

特集：公衆衛生と臨床研究への CDISC 標準導入

<総説>

用語集の整備と情報モデルの開発, そして標準化

小林慎治

京都大学 EHR 共同研究講座

Development of terminology and an information model for standards

Shinji Kobayashi

Department of EHR, Kyoto University

抄録

インターネットで大規模に収集されたデータを用いて、機械学習をベースとした人工知能技術が発展し、創造的価値が生み出されつつある。医療分野においても国家レベルで診療データを収集し、保健行政や臨床研究、創薬に役立てていくとする試みが世界各国で進められてきた。

日本においても診療データの二次利用に向けて法制度が進められると同時に、普及が進んだ電子カルテから診療情報を大規模に収集する事業が国家的プロジェクトとして進められている。しかしながら、電子カルテのデータを収集すること、そしてそのデータを活用していくことは実際にはそう簡単ではない。

関心のあるデータを特定の用語を使って高速に検索してデータを収集することは電子化のメリットの一つであるが、診療データに記録されている用語が標準化されておらず、効率よくデータを検索できないなどの問題がある。そのため医学用語集の整備と標準化が進められており、複数の用語でも効率よく対応できるシソーラスやタキソノミーの開発や、知識工学を応用したオントロジーの開発も進められている。さらに、概念の構造と用語の関係を明示した情報モデルの開発により、質の高い情報基盤の整備へと発展しつつある。

本稿では臨床情報を機械的に処理する際の問題点とその解決法として、用語と情報モデルについて解説し、その標準化について概説する。

キーワード：用語, 概念, 用語集, 情報モデル, 標準

Abstract

Recently, so called Big Data has been remarked upon as a resources of machine learning-based artificial intelligence. Moreover, healthcare data have been integrated for clinical use and secondary use, including clinical research, public health, and so on.

However, the lack of interoperability has blocked the reuse of clinical data. Standardization of terminology will improve the efficiency of data collection, data quality, and reuse of data.

連絡先：小林慎治

〒606-0805 京都市左京区下鴨森本町15 生産開発科学研究所

Research Institute for Production Development, 15 Shimogamo-morimoto-cho, Sakyo-ku, Kyoto, Japan.

Tel: 075-703-8015

Fax: 075-703-8012

E-mail: skoba@moss.gr.jp

[令和元年5月31日受理]

In this article, we describe the problems that exist in processing clinical data with machines using terminology and concept models.

keywords: term, concept, terminology, concept model, standard

(accepted for publication, 31st May 2019)

I. はじめに

GoogleやAmazonなどに代表される国際的IT企業が、世界中に広がるインターネットを利用して大規模にデータを収集し、その「ビッグデータ」をビジネスに活用し成功を収めているがデータを処理し活用していく手法として旧来の統計的手法だけではなく、深層学習 (Deep learning) に代表される機械学習を用いていることも話題となっている。大規模データと深層学習を応用した人工知能 (AI; Artificial Intelligence) の研究が急速に進み、将棋や囲碁などで人間を凌駕する能力を示したことにより、その脅威と恩恵について議論も起きている。一方で情報漏洩などの問題も頻発しており、データの収集と活用に社会的関心が高まっている。

このような背景の中、医療分野でも診療情報を収集し、「ビッグデータ」として活用することに期待が寄せられている。国家レベルで国民の医療データを収集し、保健医療行政に役立てていこうという試みは北欧にて1960年代より先行して行われてきたが、2000年代以降は医療の電子化が進展するに伴い、各国で同様の試みが行われている。「医療ビッグデータ」を研究分野に活用したり、AI導入による医療の高度化についての研究も進められつつある。

本邦においては、2007年より医療保険請求がオンライン化されており、保険情報に関しては大規模データ収集の整備が進められてきた。NDB (National Database) は厚生労働省が管理しているレセプト、特定健診のデータベースであり、医療費の適正化のために活用されている他、研究者に対しては厳格なルールのもとで研究への利用を認めている。

その他、外科学会が中心となって手術記録を収集しているNCD (National Clinical Database) 事業や独立行政法人医薬品医療機器総合機構 (PMDA) が中心となって薬剤投与ならびに薬剤に関連する有害事象について情報収集事業 (MID-NET) を行っている。

筆者らもAMED公募事業として「全国共同利用型国際標準化健康・医療情報の収集及び利活用に関する研究」(通称: 千年カルテ) に取り組んできた。すでに30病院を接続しており、総病床数として約5万床規模の臨床情報データベースが構築される見込みである。

電子カルテも普及期に入り、診療情報の多くが電子的に保存されるようになってきた。しかしながら、インターネットに代表される情報通信技術が進歩した現在においても医療データを収集し、活用していくためには多くの課題が残されている。

本稿では、診療情報収集と解析における問題点と、それを解決するための手法としての用語と情報モデルの標準化を中心に解説する。

II. 用語と概念

電子カルテには日常診療で発生した情報が記録される。その内容は医師や看護師による患者の病状についての観察や考察、実施した治療内容や検査結果などであり、医学用語を交えて構成される。医学用語は適切に使うことで、情報を簡潔にかつ正確に伝えることができる。たとえば、「胃前庭部がんに対する幽門側胃切除術」と言えば、執刀する外科医だけではなく、麻酔科医、看護師などにも実施される手術の内容や形式、準備する物品に至るまで情報を共有することができる。

用語は、特定の抽象化された概念を指し示すものである。例えば、胃がんには発生した部位や組織型などの差があり、患者ごとに違った特性があるが、疾患としての「胃がん」はそれらの特徴を総合的に捉えた「概念」である。

医学の進歩、医療技術に発展に伴って医学用語は増えつつあり、指す意味もまた時代に合わせて変化している。日本語や英語、場合によってはドイツ語やラテン語といった多国語が混在した医学用語が記載されていることも診療記録の特徴の一つであり、地域や学閥によって方言のような用語の差異も見られる。用語の違いは臨床現場で混乱を招くことがあるため、日本医学会、日本医師会は「医学用語集」を編纂して、用語の標準化、統一化を進めている。病名や薬品、検査項目などの用語を辞書的にまとめたものを用語集 (Terminology) とよび、近年ではICD10 (International Code of Disease version 10) [1] などのようにコンピュータで処理のために、各用語に対してそれぞれ固有のコード (アルファベットや数字の組み合わせ) が付与されるようになってきた。

日本では保険請求のために病名や処置、検査、薬剤について用語集の整備が進められてきた。保険病名はICD10をベースとして、検査についても日本臨床検査医学会が臨床検査分類コードを整備していて、現在は第10版 (JLAC10) が公開されている。その他、薬剤のHOTコードなど、様々な用語集がコンピュータ処理可能な形式で構築されており、マスターとも呼ばれる (表1)。

前述の通り、保険病名については保険請求のために電子的にデータが収集されており大規模データの収集が可能である。しかし、電子カルテから患者が実際に罹患している病名をすべて抽出することはそう簡単ではない。

表 1 国内で使用される主な医学用語集

名称	詳細
ICD10	WHO が策定した各国の死因統計に使われる病名集。
標準病名マスター	保険請求の時に使用される病名集。ICD10 準拠。
JLAC10	日本臨床検査医学会による標準検査項目集。
薬価基準収載用コード	保険請求に使われる厚生省が制定した医薬品集
医薬品 HOT コードマスター	JAN コードや薬価基準収載用コードを対応付けた薬品コード集

まず、電子カルテ内で「レセプト請求」に用いられる項目名と、「診療情報提供書」で用いられる項目名では同じ「病名」という概念を指しているが項目名称が必ずしも統一されていないことがある（図 1）このような場合を配慮すると、該当しそうな複数の用語を使って検索を行う必要がある。

類義語に対しても網羅的に検索可能なようなシソーラス（類義語辞書）の整備は医学用語では古くから行われてきた。機械処理ができるように類義語を系統的に収載した医学用語集として SNOMED CT [2] が国際的協力のもとで整備されつつある。

SNOMED CT は米国病理学会を中心に開発され、一般医学用語を対象とした用語集 (Terminology) としてイギリス、デンマークの協力を経て発展してきた。2019 年 5 月現在 38 カ国が開発団体に加盟しており、医学に関連した約 35 万用語を収載している。英語をベースとして、用語体系が開発されているが、フランス語やスペイン語などへの翻訳も進んでおり、同じ概念であれば言語が変わっても同一のコードが割り振られるため、共通言語基盤としても利用されている。しかし、日本は SNOMED CT の国際的開発団体である非営利活動法人 SNOMED

International には加盟しておらず日本語訳も公式には進んでいない。加盟国は無償で SNOMED CT を使用できるが、日本で利用するためには、目的と組織の規模に応じた有償ライセンスを購入する必要がある。日本語訳も不十分な SNOMED CT を日本の医療現場で利用するメリットは現在のところあまりない。しかし、臨床研究や創薬分野では、日本単独で国際的な共同研究が主体となっている。したがって、国際的な用語集を共通言語基盤として日本でも利用せざるをえなくなる可能性がある。

III. 用語と概念、関連性

これまで用語と概念、類義語について説明してきた。用語はそれぞれ独立した概念を示すものであるが、概念の間には多様な関係性がある。

たとえば、タキソノミーは生物学的系統分類のような階層的な概念関係を示している。階層的な概念関係は一般的な病名集でも取り入れられており、ICD10 は疾患コードの最初のアルファベット 1 文字が疾患カテゴリーを意味している [3]。A, B で始まる疾患コードは感染症に分類されており、C は新生物（腫瘍）、D は血液疾患、E は代謝性疾患などと系統的に分類されている。ICD10 で「肺がん」は「気管支および肺の悪性新生物」として C34 というコードが割り振られており、部位によって C34.0-9 までのコードが割り振られている。このようなコードが割り振られていれば、病名データベースから「C34」という ICD10 コードを検索して「肺がん」の患者を抽出することは容易である。しかし、コードが割り振られておらず病名がそのまま登録された病名データベースから「肺がん」をまとめて検索するためには、「肺ガン」「肺癌」といった漢字、かな表記の違いだけではなく、「肺小細胞がん」「非小細胞肺がん」「肺扁平上皮がん」などの組織型の違いも網羅する必要がある。「がん」患者全般について検索をかけるために、個別のがん（口腔がん、食道がん、胃がん、膵臓がん、肝臓がん、腎臓がん、etc）のすべてのコードを網羅的に検索してその合計を計算することは煩雑な作業となるが、ICD10 の階層的な概念構造を利用して、「C」で始まるコードを検索することは容易である。病名集の他、薬剤集には抗生物質や抗がん剤などの種別で系統的に分類されてい

レセプト請求 病名 保険記号・番号 公費別 … …	診療情報提供書 傷病名 紹介目的 治療経過 … …	病理診断報告書 診断名 検体 提出日 … …
病理診断報告書 Diagnosis Specimen … …	画像診断報告書 総合所見 方法 緊急 … …	退院時要約 入院契機病名 最終診断病名 DPC 病名 … …

図 1 臨床現場で使われる「病名」とその類義語

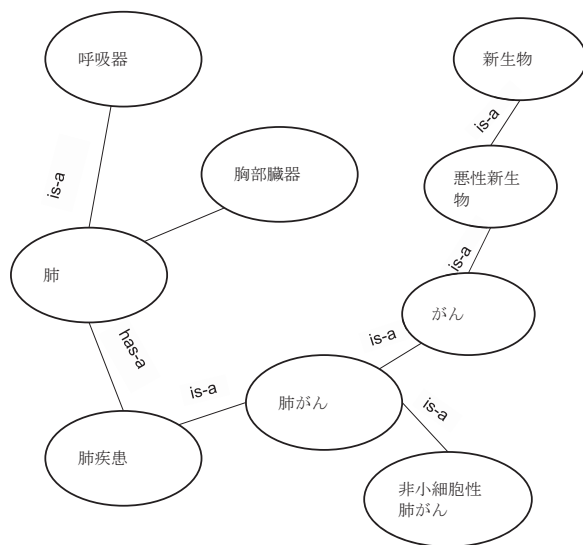


図2 肺がんのオントロジー図

るものもあるなど、タキソノミー的な階層構造は医学用語集で広く取り入れられている。

このようにタキソノミーは対象となる用語集やその目的のために階層構造で用語が示す概念を整理するが、実際には用語が示す概念の関係は階層の関係だけではなく様々な関係性を持つ。「肺がん」は臓器別では「肺疾患」に分類されるが、「がん」であるために「悪性新生物」にも分類される。図2に示すように、「非小細胞肺がん」は「肺がん」の一つの組織型であり、「肺がん」は「がん」の中で肺という臓器に発生したものである。「がん」は悪性新生物のなかで上皮細胞に発生したものである。

このように概念が示すものとその関連性について明示的に整理したものを「オントロジー」と呼び、機械処理可能な知識表現として近年注目されている。日本でも医療分野のオントロジー開発は東大医療情報学大江教授により主導的に進められてきた。

オントロジーは人工知能技術への応用が注目されており、全米クイズ選手権で優勝したり、難治性の白血病の診断を支援するなど話題となったIBM Watsonの質疑応答エンジンでも用いられている [4,5]。

IV. 用語と形式、情報モデル

用語とそれが示す概念の統一と標準化、そして機械処理について概説してきた。標準化した用語を用いることは大規模データ収集とその活用を行うためには必須であるが、それだけでは必ずしも十分ではない。図3に示すように意味的には同じ概念で、同じように用語を使用しても表現形式が異なることがある。このような形式の違いは臨床で人間が扱う範囲では問題となることは少ないが、機械的に大規模データを収集するためには障害となりうる。

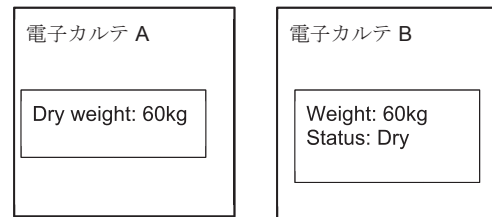


図3 Dry weightに関するデータ表現の差、下記より日本語訳し改変[6]

OBX [1 [NM [^Systolic blood pressure] |136 [mm[Hg] |90-120 [H] ||] F
OBX [2 [NM [^Diastolic blood pressure] |86 [mm[Hg] |60-80 [H] ||] F

収縮期血圧と拡張期血圧を分けた記録

OBX [1 [ST [^Blood pressure] |136/86mmHg | || ||] F

収縮期血圧と拡張期血圧をまとめた記録

図4 同じHL7 Ver 2.5血圧データ（一部抜粋）

腎不全の患者は血中の老廃物や余剰の水分を尿として排泄することができないため、代用として血液透析が行われるがその場合の目標体重として余分な水分がない状態の体重をdry weightと一般的に呼ばれている。単に用語で“Dry weight”とするのか、体重 (weight) を測定した状態 (status) をdryとして表現するのかについては、現状ではそれぞれの電子カルテや透析部門システム次第である。構造の違うデータを同じように扱うことは困難であり、データの集積には大きな障害となる。

したがって、データを収集して活用するためには用語を統一するだけではなく、その取り扱う形式もまた統一する必要がある。医療情報を機械的に伝達する形式は1970年代より開発が進められてきた。代表する標準規格としてHL7 [7] などがある。しかし、これらの規格を持っても、データ形式を共通化してデータ収集を行うことは容易ではない。血圧データを同じHL7 Version 2形式のメッセージで表現する場合に、異なる表現を取ることがあり得る (図4)。

「血圧」を表すデータには「収縮期血圧」と、「拡張期血圧」とがあり、通常はその2つの値が1つのセットとして扱われる。さらに、専門的な検査を行う場合には測定した「部位」や「機器」などの「測定方法」、測定時の「運動条件」や「体位」といった「状態」も記録される。このように「血圧」という概念に対応するデータは複数の項目と構造でなりたっている (図5)。

このように特定の概念を表す情報について構造を明らかに定義したものを「情報モデル」という。共通の情報モデルを利用することで、データ表現の正確さを向上させることができ、異なるフォーマットでも情報を相互に運用することができるようになる。医療分野でのデータモデルの開発は1990年代よりオブジェクト指向技術を取り入れる形で発展してきた。ヨーロッパとオーストラリ

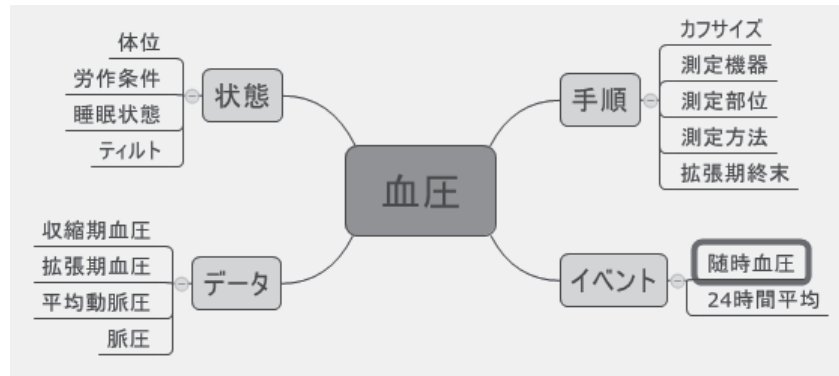


図5 「血圧」を表す情報モデルのマインドマップ図

アを中心に医療分野で共通の情報モデルを設計し、共有するopenEHRプロジェクト [8] が発足し、その成果がISO 13606標準の根幹となるarchetypeである [9]。現在では500を超える情報モデルがopenEHRプロジェクトのClinical Knowledge Manager (<https://www.openehr.org/ckm/>) に登録されており、クリエイティブ・コモンズライセンス (CC-BY-SA) のもとで自由に利用することができる。図5に示した血圧のほか、日常臨床で扱われるデータについて約450の情報モデルが公開されており、その多くは日本でも利用可能である。

情報モデルを設計して、相互運用性を高めようとする試みは、openEHRだけではなく、HL7 CDAやHL7 FHIR [10]、CDISC [11] といった他の医療標準規格開発団体で行われておりそれぞれの標準規格に反映されている。CDASHはCDISC標準で臨床調査票 (Clinical Research Form) を構築するための情報モデルについて取り決めた規格である。

2010年には医療分野での標準化団体 (CDISCを含む) とそれに関連する開発ベンダーが一同に介して、情報モデルを共有しようとする団体CIMI (Clinical Information Modeling Initiative) [12] が発足した。CIMIは現在HL7の分科会として、共通する情報モデルの開発を行っており、openEHRによって開発されたarchetype技術を利用して

V. 日本での情報モデル普及への課題

残念ながら日本では情報モデルそのものへの理解が進んでいない。ICD10を利用して病名を横断的に検索することは可能であるが、重症度や病期、診断根拠に関する用語とそれを格納する情報モデルが整理されていないため、これらを指定して検索することは非常に難しい。治験データの国際標準であり医薬品医療機器総合機構が治験データの電子的提出の標準としたCDISCにおけるCDASH規格は情報モデルの理解を前提としている。また、2010年に開発が始められて以後、爆発的に普及が進んでいるHL7 FHIRもまたResourceやProfileと呼ばれる情報モデルを中心に組み立てられている。したがって情

報モデルについての理解がなければ、これらの標準や技術を理解することはできないし、正當に評価も行うことすら困難である。

臨床情報モデルについての議論は、医療についての知識と情報学的素養の両方が必要となる。標準として利用可能なモデルを設計するには経験も必要であり時間もかかる一方で、実際の運用が目に見えづらいために敬遠されがちである。日本医療情報学会でもモデリングについての分科会を立ち上げるなどの取り組みが行われており、我々もopenEHRの普及活動の一環として臨床情報モデルについての講習会を開いてきた。標準的な臨床情報モデルの開発は国際的な協力のもとで進んでいるが、日本でそれを活用するためには情報モデルについての知識をさらに広めていく必要がある。

VI. まとめ

医療情報の標準規格を構成する用語体系と情報モデルについて概説した。日本ではICD10に準拠した病名集が普及するなど用語体系の取り込みは積極的に行われているが、国際的な用語体系や情報モデル開発にどのように参加していくかについてはまだ十分に検討できるほど情報モデルについての知識が普及していない。

本稿が日本において情報モデルについての知識が普及するきっかけとなることを切に願ってやまない。

引用文献

- [1] WHO. ICD-10 online versions. <https://www.who.int/classifications/icd/icdonlineversions/en/> (accessed 2019-05-15)
 - [2] SNOMED International. SNOMED. <http://www.snomed.org/> (accessed 2019-05-14)
 - [3] 標準病名マスター作業班. ICD10対応標準病名マスター (ICD10分類). http://www.dis.h.u-tokyo.ac.jp/Scripts/ICD10Categories/default2_ICD.asp (accessed 2019-05-27)
- Hyojun Byomei Master Sagyohan. [ICD 10 taio hyojun

- byomei master (ICD10 bunrui.)] http://www.dis.h.u-to-kyo.ac.jp/Scripts/ICD10Categories/default2_ICD.asp (accessed 2019-05-27) (in Japanese)
- [4] IBM Community. NLP-driven Ontology Modeling (Language Processing). https://www.ibm.com/developerworks/community/blogs/nlp/entry/nlp_driven_ontology_modeling8?lang=en (accessed 2019-05-20)
- [5] IBM. Clients the stars of World of Watson conference - Client Success Field Notes. <https://www.ibm.com/blogs/client-voices/client-stories-world-of-watson-conference/> (accessed 2019-05-24)
- [6] Stan Huff. Practical Modeling Issues: Representing Coded and Structured Patient Data in EHR Systems. AMIA 2014 Annual Symposium; 2014.11.15-19; Washington D.C., USA. <https://www.amia.org/sites/amia.org/files/AMIA-symposium-2014.pdf> (accessed 2019-05-24)
- [7] HL7 International. Health Level Seven International. <http://www.hl7.org/index.cfm> (accessed 2019-05-26)
- [8] The openEHR. 2015. <http://www.openehr.org/programs/specification/latestreleases> (accessed 2019-05-26)
- [9] ISO. ISO 13606-2 :2008 Health Informatics – Electronic health record communication – Part 2: Archetype interchange specification. 2008. <https://www.iso.org/obp/ui/#iso:std:iso:13606:-2:ed-1:v1:en> (accessed 2019-05-26)
- [10] Bender D, Sartipi K. (2013, June). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. 2013 IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS); 2013.6.20-22; Porto, Portugal. IEEE. p.326-331. <http://dx.doi.org/10.1109/CBMS.2013.6627810> (accessed 2019-05-26)
- [11] Daniel C, Sinaci A, Ouagne D, Sadou E, Declerck G, Kalra D, et al. Standard-based EHR-enabled applications for clinical research and patient safety: CDISC - IHE QRPH - EHR4CR & SALUS collaboration. AMIA Jt Summits Transl Sci Proc. 2014;2014:19-25.
- [12] HL7 International. Clinical Information Modeling Initiative. <https://www.hl7.org/Special/Committees/cimi/index.cfm> (accessed 2019-05-28)