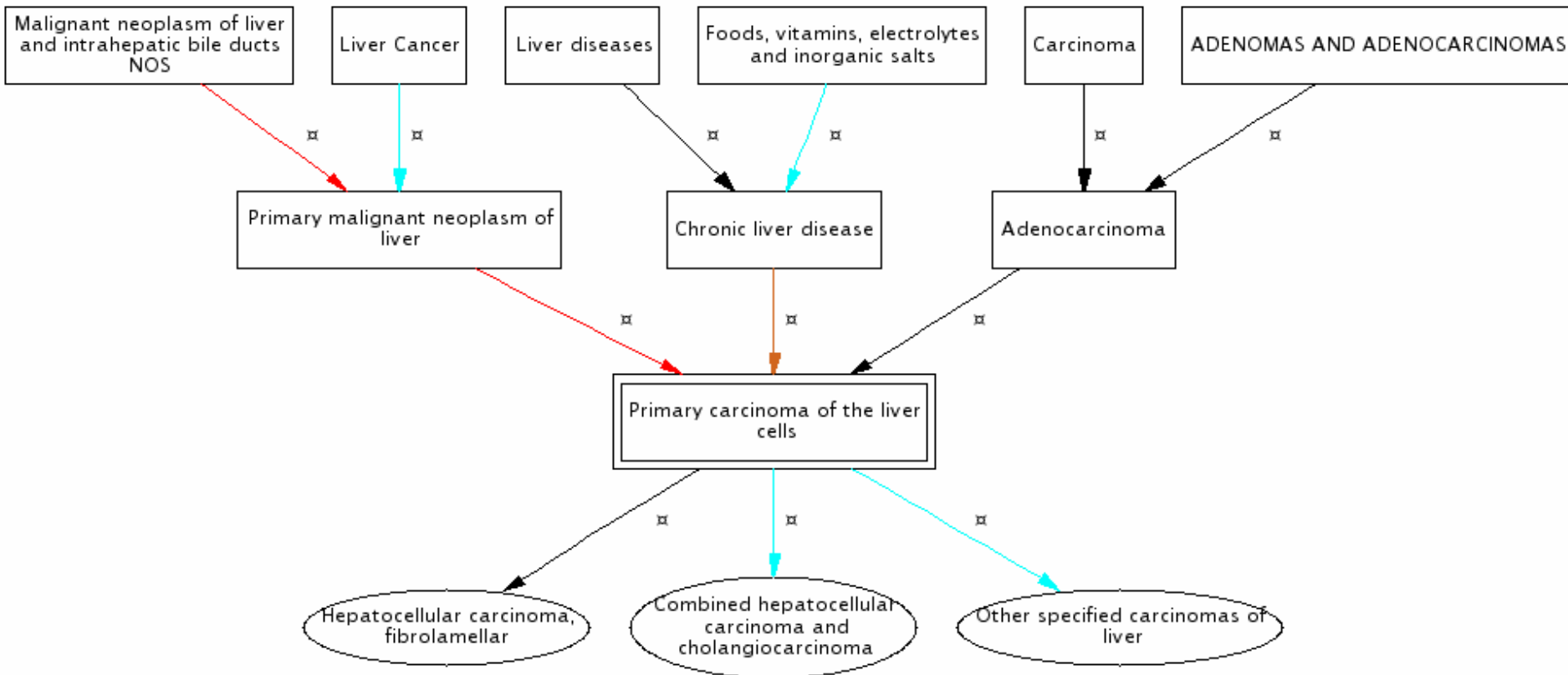


isa
 associated_with
 physically_related_to
 part_of
 consists_of
 contains
 connected_to
 interconnects
 branch_of
 tributary_of
 ingredient_of
 spatially_related_to
 location_of
 adjacent_to
 surrounds
 traverses
 functionally_related_to
 affects
 manages
 treats
 disrupts
 complicates
 interacts_with
 prevents
 brings_about
 produces
 causes

[associated_with] (continued)
 [functionally_related_to] (continued)
 performs
 carries_out
 exhibits
 practices
 occurs_in
 process_of
 uses
 manifestation_of
 indicates
 result_of
 temporally_related_to
 co_occurs_with
 precedes
 conceptually_related_to
 evaluation_of
 degree_of
 analyzes
 assesses_effect_of
 measurement_of
 measures
 diagnoses
 property_of
 derivative_of
 developmental_form_of
 method_of
 conceptual_part_of
 issue_in



Semantic navigator



Haversian System = 骨単位 (コツタンイ)
(A10.165.265.507, C0824691, S0046764, 02606)



MeSH node

UMLS

統制語:Havers系/ハバース系

辞書	code	典拠
骨単位 (コツタンイ)	0.000	1) MeSH 1997 医学件名標目表 英和版 2) 解剖学辞典
オステオン (オステオン)	1.001	1) 解剖学辞典

以下参考情報

辞書	code	典拠
Havers系 (Haversケイ)	2.001	1) MeSH 1997 医学件名標目表 英和版
Haversian System (Haversian System)	2.002	1) 医学用語シソーラス 第3版 2) MeSH ANNOTATED ALPHABETIC LIST 1997 3) MeSH ANNOTATED ALPHABETIC LIST 2001
Osteon (Osteon)	2.003	1) MeSH ANNOTATED ALPHABETIC LIST 1997 2) MeSH ANNOTATED ALPHABETIC LIST 2001
ハバース系 (ハバースケイ)	2.004	1) MeSH 1997 医学件名標目表 英和版 2) 日本医学会医学用語辞典 英和

CUI-Source and Definition

C0824691-MSH2002_06_01

A circular structural unit of bone tissue. It consists of a central hole, the Haversian canal through which blood vessels run, surrounded by concentric rings, called lamellae.

STR in the same CUI

STR	TS	STT
Osteon	P	PF
osteon	P	VC
Osteon, NOS	P	VO
Osteons	P	VP
Haversian System	S	PF
Osteonum	S	PF
haversian system	S	VC
Haversian system	S	VC
System, Haversian	S	VW

MeSH entries for same CUI

MeSH entries for same CUI	TRACE
A10.165.265.507	4:Haversian System 3:Bone and Bones 2:Connective Tissue 1:Tissues



JapaneseTerm	CUI	UMLS term(P and PF only)
肝臓-人工	C0336562	Liver, Artificial
肝臓エキス	C0023899	Liver Extract
肝臓グリコーゲン	C0023902	Liver Glycogen
肝臓グリコゲン	C0023902	Liver Glycogen
肝臓ジストマ	C0009022	Clonorchis sinensis <1>
肝臓移植	C0023911	Transplantation of liver
肝臓学	C0017163	Gastroenterology
肝臓機能検査	C0023901	Liver Function Tests
肝臓結核	C0041313	Tuberculosis, Hepatic
肝臓検査	C0023901	Liver Function Tests
肝臓梗塞	C0151731	Hepatic infarction
肝臓紫斑病	C0030781	Peliosis Hepatis
肝臓疾患	C0023895	Liver diseases
肝臓腫瘍	C0023903	Liver neoplasms
肝臓症	C0877372	Hepatositis
肝臓抽出物	C0023899	Liver Extract
肝臓摘除	C0040507	Total hepatectomy



	STR in the same CUI	TS	STT	Source	
JapaneseTerm	Liver neoplasms	P	PF	MTH	UMLS term(P and PF only)
肝臓-人工	Liver Neoplasms	P	VC	MSH2002_06_01	Liver, Artificial
肝臓エキス	Neoplasms, Liver	P	VCW	MSH2002_06_01	Liver Extract
肝臓グリコーゲン	Neoplasm of Liver	P	VO	COS95	Liver Glycogen
肝臓グリコゲン	Liver neoplasm	P	VS	MDR50	Liver Glycogen
肝臓ジストマ	Liver Neoplasm	P	VS	MSH2002_06_01	Clonorchis sinensis <1>
肝臓移植	liver neoplasm	P	VS	AOD2000 CSP2002	Transplantation of liver
肝臓学	Neoplasm, Liver	P	VWS	MSH2002_06_01	Gastroenterology
肝臓機能検査	Neoplasm liver	P	VWS	MDR50	Liver Function Tests
肝臓結核	NEOPLASM LIVER	P	VWS	CST95	Tuberculosis, Hepatic
肝臓検査	Tumor of liver	S	PF	RCDAE	Liver Function Tests
肝臓梗塞	Hepatic neoplasia	S	PF	MDR50	Hepatic infarction
肝臓紫斑病	Hepatoma <2>	S	PF	MTH	Peliosis Hepatis
肝臓疾患	Neoplasms, Hepatic	S	PF	MSH2002_06_01	Liver diseases
肝臓腫瘍	Hepatic tumor	S	PF	RCDAE	Liver neoplasms
肝臓症	liver tumor or cancer	S	PF	AOD2000	Hepatosi
肝臓抽出物	HEPATIC NEOPLASIA	S	VC	CST95	Liver Extract
肝臓摘徐					Total hepatectomy

2003/9/30

48



統制用語の紹介：まとめ

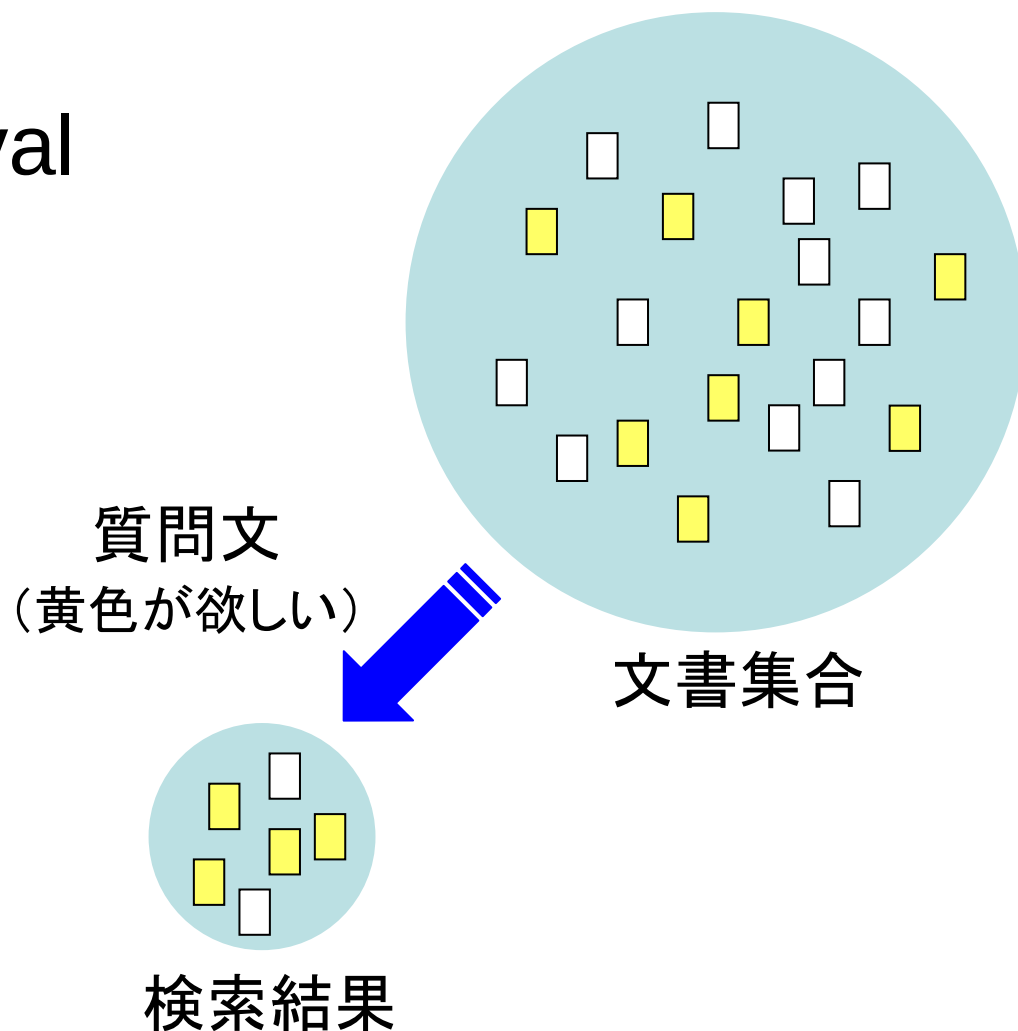
- ICD10 病名・死因統計
- MeSH 文献分類&検索
- SNOMED-CT 臨床用語
- UMLS シソーラスの統合
- 日本語医学用語
 - ICD10対応電子カルテ用標準病名マスタ
 - 医学用語シソーラス(MeSHに対応)



テキスト検索の基本手法

Information Retrieval

- ブーリアンモデル
- ベクトルモデル
- 確率モデル



記号の定義

文書は特定の用語（インデックスターム）であらかじめ索引付けされているものとする。

- インデックスターム（語） $K = \{k_1, k_2, \dots, k_t\}$
- 文書とタームで決まる重み $w_{i,j}(k_i, d_j)$
- 文書ベクトル $\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$



精度 (precision)と再現率 (recall)

	True: Relevant	True: Not-Relevant
Algorithm: Relevant	True Positive (TP)	False Positive (FP)
Algorithm: Not Relevant	False Negative (FN)	True Negative (TN)

何かを検索した結果の評価基準

$$\text{精度} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

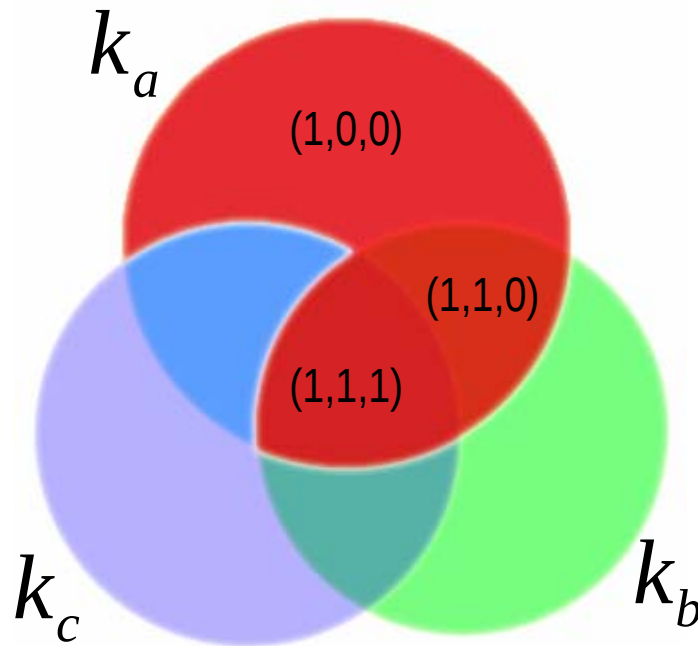
完璧な検索アルゴリズムはFP=FN=0



ブーリアンモデル

- 質問: a かつ (b または $\neg c$)

$$\begin{aligned} q &= k_a \wedge (k_b \vee \neg k_c) \\ &= (1, 1, 1) \vee (1, 1, 0) \vee (1, 0, 0) \end{aligned}$$



文書の重み付けが
インデックス語の有無、
つまり 0 か 1 だけに依
存するモデル



ブーリアンモデル

- 各文書が、各タームに「関連するかしないか」しか記述することができない。
- 「部分的に一致する」は不可能。
例えば k_b を含む文書 $d_j = (0, 1, 0)$ は先の *query* に関係なしとされる。
- 結果 → あまりにも大量 or あまりに僅か



ベクトルモデル

- 対象を良く記述する要素: intra-cluster

文書 d_j に含まれる用語 k_i の頻度

- 対象を他と区別する要素: inter-cluster

文書集合全体での用語 k_i の頻度の逆数



ベクトルモデル

- 文書 d_j に含まれる用語 k_i の頻度

term frequency (tf)

文書内の用語 l に対して

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$$

- 全文書中で用語 k_i を含む文書数の逆数

inverse document frequency (idf)

N は全文書数、
 n_i は k_i を含む文書数

$$idf_i = \log \frac{N}{n_i}$$

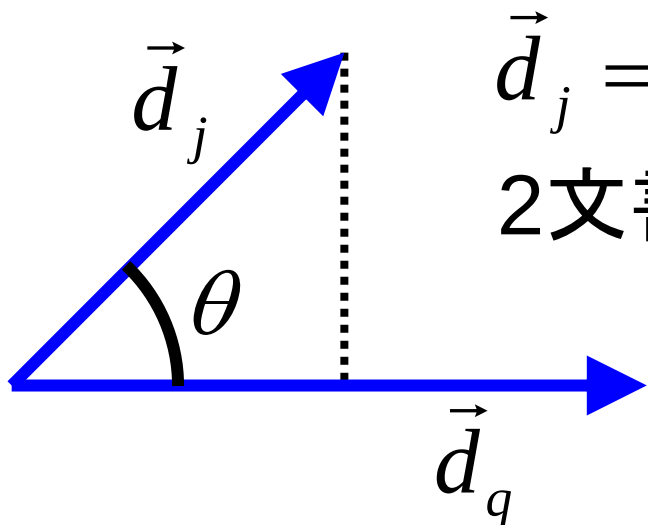


ベクトルモデル

- 両者を掛け合わせたものを重みにする。

tf-idf 法

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$$



$$\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

2文書の類似度 = 2ベクトルの内積

$$\vec{d}_j \bullet \vec{d}_q = |\vec{d}_j| \times |\vec{d}_q| \times \cos(\theta)$$



ベクトルモデル

- 検索の性能向上
- 質問文に類似した文書検索が可能
- 類似度の高い順に呈示できる
- 現在でも一般的な検索モデルである



確率モデル

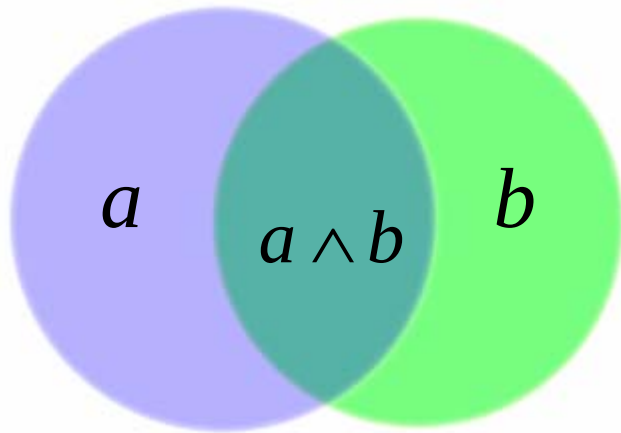
- $w_{i,j}$ (文書における語の重み)はbinary
- R を質問 q に関連する既知の文書集合、
- $\bar{R}$ をその補集合とする(関連のない文書)。
- $P(R | \vec{d}_j)$ を文書 $\vec{d}_j$ が質問に関連する確率
- $P(\bar{R} | \vec{d}_j)$ を文書 $\vec{d}_j$ が質問に関連しない確率

$$\text{類似度 } sim(d_j, q) = \frac{P(R | \vec{d}_j)}{P(\bar{R} | \vec{d}_j)}$$



条件付確率

$p(a|b)$: b という条件のもとで a である確率



$$p(a \wedge b) = p(a)p(b | a)$$

$$p(a \wedge b) = p(b)p(a | b)$$

$$\therefore p(a | b) = \frac{p(a)p(b | a)}{p(b)}$$

(Bayes の定理)



確率モデル

$$\text{sim}(d_j, q) = \frac{P(R | \vec{d}_j)}{P(\bar{R} | \vec{d}_j)} = \frac{P(\vec{d}_j | R) \times P(R)}{P(\vec{d}_j | \bar{R}) \times P(\bar{R})} \quad (\text{Bayes})$$



$$\text{sim}(d_j, q) \approx \frac{P(\vec{d}_j | R)}{P(\vec{d}_j | \bar{R})} = \frac{\left(\prod_{w_{i,j}=1} P(k_i | R) \right) \times \left(\prod_{w_{i,j}=0} P(\bar{k}_i | R) \right)}{\left(\prod_{w_{i,j}=1} P(k_i | \bar{R}) \right) \times \left(\prod_{w_{i,j}=0} P(\bar{k}_i | \bar{R}) \right)}$$



$$\text{sim}(d_j, q) \approx \sum_{i=1}^t w_{i,q} w_{i,j} \left(\log \frac{P(k_i | R)}{1 - P(k_i | R)} \right) \left(\log \frac{1 - P(k_i | \bar{R})}{P(k_i | \bar{R})} \right)$$



確率モデル

- $P(k_i | R)$ を、検索された文書中で k_i を含む割合にて代用する(関連文書→検索結果)。
- $P(k_i | \bar{R})$ を、検索されなかった文書中で k_i を含む割合にて代用する(非関連文書→検索されなかった文書)。これを再帰的に繰り返す。

$$\therefore P(k_i | R) = \frac{V_i}{V} \quad P(k_i | \bar{R}) = \frac{n_i - V_i}{N - V}$$



確率モデル

- 利点
 - 質問への類似度が高い順に呈示できる
- 欠点
 - 最初の推測(Rの設定)が任意
 - 重みがbinaryである
 - ベクトルモデルに劣る

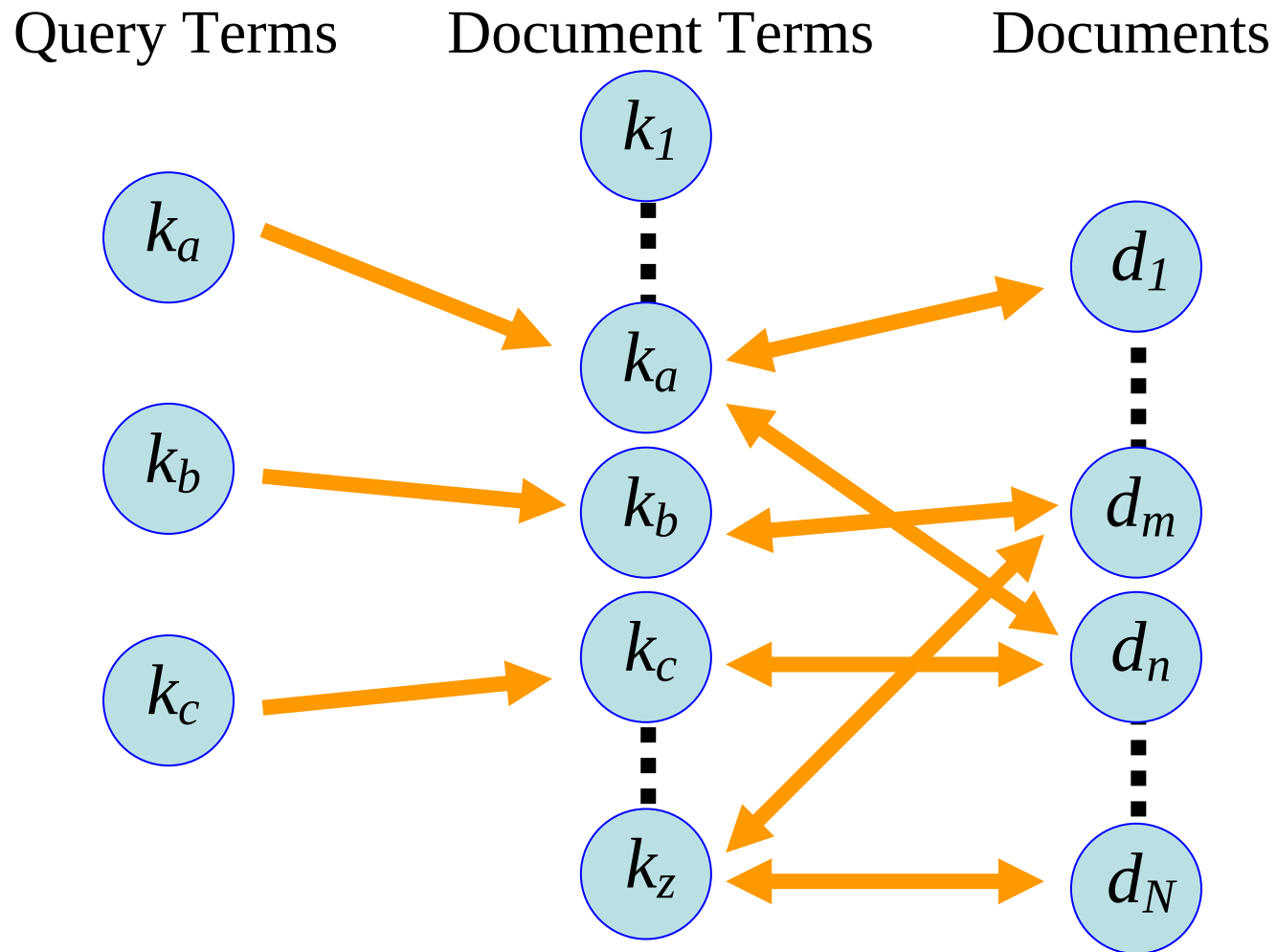


Information Retrieval、その他

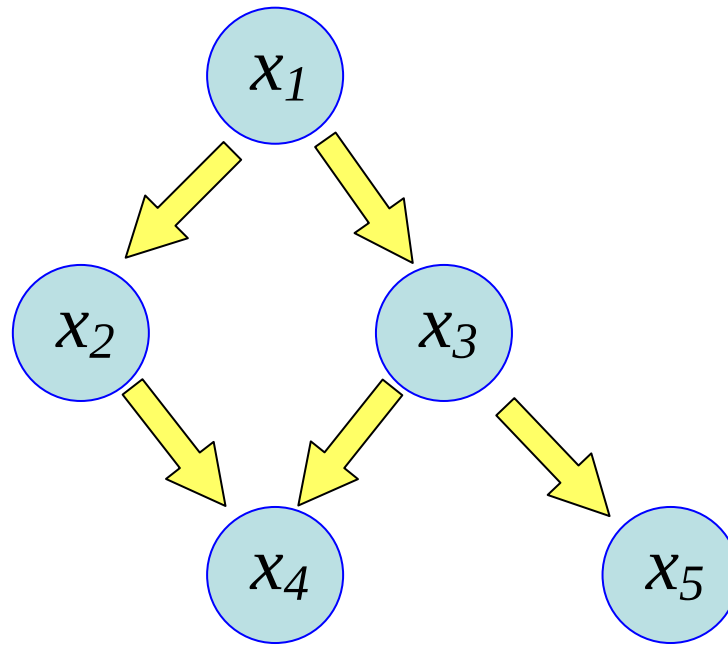
- Fuzzy Set Model
- Extended Boolean Model
- Generalized Vector Space Model
- Latent Semantic Indexing Model
- Neural Network Model
- Bayesian Networks
- Inference Network Model
- Belief Network Model



Neural Network



Bayesian Network



$$\begin{aligned} &P(x_1, x_2, x_3, x_4, x_5) \\ &= P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(x_5 \mid x_3) \end{aligned}$$



IRのまとめ

- 各文書を特定のタームでインデックスする。
- 質問文も同様にインデックスする。
- ブーリアンモデル、ベクトルモデル、および確率モデルが古典的な情報検索の手法。
- *tf-idf* 法。
- Bayesの定理。



文献

- 財団法人 医療情報システム開発センター (MEDIS-DC)
<http://www.medis.or.jp>
- UMLS & MeSH
<http://www.nlm.nih.gov/database/database.html>
- Coletti MH, Bleich HL, Medical Subject Headings Used to Search the Biomedical Literature. JAMIA. 2001;8(4):317-323.
- SNOMED® Clinical Terms Guide Technical Implementation Guide July 2002 Release Version 8 (2002-07-26)
- UMLS Knowledge Sources 14th Edition-January Release 2003AA Documentation
- Modern Information Retrieval, Ricardo Baeza-Yates, Ribeiro-Neto, Addison-Wesley 1999

