様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　１１３０１
研究種目：　基盤研究(A)（一般）
研究期間：　2014～2016
課題番号：　２６２４００４４

研究課題名（和文）階層的オミックス医療情報のデータ交換形式と多軸的・多面的なオントロジーの整備

研究課題名（英文）Development of both the data exchange format of hierarchical omics data and the multi-axis and polyphenic ontology

研究代表者

中谷　純（NAKAYA, Jun）

東北大学・医学系研究科・大学院非常勤講師

研究者番号：９０４２０４６３

交付決定額（研究期間全体）：（直接経費）　　33,700,000円

研究成果の概要（和文）：本研究課題では、階層化されたオミックス医療情報のデータ交換を可能とするデータ交換フォーマットを構築した。ゲノムからフェノームに至る階層オミックス情報、医療情報、環境情報といった多次元・多因子・多階層にわたるオミックス医療情報について、そのデータ要素項目と意味関係に分離して記述することのできるデータ交換フォーマット方式の開発・標準化する案を改訂した。ISOにて国際標準化を進めた。

研究成果の概要（英文）：In this study, we formulated data exchange format of omics medical data including hierarchical omics data and clinical data. We developed novel data exchange format of multidimensional, multifactorial, multi layered data of hierarchical omics data (from genome to phenome), clinical data, and environmental data, which are separately described by data elements and semantics relationships. We are working on international standardization of our data exchange formats.

研究分野：　生命・健康・医学情報学

キーワード：　オミックス　ICD　臨床情報モデル　全ゲノムシーケンス　マークアップ言語　マルチエントリーポイント　国際標準化

１．研究開始当初の背景

次世代シーケンスの技術革新にともない、個人ゲノムがシーケンスされ、その配列の変異解析とその変異に基づく疾患リスク予測が急速に進展している。研究代表者はすでに、ISO において正式に国際標準と認められた日本発信の技術である ISO 25720 GSVML(Genomic Sequence Variation Markup Language)の策定を行ってきた。GSVML は、臨床ゲノム応用を目的としたゲノム配列の変異の標準データフォーマットであり、SNP(Single Nucleotide Polymorphism)などの遺伝子変異データと臨床データ、その付加データのデータ交換を可能とする標準データフォーマットである。申請者がリーダーとなって、ISO 8 カ国（アメリカ、イギリス、カナダ、オーストラリア、イスラエル、韓国、イタリア、日本）および HL7 CG SIG の協力を得て、策定された。この GSVML により、加速する個人ゲノムの遺伝子変異解析と疾患リスク予測を支えるデータベース基盤を構築することが可能になっている。

一方で、ゲノムのみならず、トランスクリプトーム、プロテオームなどの網羅的分子生物情報と医療情報、すなわちオミックス医療情報の蓄積も急速に進展しており、すでに医療への応用も始まりつつある。臨床・病理情報や生活習慣などの環境情報に、ゲノム、トランスクリプトーム、プロテオームなどのオミックス情報が加わったオミックス医療情報により、個人の遺伝的素因、病態、環境に特化した、新たな医療である個別化医療を可能になることが期待され、研究代表者や研究分担者はオミックス医療を提唱している。

しかしながら、ゲノムの遺伝子変異データに関しては GSVML のデータフォーマットが策定されているものの、ゲノムの遺伝子変異データのみならず、トランスクリプトーム、プロテオームなどのさまざまなオミックス情報と臨床情報、すなわちオミックス医療情報のデータ交換を可能とするデータフォーマットはいまだ整備されていない。

研究代表者は、これまで世界保健機構(WHO)による国際疾病分類(International Classification of Diseases: ICD)の策定にたずさわってきた。現在のバージョンである ICD10 は 1990 年の第 43 回世界保健総会で採択されたが、WHO においてその改訂が進行しており、2014 年に ICD11 が採択される予定である。申請者はその改訂にたずさわっている。また、これまでに、ICD11 のサブ情報モデルとして、臨床コンテンツモデルと整合性のある ICD11 オミックス情報モデルのβ版(ICD11 Omics Sub Information Model β: iCOS β)を提案してきた。iCOS βは主に網羅的分子データに関連した内容を記述するモジュール OML(Omics Markup language)、ICD11 コンテンツモデルに対応するモデリングモジュール ICD11entity モジュール、OML と ICD11 entity の間の転写関係を規定する双方向の転写モジュール Transcription Module から構成されている。この iCOS βにより、ICD11 のサブ情報モデルとして、ゲノム配列の変異データのみならず、主にトランスクリプトームのオミックス情報を記述することができる。

しかし、オミックス情報と臨床情報、すなわちオミックス医療情報の、情報モデルを、ICD11 のサブ情報モデルとしてではなく、独立した情報モデルとして構築する必要がある。ゲノムの遺伝子変異情報と臨床情報、その付加情報の情報モデルである GSVML を拡張し、ゲノムの遺伝子変異情報に始まり、トランスクリプトーム、プロテオームからフェノームに至る階層オミックス情報、医療情報、環境情報といった多次元・多因子・多階層にわたるオミックス医療情報を、そのデータ要素項目と意味関係に分離して記述することのできるデータ交換フォーマット方式を開発・標準化する必要がある。

２．研究の目的

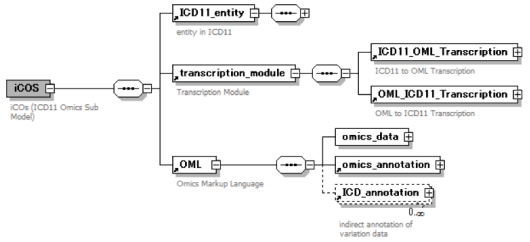
本研究課題では、階層化されたオミックス医療情報のデータ交換を可能とするデータ交換フォーマットを構築する。ゲノムからフェノームに至る階層オミックス情報、医療情報、環境情報といった多次元・多因子・多階層にわたるオミックス医療情報について、そのデータ要素項目と意味関係に分離して記述することのできるデータ交換フォーマット方式を開発・標準化する。データ要素項目のエントリーポイントは、任意に設定可能なマルチエントリーポイント方式とする。要素データと意味関係を分けた情報モデルとし、意味関係としては、要素データの階層関係を表現可能とし、多軸的・多面的なオントロジーを整備する。意味関係については、その関係性を確率表現とともに記述した上で、意味関係のみを抽出して独立に表現可能とする機能を開発収載する。

３．研究の方法

(1) 要素データの情報モデルの予備的検討

研究代表者が提案した、ICD11 オミックス情報モデルのβ版 (ICD11 Omics Sub Information Model β: iCOS β)は、OML(Omics Markup language)、ICD11 Entity、Transcription Module から構成され、OML モジュールは、主にオミックス情報に関連した内容を記述するモジュール、ICD11entity モジュールは ICD11 コンテンツモデルに対応するモデリングモジュール、Transcription Module は、OML と ICD11 entity の間の転写関係を規定する双方向の転写モジュールである。

この iCOS βにより、ICD11 のサブ情報モデルとして、ゲノム配列の変異情報と臨床情報のみならず、主にトランスクリプトームのオミックス情報と臨床情報を記述すること



ができる。初年度はまず、この iCOS βに基づき、要素データの情報モデルの予備的検討を行う。トランスクリプトームデータの情報モデルについては、FGED (Functional Genomics Data) Society において策定されている MIAME(Minimum Information About a Microarray Experiment) 形式との整合性について検討する。トランスクリプトーム以外のオミックス情報と医療情報の情報モデルについても検討する。

(2) 要素データの階層的な意味関係の予備的検討

オミックス情報および臨床情報を構成する要素データについて、包括的な意味関係は現在存在しないが、GSVML、NCK (Normalized Clinical Knowledge)にくわえて、National Center for Biomedical Ontology で収集されているオントロジーを参考にして、これらで用いられている用語を基に、用語間の関連性を解析する。関連性検討に際して、それぞれの用語の意味を多方面から分解し多軸的な意味情報（メタデータ）として記述し、共通の軸を抽出する。抽出した共通軸を中心として、非共通軸を重複のない形で付加し、すべてのメタデータを記述しうる構造を構築する。これにより予備的検討を行う。

(3) 要素データの情報モデルの構築

前年度の予備的検討を踏まえて、要素データとなる、ゲノム、トランスクリプトーム、プロテオームからフェノームに至るさまざまなオミックス情報と臨床情報、その付加情報のデータ交換を可能とする情報モデルを構築する。XML 形式の情報モデルとする。データ要素項目のエントリーポイントは、任意に設定可能なマルチエントリーポイント方式とする。ゲノムの遺伝子変異情報、トランスクリプトーム情報、プロテオーム情報、シグナローム情報、メタボローム情報、オルガノーム情報、パスウェイ情報、環境因子として生活習慣情報、臨床情報、フェノーム情報を新たに定義する。

(4) ICD11 コンテンツモデルとの整合性検証

ICD11 コンテンツモデルとの整合性検証は、要素データの情報モデルで定義される項目の意味的整合性とオミックス情報記述性との整合性の観点から行う。また、臨床オミックス情報モデル作成に必要な構成要素、他の標準臨床医学情報モデルとの連携性、概念対応の違いなど、臨床オミックス情報モデルを

フェノーム	XML形式
臨床情報	HL7, HL7 CDA (Clinical Document Architecture) R2, DICOM形式
環境因子 生活習慣情報	XML形式
オルガノーム	iCOS β Organome Module
メタボローム	iCOS β Metabolome Module
パスウェイ	SBML (Systems Biology Markup Language)形式
シグナローム	iCOS β Signalome Module
プロテオーム	iCOS β Proteome Module
トランスクリプトーム	iCOS β Transcription Module
ゲノム(遺伝子変異)	ISO 25720 GSVML(Genomic Sequence Variation Markup Language)

ICD11 の中で実効させるために必要な情報を明らかにする。

(5) 要素データの階層的な意味関係の構築

前年度の予備的検討を踏まえて、遺伝子変異、トランスクリプトーム、プロテオームからフェノームに至るオミックス情報の階層関係を明らかにし、多軸的・多面的なオントロジーを整備する。Protege や LexWiki というオントロジー構築ソフトウェアを用いて行う。これらの要素データの階層的な意味関係を構築するにあたり、パスウェイ情報を用いる。さらに、意味関係の確率表現と統計集計機能の開発収載について検討する。

(6) WHO-TAG-HIM 会議における情報交換

WHO-TAG-HIM (Topic Advisory Group for Health Informatics and Modeling)は、WHO の諮問委員会で、ICD の電子化における実務上の決定権をもっている。ネットワークミーティングが随時ジュネーブで行われるので、そのミーティングに出席し、ICD のモデル化作業に関する情報収集と意見交換を行う。また、随時、WHO-FIC, 内科 TAG などへも参加し、情報交換を行う。

４．研究成果

本研究課題では、階層化されたオミックス医療情報のデータ交換を可能とするデータ交換フォーマットを構築した。ゲノムからフェノームに至る階層オミックス情報、医療情報、環境情報といった多次元･多因子･多階層にわたるオミックス医療情報について、そのデータ要素項目と意味関係に分離して記述することのできるデータ交換フォーマット方式の開発･標準化する案を改訂した。

データ要素項目のエントリーポイントは、任意に設定可能なマルチエントリーポイント方式を開発した。要素データと意味関係を分けた情報モデルとした。

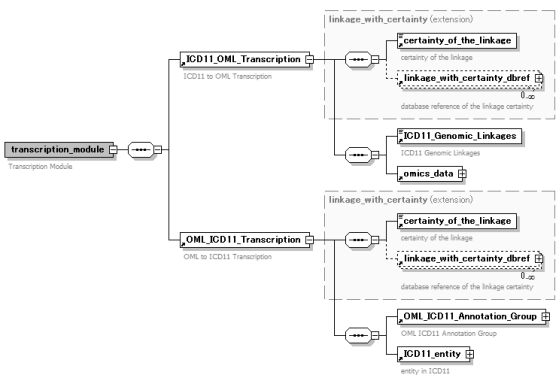
要素データとしては、ゲノムの遺伝子変異情報については、ISO 25720 GSVML(Genomic Sequence Variation Markup Language)にしたがうものとした。トランスクリプトーム情報については、iCOS β Transcirptome Module を基にし、FGED (Functional Genomics Data) Society において策定されている MIAME(Minimum Information About a Microarray Experiment) 形式との整合性をもたせるように検討した。プロテ

オーム情報、シグナローム情報、メタボローム情報、オルガノーム情報については iCOS β Proteome, Metabolome, Organome Module を基に検討した。パスウェイ情報は SBML(Systems Biology Markup Language)形式との整合性をもたせるように検討した。環境因子として生活習慣情報はXML形式で、新たな定義を検討した。臨床情報は HL7, HL7 CDA (Clinical Document Architecture) R2, DICOM 形式にしたがう。フェノーム情報も XML 形式で新たな定義を検討した。

意味関係としては、要素データの階層関係を表現可能とし、遺伝子変異、トランスクリプトーム、プロテオームからフェノームに至るオミックス情報の階層関係を考慮した、多軸的･多面的なオントロジー案を改訂した。

これらの要素データの階層的な意味関係を構築するにあたり、パスウェイ情報を用いることを検討した。**パスウェイ情報により遺伝子変異、トランスクリプトームからプロテオーム、シグナロームデータ、メタボロームデータは階層的な意味関係をもたせることが可能となる。**

意味関係については、その関係性を確率表現とともに記述した上で、意味関係のみを抽出して独立に表現可能とする機能を開発収載した。**Transcription module は、ICD11_OML_Transcription および OML_ICD11_Transcription のモジュールから構成することができる。それぞれのモジュールは、その投射方向により機能分割することができる。ICD11_OML_Transcription は、ICD11 から OML への投射を行う場合に使用するモジュールであり、OML_ICD11_Transcription は、OML から ICD11 への投射を行う場合に使用するモジュールである。ICD11_OML_Transcription モジュールは、ICD11GenomicLinkage という ICD11 Entity の項目とオミックス情報との間の関係を一方向に規定する。関係の重みづけは、確信度という形で用意してあるが、特殊な定義を用意して、参照する形で使用することもできる。OML_ICD11_Transcription モジュールは、OML ICD11 Annotation Group という OML の項目と ICD11 Entity との間の関係を一方向に規定することを検討した。こちらも、関係の重みづけは、確信度という形で用意してあるが、特殊な定義を用意して、参照する形で使用することもできる。**



研究代表者がたずさわっている世界保健機構(WHO)による国際疾病分類(International Classification of Diseases: ICD)の ICD11 について、そのサブ情報モデルとして、臨床コンテンツモデルと整合性のある ICD11 オミックス情報モデルのβ版(ICD11 Omics Sub Information Model β: iCOS β)を提案してきた。iCOS の開発のなかで逆転写情報モデル、OML、WGMLといった多軸構造対応モデルの施策を行った。データフォーマットは、GSVML の発展型として OML(Omics Markup Language)、WGML(Whole Genome Markup Language)を開発し、これらの国際標準化活動を開始した。OML、WGML については、ISO TC215 WG2 において、GSVML の付加構造として日本の JISC から国際標準化に向けた提案を行い、OML は国際標準(IS)、WGML は Technical Report(TR)となるように進めた。また、このプロジェクトと連動し、WHO においては、ICD11 のジェノミクス関連サブモデルである iCOS が WHO-FIC 会議で 2 年連続で取り上げられ、プレゼンテーションアワードを受賞した。iCOS は、ICD11 のサブプロジェクトとして独立し、WHO-FIC ITC において、主要課題として検討されている。

また、iCOS について、臨床研究の国際的データ交換基準の事実上の標準である CDISC、HL7 と連携した、ICD11 コンテンツモデルとの整合性検証を行った。OML (Omics Markup Language)、WGML (Whole Genome Markup Language)の多軸構造対応モデルの国際標準化活動は継続し、OML は国際標準(IS)、WGML は Techinical Report (TR)としての標準化を継続して進めた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 16 件）

① Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F Lauledеrkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MW, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JO, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. The Human Phenotype Ontology in 2017. Nucleic Acids Res. **査読有**、2017. 45(D1) :D865-D876. 10.1093/nar/gkw1039

② Mizuno S, Ogishima S, Nishigori H, Jamieson DG, Verspoor K,Tanaka H, Yaegashi N, Nakaya J.The Pre-Eclampsia Ontology: A Disease Ontology Representing the Domain Knowledge Specific to Pre-Eclamsia.PLos One. **査読有**、2016. 11(10):e0162828. 10.1371/journal.pone.0162828

③ Nakaya J, Kimura M, Ogishima S, shabo A, Kim IK, Parisot C, de Faria Leao B. Future Direction of IMIA Standardization. IMIA Yearbook of Medical Informatics. **査読有**. 2014. 9:105-109. 10.15265/IY-2014-0010

④ Michio Kimura, Jun Nakaya, Hiroshi Watanabe, Toshiro Shimizu, Kazuyuki Nakayasu.、A Survey Aimed at General citizens of the US and Japan about Their Attitudes toward ElectronicMedical Data Handling. International Journal of Environmental Research and Public Health . **査読有**、2014. 11:4572－4588. 10.3390/ijerph110504572

⑤ **中谷　純、医学情報学分野の視点から見た未来型医療と東北メディカル・メガバンク事業、日本歯科医師会雑誌、査読有、66 巻、2014、ページ数（NA）**

⑥ Jun Nakaya, Takeshi Imai, Michiyo Kamata, Kaei Hiroi, Hiroshi Tanaka.、Clinical Omics sub information model for ICD11(iCOs). WHO-FAMILY OF INTERNATIONAL CLASSIFICATIONS NETWORK ANNUAL MEETING. **査読有**. **巻**（NA）、2014. C318

⑦ Jun Nakaya, Takeshi Imai, Michiyo Kamata, Kaei Hiroi, Hiroshi Tanaka. Clinical Omics sub information model for ICD11(iCOs). WHO-FAMILY OF INTERNATIONAL CLASSIFICATIONS NETWORK ANNUAL MEETING. **査読有、巻**（NA）. 2014. C319

〔学会発表〕（計 9 件）

① Jun Nakaya, Takeshi Imai, Kaei Hiroi, Mika Watari, Hiroshi Tanaka、Progress Around Clinical Omics sub Information model for ICD(iCOs). WHO-FAMILY OF INTERNATIONAL CLASSIFICATIONS NETWORK ANNUAL MEETING. 2015 **年** 10 **月** 17 **日**～10 **月** 23 **日**、Manchester(UK)

② **中谷　純、招待講演「未来型医療、地域医療連携における医療介護福祉情報の変化」、第** 41 **回日本診療情報管理学会学術大会、**2015 **年** 9 **月** 17 **日**～18 **日、岡山コンベンションセンター（岡山市）**

③ **中谷　純、大会長講演「羅針盤：医療情報学の位置」、第** 19 **回医療情報学会春季学術大会、**2015 **年** 6 **月** 12 **日**～13 **日、仙台国際センター（仙台市）**

④ Jun Nakaya, Takeshi Imai, Michiyo Kamata, Kaei Hiroi, Hiroshi Tanaka.、Clinical Omics sub information model for ICD11(iCOs). WHO-FAMILY OF INTERNATIONAL CLASSIFICATIONS NETWORK ANNUAL MEETING. 2014 **年** 10 **月** 11 **日**～17 **日、**Balcelona(Spain)

⑤ **井上　隆輔, 中山 雅晴, 中谷　純、病院情報システムにおける薬剤近畿情報の取り扱い、第** 18 **回医療情報学会春季学術大会、**2014 **年** 6 **月** 6 **日**～7 **日、岡山コンベンションセンター（岡山市）**

〔図書〕（計 1 **件）**

① **清水佳奈、山本奈津子、川嶋実苗、片山俊明、荻島創一、羊土社、改正個人情報保護法でゲノム研究はどう変わるか？-個人識別符号・要配慮情報としてのゲノムデータ、**2017、35(4):660(600-605)

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
中谷　純（NAKAYA, Jun）
東北大学・医学系研究科・大学院非常勤講師
研究者番号：　９０４２０４６３

(2)研究分担者
永家　聖（NAGAIE, Satoshi）
東北大学・東北メディカル・メガバンク機構・助教
研究者番号：００７２６４６６

荻島　創一（OGISHIMA, Soichi）
東北大学・東北メディカル・メガバンク機構・准教授
研究者番号：　４０４４７４９６

今井　健（IMAI, Takeshi）
東京大学・大学院医学系研究科（医学部）・准教授
研究者番号：　９０４０１０７５