

特集：Evidence Based Public Health: ICT/AI を活用したこれからの保健医療

<総説>

Real World Data を活用する観察研究データベースの考察

木村映善

国立保健医療科学院統括研究官

Considerations of the observational research
database utilizing Real World Data

Eizen Kimura

Research Managing Director, National Institute of Public Health

抄録

悉皆的に収集されたReal World Data (RWD) を用いた観察研究からエビデンスを導出できるような取り組みが求められている。データベース設計に関する課題として、標準情報モデルへの統合、統制用語へのマッピング、各施設の測定結果などの組織間較正、患者個体の識別・追跡性の確保およびデータソースを巡るバイアスについて提示した。これらの課題を踏まえて、RWDからのETL手法の標準化、標準化されたデータベース構造の定義、コホートを適切に定義する方法論、RWDデータの素性の特長と品質改善、そしてRWDを収集する対象と時系列上の網羅性の確保が、医療ビッグデータ時代に課せられた優先度の高い研究テーマであることを提示した。標準化されたデータベースとコホートを適切に定義する方法論について焦点をあて、先行研究としてeMERGEプロジェクトにおけるPheKBというフェノタイピング手法を公開するリポジトリと、OHDSIプロジェクトにおけるOMOP CDMによるデータベースの構造の標準化と標準統制用語集の取り組みについて紹介した。今後の我が国における取り組みの一案として、国内事情を踏まえたDWHの取り組みと国内外のCDMと統制用語集のマッピングの取り組みとの間に責任分界点を設定したプロジェクトを立ち上げることを提案する。

キーワード：データヘルス，リアルワールドデータ，データウェアハウス，フェノタイピング，二次利用

Abstract

Data health initiatives require that we are able to derive evidence from an observational database using Real World Data (RWD) collected exclusively. As the issues related to database design, RWD requires integration into the standard information model and controlled terminologies, inter-organization calibration on such as measurement results, identification and tracking of patient individuals, and bias problems related to data sources. Regarding these issues, standardization of ETL method from RWD, definition of standardized database structure, developing proper methodology to define cohorts, identification of the sources of RWD data and development of quality measures, and ensuring coverage on all cohorts and time series are the highest priority research topics of the medical big data era. We focused on methodologies for defining a

連絡先：木村映善

〒351-0197 埼玉県和光市南2-3-6

2-3-6, Minami, Wako, Saitama 351-0197, Japan.

Tel: 048-458-6213

E-mail: kimura.e.aa@niph.go.jp

[平成30年5月1日受理]

common data model for an observational research database and proper phenotyping for the medical informatics concerned. We have investigated the PheKB repository that publishes phenotyping methods in the eMERGE project, and the OMOP Common Data Model and standard vocabulary developed by in OHDSI project. We also propose that rather than trying to implement international CDM in a single step, we should set up two projects. One is promoting the development of a CDM consortium so that electronic medical vendors properly integrate their EHR system and DWH regarding Japanese standard terminologies. The other is that the researcher should develop the mapping tools that map between domestic and international CDM and Terminologies leveraging the CDM.

keywords: data health, Real World Data, data warehouse, phenotyping, secondary usage

(accepted for publication, 1st May 2018)

I. データヘルスの課題

2017年1月にデータヘルス改革推進本部が設置され、医療ビッグデータの分析をととしての保険者機能強化、科学的介護等を実現するための保険医療分野のデータ活用基盤の構築を一体化された戦略として取り組むことが標榜された。現状では、データヘルスとは主にレセプト情報や特定健診結果をもとに保険加入者の健康情報を分析し、エビデンスにもとづく保健事業を実施する試みである。しかし、現状の医療情報の多くは診療報酬請求情報であり、疾患概念的に制約がある。また、時系列的な網羅性にも欠けている。例えば、脳卒中・虚血性心疾患の患者の半数以上は、自覚症状がでるまで医療機関を受療しないことが確認されている。重篤化するまでに、生活習慣病のデータ収集に参加している医療機関を受診しない患者が多数存在するため、患者個体の病態の遷移を詳細に把握・追跡するには、時系列的網羅性の欠如がある。すなわち、生活習慣病に関する各指標の実態と受療行動、治療プロセスの全体像を可視化し、疫学的エビデンスを構築するための医療情報が不足している。故に、データの収集範囲を特定機能病院から診療所までの至る所に拡げ、医療情報の時系列的、疾患概念的な網羅性を高めて、疫学的エビデンスに貢献する医療情報を収集する必要がある。

II. RWD への期待と課題

医療におけるエビデンスの獲得はランダム化比較試験(RCT: Random Control Trial)をゴールドスタンダードとして取組まれてきたが、RCTの遂行は非常に多くのコストと時間を要する。また、RCTにおいて収集されたデータは収集範囲が限定的であるため、他の因果関係の解明のために再利用することが困難である。一方、近年における情報処理技術の進歩は、データ収集のコスト削減や悉皆性を確保し、医療ビッグデータの蓄積と分析を可能にすることが期待されている。すなわち、電子カルテ、診療報酬請求、処方箋などの日々の診療業務から発生するデータであるReal World Data (RWD) [1]や、公衆衛生関連行政、臨床研究の過程で収集される医療情報と

いった、諸々の観察データ群 (Observational Data) を統合し、探索的分析やエビデンスの抽出を可能にするデータベースの構築及び分析手法の開発が期待されている。

ただし、RWDをデータベースに取り込んで分析するには、いくつかの前提条件をクリアしなければならない。RWDといっても様々な種類のデータがあり、データ構造や利用しているマスタも異なる。また、医療機関によって採用している検査機器・試料も異なるため、複数組織から収集した検査結果を比較するための校正処理も必要になる。すなわち、多種多様な背景を持つRWDを有機的に連携させ、データの品質を高いものにちつつ統合的に分析できるデータベースを構築するにあたり、下記に述べる課題がある。

1. データベース設計に関する課題

1) 標準データモデルへの統合

各々のRWDが準拠しているデータ構造を、Common Data Model (CDM) の構造にあわせて変換する。CDMは、ある目的を達成するための分析手法にあわせて設計された共通データモデルであり、データベースの場合はスキーマの形で実装される。世の中には様々なプロジェクトがあり、その数だけの種類のCDMが存在する。ここでいう「Common」とは、分析に関わるステークホルダー間における共通理解・合意という意味合いであり、ありとあらゆる場面に対応できるような共通規格を志向しているものではない。

2) CDMで利用する統制用語の管理、統合

我が国では医療情報標準化推進協議会の医療情報標準化指針に記載されている標準マスタ[2]を利用してデータを表現する方向で普及活動が進められているが、医療機関独自のマスタで運用しているところが多く、全ての医療機関が標準マスタに準拠したデータを提供するまでの過渡期にある。標準マスタ群のバージョン更新(例: ICD 10 -> ICD 11, JIAC (臨床検査項目標準マスタ) ver.10 -> ver.11)も控えている。さらに、臨床研究によっては薬剤の個別の製品コード・薬剤名称ではなく薬剤の分類を要求するケースがあり、データ提出側が正確な薬剤の分類ができるように、薬剤コードと臨床研究上の分類との対応表を提供する必要がある。CDMで利用

する標準的概念を網羅した標準統制用語集を策定し、各RWDが準拠しているマスタから標準統制用語集に変換するためのマッピングや、臨床研究における薬剤分類と薬剤コードのマッピング等を処理するツールの開発・提供が必要である。

3) 各施設の測定結果の組織間校正

各医療機関で異なる測定機器・試薬を採用している。また同一医療機関であっても測定機器・試薬を変更することにより、測定結果の分布や精度のバラツキがあり、複数の医療機関から収集した検査結果を単純に統合して分析することはできない。各々の測定結果の分布状況を踏まえて、標準的な分布モデルに変換する必要がある。

4) 患者個体の識別・追跡性の確保

データの精度、個人の追跡性を確保するために、個人を識別するIDの悉皆性、一意性を担保する必要がある。

2. データソースを巡るバイアス問題

我が国では、医療情報交換を推進するために、平成18年度に厚生労働省電子的診療情報交換推進事業(SS-MIX: Standardized Structured Medical Information eXchange)が立ち上げられた[3]。SS-MIX標準化ストレージ(以下SS-MIX)は、標準医療情報規格にもとづいた各種データを規則に従って階層化ファイルシステム上に展開することで、データのオンライン・オフラインにおけるやり取りのプロセスを標準化し、情報交換を容易にすることを目的としている。このSS-MIXを利用して、AMED事業[4]、診療録直結型全国糖尿病データベース事業[5]等で研究参加施設からの医療情報の収集が試みられている。

ただ、SS-MIXを取り巻く環境の構造的問題があり、データ提供者およびデータ収集者双方に運用上の負担が生じている。電子カルテシステム、オーダーリングシステムからのデータは、HL7 2.xメッセージファイルとしてSS-MIX標準化ストレージに保存される。HL7形式メッセージ以外の医療情報(DICOM, HL7 CDA等)は拡張ストレージに保存される。つまり、SS-MIXは階層化ファイルシステムを利用したファイルの保管庫にすぎず、データベースではない。従ってデータ分析者は、各々の分析目的のために構築したデータベースの構造にあわせて、SS-MIXに格納されたファイル群からデータを抽出・変換・加工してデータベースに取り込む処理(ETL: Extract, Transform, Load)を開発する必要がある。プロジェクトごとに独自にデータベースが構築されるため、同じデータソースを用いても、分析者によってデータベース構造やコホート定義手法が異なることによる新たなバイアスが生まれる可能性がある。また、SS-MIXに格納されるデータは医療情報システムのトランザクションデータから自動的に生成されるものが主流であり、分析者においてデータの修正や較正があっても、SS-MIXに格納されたデータに遡って修正されることは稀である。異なる分析者が同じSS-MIXをソースとすると

に、先行して取組んでいる分析者によるデータ確認・修正の経験が再利用される機会が少ない。そして、分析者達が独自の分析用データベースを構築することは、バイアスを増やす。同じ分析シナリオでも利用する観察研究データベースが異なると、異なる臨床的結論を導出してしまふ事例が報告されている[6]。また、診療報酬請求情報を利用した2つの異なる研究グループから、同じ種類のデータを用いたにもかかわらず臨床的に正反対の結論を下した論文が[7, 8]発表されている。以上のように、複数の観察研究データベース間でバイアスが生じることが報告されており、かつ同じデータソースであってもデータ抽出手法、データベース構造、コホート抽出方法、分析手法が異なることによるバイアスが発生している。RWDにもとづく観察研究データベースを用いた臨床研究にあたっては、医療情報・臨床疫学の研究者の双方が協力して取り組まなければならないテーマが屹立している。

まとめると、(1)RWDからのETL手法の標準化、(2)標準化されたデータベース構造の定義、(3)コホートを適切に定義する方法論、(4)RWDデータの素性の特定と品質改善、そして(5)RWDを収集する対象と時系列上の網羅性の確保が、医療ビッグデータ時代に課せられた優先度の高い研究テーマと言える。

III. RWD を活用する研究データベースと phenotyping プロセスの標準化の必要性

ビッグデータ時代においてRWDを活用するデータベースの課題を5つ提示したが、本稿では(2)標準化されたデータベース構造の定義、(3)コホートを適切に定義する方法論について論じる。(1)は、データベースの構造を標準化すれば、派生的に解決していくものであるし、(4)は(2)(3)の方法論が確立されてからの段階であり、(5)については医療政策やビジネスモデルの進展に密接に関係するものである。保健医療情報管理研究分野においては、データベースの構造の定義と、その定義によって影響されるコホートの定義手法に主な関心がある。

データヘルスに貢献できるデータベースを構築するには、多種多様なRWDを収集し、かつ各々のRWDをCDMや標準統制用語に準拠するように変換した上で、データベース上で統合する必要がある。また、そのデータベースの構造は研究を志望する者すべてに公開されて自由に利用できるようにする。検証者が研究内容を再現し追試できるように、コホート抽出のための条件・プログラムを論文のAppendixあるいは公開リポジトリに掲載する。そうすることによって、RWDを用いた臨床研究において透明性と再現性が担保される。これまでの臨床研究では、ランダム化比較試験、オーソドックスな統計解析手法、計算結果の精度保証がなされた統計処理ソフトウェアによって研究の信頼性が確保されていた。しかし、RWDを利用した観察研究にあたって、ランダム化比較

試験に代わるマッチング手法の開発の他に、ETLプロセス、データクレンジング、分析者が使用するデータベースの構造およびコホートを定義するプログラムの差異による新たなバイアスが発生している。新たなバイアスを乗り越えるために、標準化されたデータベースの構造定義およびコホートの抽出手法の開発をPDCAでまわしていく枠組み作りが喫緊の課題となっているが、我が国ではまだ存在しない。

主に電子カルテからのデータから患者の表現型を抽出（EHR-based phenotyping）、すなわち関心のある臨床的特徴を持つ症例を同定する手法に関して、明確に定義した文献はないが、データからコホートを同定する手法を“phenotyping”とする解釈に立って論じる[9]。以下、米国における観察研究データベースとphenotypingの標準化のイニシアチブについて取り組んでいるeMERGEとOHDSIプロジェクトについて紹介する。

IV. eMERGE プロジェクトにおける phenotyping の事例データベース

2012年 NIHはData and Informatics Working Group (DIAG) による生物医学ビッグデータによってもたらされる機会と課題を解決するためのNIH内を縦断した広範かつ包括的なプログラム確立の要請をうけて、BD2K (Big Data to Knowledge) イニシアチブを立ち上げた[10]。B2DKの目標は、(1)生物医学ビッグデータの蓄積、アクセス、共有、利用の推進、(2)データ解析手法とソフトウェアの開発と普及、(3)生物医学ビッグデータとデータサイエンスに関わる教育の強化、(4)データサイエンスのCOE (Center of excellence) の立ち上げ[11]である。B2DKに先行して、米国NIHの研究所である国立ヒトゲノム研究所 (The National Human Genome Research Institute : NHGRI) はゲノム医療実現に向けて6つの領域

The screenshot shows the PheKB (Phenotype Knowledge Base) website. The header includes the PheKB logo, a tagline 'a knowledgebase for discovering phenotypes from electronic medical records', and links for 'Login' and 'Request Account'. A search bar is also present. The main navigation bar has links for 'Home', 'Phenotypes', 'Resources', and 'Contact Us'. The breadcrumb trail indicates the path: 'Phenotypes » Phenotype 737'. The title of the selected phenotype is 'Rheumatology Auto-Immune characteristics'. Below the title are tabs for 'Phenotype', 'Data Dictionaries', and 'Implementations/Datasets'. The 'Phenotype' tab is active, displaying a message from Rachel (rknevel@bwh.harvard.edu) regarding the identification of cases with auto-immune rheumatologic phenotype (for NT198) and requesting information about auto-antibody and drug information. The message also mentions data dictionaries and files. Below the message, there are sections for 'Phenotype ID: 737', 'Status: Final', 'Type of Phenotype: Other Trait', 'Data Modalities and Methods Used: Laboratories, Medications', 'Files: Antibody list of alternative names, Drug list alternative names', 'Institution: Harvard Medical School', 'Date Created: Wednesday, June 28, 2017', and 'Age: Adult'.

図1 リウマチ自己免疫性疾患のphenotyping公開（本例では抽出条件としての抗体、医薬品のリストを提供）

のプログラム (Cancer Genomics, Pharmacogenomics, Genomic Medicine 'Test Drive' Programs, Newborn Genomic Analysis, Clinical Genomics Information Systems, Ultra-Rare Genetic Disease Diagnostics) を提供している。その中で、2007年に電子カルテのデータを活用してゲノム医療を促進するためのeMERGE (electronic Medical Records and Genomics) コンソーシアムが設立され、NHGRIによって研究資金が提供された[12]。トランスレーショナルリサーチにおいて研究段階から実際の医療・政策に移行するまで、(T0) 記述と発見、(T1) 発見から医療への応用 (治験、介入)、(T2) エビデンスの発見、(T3) 診療ガイドライン形成、(T4) 医療のアウトカム評価の5段階のフェイズがある。各々において疫学研究は重要な役割を果たしている[13]が、eMERGE-IIは主にT0フェイズ段階の研究プロジェクトとして位置づけられ、3つの主要な目的を掲げている。(1)電子カルテシステムのデータを使い、ロバストなphenotypingを確立する、(2)先述したphenotypingを用いて、全ゲノム相関解析 (Genome Wide Association Analysis : GWAS) を試みる、(3)電子カルテシステムを利用してGWASを実施し、広範囲にわたってデータを共有ことの倫理、法的、社会的影響の考察、である[12]。Coordinating Centerである Vanderbilt University中心にフェノタイピング手法の開発が行われ、その時の知見をもとにPheKB (Phenotype Knowledge Baseというフェノタイピング手法のデータベースを構築した[14]。PheKBは研究者が開発したフェノタイピング手法をデータベースに登録・公開し、手法について相互レビューを行う研究者コミュニティのための事例データベースである。これまで委細な抽出条件・ロジックは論文には記述されておらず、あったとしてもAppendix等に掲載されていることが多かった。PheKBによってフェノタイピング手法の検索や詳細な内容の閲覧が可能になり、フェノタイピング手法の妥当性に関する検討が容易になっている。2018年4月現在では、PheKBに収載されている最近のフェノタイプ手法として、1型・2型糖尿病、リウマチ自己免疫性疾患 (図1)、鎌状赤血球症[15]、Consistent Care、ピーナッツアレルギーが紹介されている。各々のフェノタイプ手法の記述として、抽出する条件としての病名[16]・抗体・医薬品、CPT (Current Procedure Terminology : 医療通用手技用語集) のリスト提供、ならびに自然文章による抽出ロジックが提供されている。背景となるデータベースや抽出ロジックの実装方法も様々であり、ソースコードが公開されていないものもある。従って、phenotypingについての検討ができる環境を提供しているが、phenotyping手法の記述、定義の標準化には至っていない。

我が国では、このPheKBに収載されているフェノタイピング手法を本邦のマスタやSS-MIXの事情に合わせて移植・改良したもので評価し、海外で発表されている成果と遜色のない高い精度で実現したことが報告されている[17]。ただ、前述したとおり、現状ではSS-MIX導入医

療機関においてデータ品質を高い状態にする運用がなされていないことが多いこと、ファイルシステム上にデータファイルを配置しているのみであるので“ビッグデータ”に対応できるスケーラビリティがないこと、phenotyping手法について共有するリポジトリが不在であるという課題が残っている。SS-MIXをビッグデータに向けた分散処理システム上に展開する試み[18]もあるが、汎用的な検索ができるデータベースとして再構成しないかぎり、phenotypeを設計する研究者にとって身近なプラットフォームとして利用できる可能性は低いと考える。

V. OHDSI プロジェクトによる CDM と Standard Vocabulary の構築

1. OHDSIの目的

Observational Health Data Science and Informatics (OHDSI) は、米国の官民共同 (PhRMA, FDA, FNIH) の取り組みとして発足したOMOP (Observational Medical Outcomes Partnership) を前身として、医療従事者、生物統計学者、コンピュータサイエンティスト、疫学研究、ライフサイエンス等の多様なステークホルダーからなる研究者集団によって推進されているプロジェクトであり、観察研究をとおして健康と疾患に関する包括的な理解をすることをビジョンとしており[19]、現在、OHDSIは6つの領域に取り組んでいる[20]。

1) データの標準化

観察研究の品質、効率性、透明性を実現し、かつ具体的なユースケースを最適にサポートできるようにデータベースの構造、内容、そして分析手法を標準化する。OMOP Common Data Model (CDM) を標準モデルとして採用し、またOMOP CDMを実装し運用するためのオープンソースのツールを開発・提供する。

2) 医薬品製品の安全性に関するサーベイランス

あらゆる医薬品製品への暴露と副作用の潜在的な関連を検出・評価する方法を開発するために、国際的なリスクの識別および分析システムを確立すること。全ての医薬品および関心のあるアウトカムをリアルタイムに探索できる大規模な分析環境を開発すること。オープンソースのエビデンス・リポジトリを確立して、複数の組織が参加して共有できるしくみをつくること。

3) 効果比較研究 (comparative effectiveness research : CER)

公衆は代替療法をうける権利があり、代替療法の有効性を既存療法と比較できるようにする。観察データからエビデンスを生成するためのオープンソースを開発する。「この医薬品は、あるアウトカムをもたらすか？」という疑問を追求するための安全サーベイランスツールとは異なり、CERツールでは「この医薬品は代替品よりも、多い、もしくは少ないアウトカムを形成し、同じ (治療) 目的に資することができるか？」といった疑問を追究する。

4) パーソナライズされたリスク予測

患者個人レベルの予測モデルは集団レベルの予測を補完し、平均的な治療効果に対する社会及び政策的課題に依っていくのみならず、過去の生活地域、病歴、健康行動にもとづいて将来のアウトカムを個別に洞察することを可能にする。このような個人レベルの予測を可能にするアルゴリズムとリスク情報を患者に直接伝えるツールを開発する。

5) データ特性

観察データから学び、信頼性かつ透明性が担保された現実世界のエビデンスを導出するために、RWDの素性について理解する必要がある。分析に使われるデータの多くは研究以外の目的で使われてきたデータ、例えば医療機関と保険者との間での診療報酬請求に関わる金銭的なトランザクションデータである保険請求や、医療従事者を支援する電子カルテシステムのデータである。これらのデータがどこから来て、どのような事由で生成されたものであるかを理解することが非常に重要である。データベースが利用に耐えるものであるか、分析にあたって考慮しなければならないデータ上の問題があるかを調査する品質評価とデータベースのプロファイリングツールを作成する。

6) 品質向上

OMOP CDM上に収録された観察データに対して体系的な品質評価をするオープンソースツールを開発する。新しい実証のエビデンスにもとづいた品質尺度の設計、開発、評価に取り組む。

2. OHDSIのCDM

OHDSIは先述したとおり、OMOPプロジェクトを前身とする経緯があり、CDMとしてOMOP CDMを採用している。大元の利用目的や設計が異なるヘテロジニアスな医療情報をOMOP CDMと標準統制用語に適合するよ

うに変換することで、最終的に複数組織・システムにまたがって横断的な検索・分析が可能なデータベースを構築できる。OMOP CDMの仕様書はCreative CommonのPublic Domainライセンスで公開されており、特別の会員登録なしに閲覧することが可能である[21]。OMOP CDM（以下CDM）は大別して6領域に分類される数十個のテーブルにより定義されている。以下、各領域について最新版のCDM Ver5.3に基づいて概説する。特にphenotypingの標準化に大きく貢献する概念周りの管理について詳説する。

1) Standard Vocabularies

複数の統制用語集からOMOP CDM内で使用される概念（Concept）へのマッピングと、その概念の利用が許可される場所（Domain）を定義するのが主な役割である（図2）。全ての統制用語集で定義されている概念はCONCEPTテーブルに登録される。各々の統制用語集における概念のIDが重複しないように、OHDSIの開発グループによって各々の概念に対して一意の管理番号が付与されている。これらの概念の中から、OMOPにおけるデータを表現するための標準的概念（Standard Concepts）が指定される。標準的概念は一つの領域（Domain）のみでの使用が認められるが、この「領域」とは、後述するStandardized Clinical Data Tablesにおける各テーブルに対応する。例えば、病名に関する概念は、疾患に関する情報を収載する領域以外（OMOP CDMではCONDITION_OCCURENCEテーブルが該当する）では使用出来ない。標準的概念を利用できる領域（データベースのテーブルに対応）を制御することにより、意図された使い方ではない形でデータが記録されることを抑止する狙いがある。DOMAINテーブルはOMOP CDMにおいて、概念を利用できる場所を制約する「領域」を定義する。現在Cost, Gender, Drug, Visit, Device等の40種類の領域が定義されている。一方、CONCEPT_CLASSで概

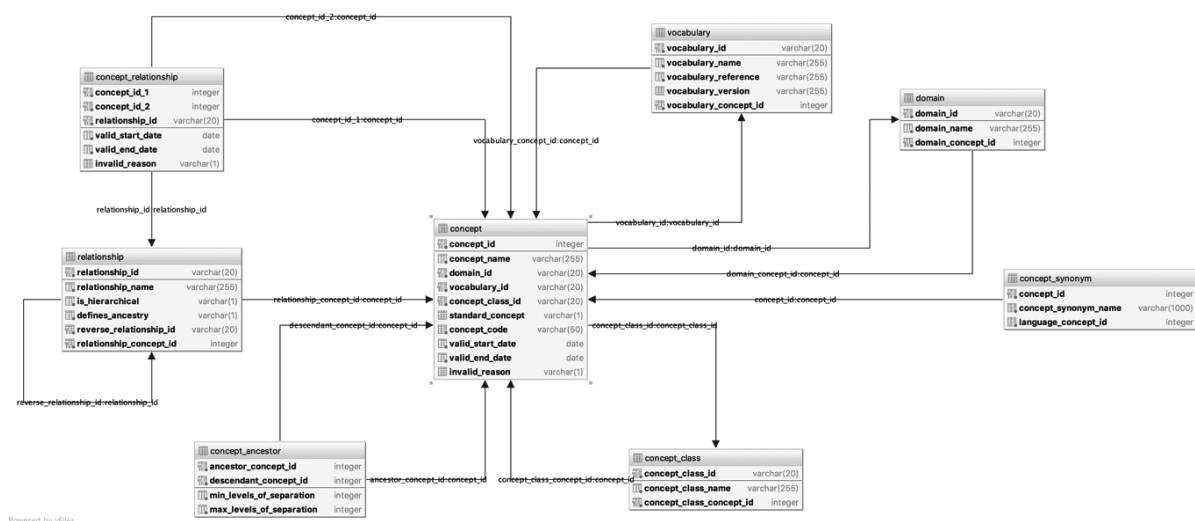


図2 概念を管理するOMOP CDMのテーブル群（CDM v5.3のスキーマを可視化したもの）

念のクラス（例、Procedure、Clinical Finding等）を定義しており、各々の概念の分類を定義している。以上のように、OMOP CDMのStandard Vocabularyにおいて複数の統制用語集から採用した概念の利用上の混乱が生じないように、概念のクラス分類、CDMにおける各項目の値域としての概念の領域を定めている。結果、phenotypeを構築する際に、条件として定義する概念を検索するにあたっての曖昧さを回避し、信頼性・再利用性の高い検索式を構築するのに貢献している。VOCABULARYテーブルは、CONCEPTテーブルに収載された概念を収録している統制用語集の一覧を管理している。OMOP CDM ver 5.3時点で独自に定義している用語集も含めて70程度の用語集が登録されており、登録された概念の素性が追跡できるように管理されている。

RELATIONSHIPテーブルは概念間の関係の種類を管理しており、実際概念間の関係性はCONCEPT_RELATIONSHIPで管理されている。後述するように、phenotypingにあたって疾患名のグルーピングや同じ効能あるいは成分を保有する医薬品を一括指定するために、このCONCEPT_RELATIONSHIPで定義されている関係性が利用される。この関係性の定義が公開されて研究者で共用されることにより、研究者によってphenotyping時に条件として指定する疾患名や医薬品の選択対象に違いがでるというバイアスを抑止している。

2) Standardized Metadata

CDM_SOURCEテーブルでCDMに取り込んだデータの発生源について管理しており、データソース名と説明、ETL手法について記述した文書へのURL、データの取り込み日、当時に準拠していた統制用語集のID等を記録する。ただし、現在はCDMに格納されたデータについて、このCDM_SOURCEとのリレーションが構成されていないため、ただちに各々のデータがどのデータソース、ETLのプロセスを経て取り込まれたかを特定できる状態にはなっておらず、設計上の改善の余地があると思われる。

3) Standardized Clinical Data Tables

観察期間内における各患者の臨床イベントは、このCDMに記録される。患者基本情報（PERSON）、各々における観察期間（一人につき、複数の観察期間の対応関係がある）（OBSERVATION_PERIOD）、生体検査（SPECIMEN）、患者の死亡日と死因（DEATH）、医療機関への来訪履歴（外来、入院、救急外来、長期ケア）（VISIT_OCCURRENCE）、診断や治療等の診療行為（PROCEDURE_OCCURRENCE）、内服・注射（DRUG_EXPOSURE）、植え込み機器（ペースメーカー・ステント・人工関節等）や機器（包帯、松葉杖、シリジ等）、その他（縫合、除細動器）（DEVICE_EXPOSURE）、医師の診断や患者の主訴にもとづく疾患や病態、症候（CONDITION_OCCURRENCE）、検査結果、バイタル、病理診断のうち定量分析結果など、数値やカテゴリカル変数で構成される構造的なデータ（MEASUREMENT）、

診療録（NOTE）、診療録をNLP処理した結果を格納するもの（NOTE_NLP）、検査・問診・処置等を通して得られた臨床に関する事実、生活習慣、病歴、家族歴等他の領域にあてはまらないものはここに格納される（OBSERVATION）、CDMにレコードとして格納されている各ファクト間の関係性を定義（例：PERSON間の親子関係、病態と処方への対応、ある処置で利用された医療機器への対応、提出された生体と測定結果への対応等）（FACT_RELATIONSHIP）を収録する。

4) Standardized Health System Data Tables

患者や医療機関の所在地の管理テーブル（LOCATION）、医療機関の種類、所在地（CARE_SITE）、医療従事者の基本情報（PROVIDER）等、個人や組織に関する情報はここで集約して管理される。

5) Standardized Health Economics Data Tables

保険の種類と有効期間（PAYER_PLAN_PERIOD）、各診療行為に関わる諸コストとして、DRUG_COST（DRUG_EXPOSUREに要したコスト）、PROCEDURE_COST（PROCEDURE_OCCURRENCEに要したコスト）、VISIT_COST（VISIT_OCCURRENCEに要したコスト）、DEVICE_COST（DEVICE_OCCURRENCEに要したコスト）等が定義されている。

6) Standardized Derived Elements

ここに分類されているテーブルはデータソースからのデータではなく、他のテーブルから計算処理などをして派生的に導出された結果を格納するためにある。RWDはトランザクションデータであり、各々のイベントは一日単位で集計されていることが多く、「イベント期間」という概念がない。しかし、コホート抽出に際してイベント期間も条件になることがある。そこで、例えばDRUG_EXPOSUREは個々のレコードが患者に投与された医薬品に対応している。連続した投与履歴から所定の規則にしたがって「投与期間」を算出してDRUG_ERAに格納する。分析者は投薬のイベント期間を確認したいときにDRUG_ERAに格納されたイベント期間を参照する。

また、COHORTテーブルはCOHORT_DEFINITIONテーブルで定義されたコホート抽出のロジックを実行して得られたコホートを格納する。

3. OHDSIのエコシステムとしてのツール群

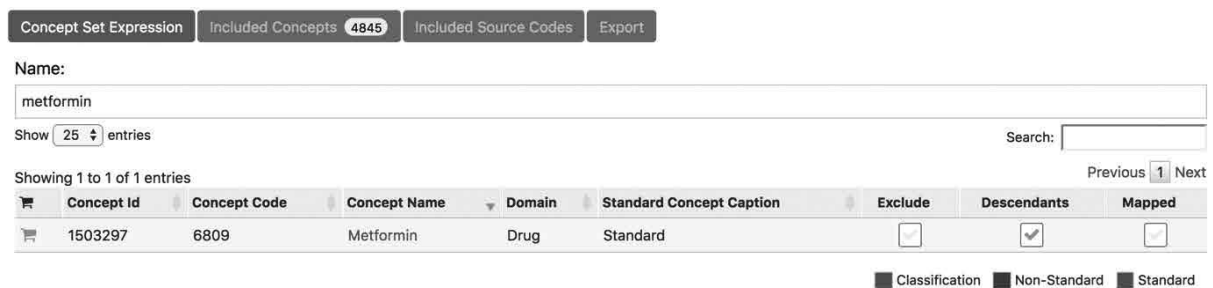
OHDSIプロジェクトでは、先述した6つの領域の研究を支援するためにオープンソースのツールを多数開発し、エコシステムを構成している。OHDSIで利用されている標準統制用語集を管理しているATHENAリポジトリ、ローカルコードからCDMで使用されている標準統制用語集へのマッピングを支援するUSAGI、各種ファイル・データベースからCDMへETLを支援するWHITE RABBIT、大規模な時系列データの特性を概観できるダッシュボードを提供するACHILLES、標準統制用語集から暴露、アウトカム、共変量を定義するためのHER-

MES, データベースや統制用語集の閲覧やコホートの定義・集団レベルの分析を提供するATLAS, コホート抽出・自己対象ケースシリーズ・ケースコントロール研究・ケース・クロスオーバー研究に利用できる統計処理ソフトウェアR用ライブラリ等が開発され, CDMを利用する研究者によって利用・レビュー・改善が行われている。本稿ではphenotypingに焦点をあてているので, コホート定義・抽出を提供しているATLASについて説明する。ATLASは, OMOP CDM v5に準拠して構築されたデータベースに対して, 統制用語集から暴露した医薬品や診断病名の概念を検索の上, 当該概念をコホート抽出の条件として指定・検索することを支援するWebベースのツールである[22]。

4. ATLASを利用したコホート抽出

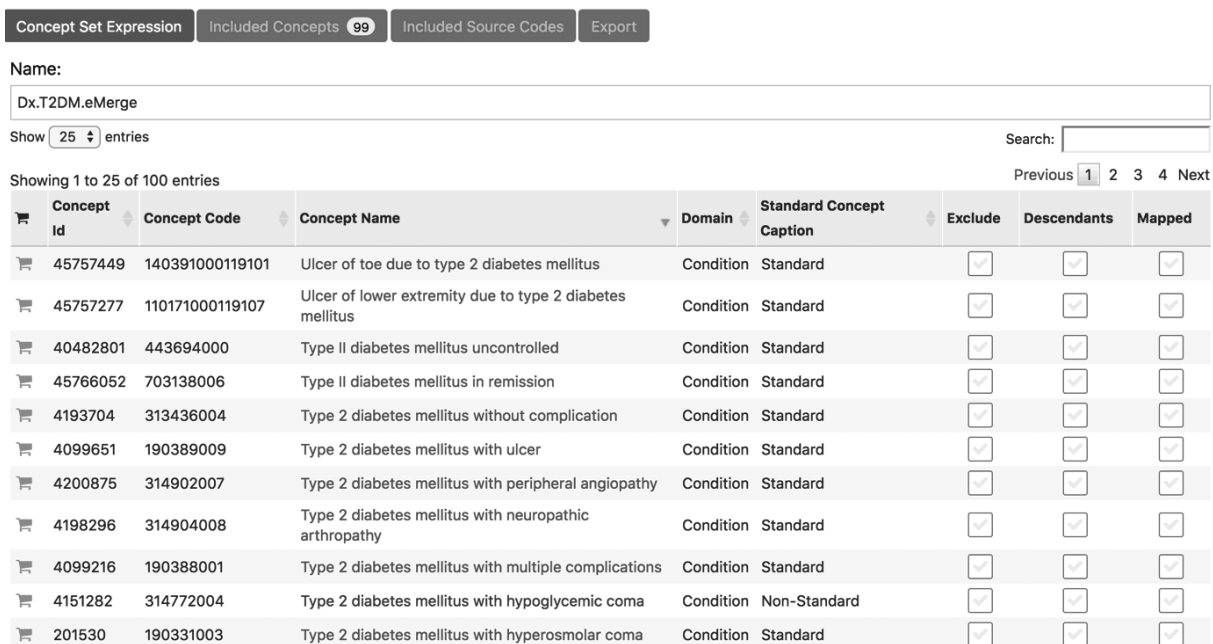
ATLASでは複雑なプログラム言語を利用することなく, グラフィカルユーザーインターフェース上で対話的にコホート定義ができるようにしている。コホート定義にあたってConcept Set (概念セット) の定義を背景として,

コホート抽出のロジックを組み立てていく。概念セットは, 抽出条件を指定する際にパラメータとして与える概念群をとりまとめたものである。デモサイトに掲載されていたコホート定義例から紹介する。「40歳以上で, メトフォルミンを初めて処方された (インデックス日付より180日以上前に当該履歴がないもの, かつ最も初期のイベントであること) 者で, かつ次の2つの条件(1)インデックス日以前に2型糖尿病の病名がつけられていることならびに, (2)ピオグリタゾンの投与歴が一度もないこと」を満たしているコホートを抽出したいとする。何かを対象とするときには, 標準統制用語集にて標準概念として指定されている概念群から選んだ概念を条件として指定する必要がある。上記のコホート定義では, 「メトフォルミン」「2型糖尿病」「ピオグリタゾン」が相当する。「メトフォルミン」はRxNorm[22]から採用された概念を利用しており, "Descendants"属性を有効にして, メトフォルミンに関連している下位概念を対象にすることで「メトフォルミン」系統の医薬品を全て対象にしている (図3)。



Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption	Exclude	Descendants	Mapped
1503297	6809	Metformin	Drug	Standard	☐	☒	☐

図3 メトフォルミンの医薬品をRxNormの階層関係を利用して一括指定している例



Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption	Exclude	Descendants	Mapped
45757449	140391000119101	Ulcer of toe due to type 2 diabetes mellitus	Condition	Standard	☒	☒	☒
45757277	110171000119107	Ulcer of lower extremity due to type 2 diabetes mellitus	Condition	Standard	☒	☒	☒
40482801	443694000	Type II diabetes mellitus uncontrolled	Condition	Standard	☒	☒	☒
45766052	703138006	Type II diabetes mellitus in remission	Condition	Standard	☒	☒	☒
4193704	313436004	Type 2 diabetes mellitus without complication	Condition	Standard	☒	☒	☒
4099651	190389009	Type 2 diabetes mellitus with ulcer	Condition	Standard	☒	☒	☒
4200875	314902007	Type 2 diabetes mellitus with peripheral angiopathy	Condition	Standard	☒	☒	☒
4198296	314904008	Type 2 diabetes mellitus with neuropathic arthropathy	Condition	Standard	☒	☒	☒
4099216	190388001	Type 2 diabetes mellitus with multiple complications	Condition	Standard	☒	☒	☒
4151282	314772004	Type 2 diabetes mellitus with hypoglycemic coma	Condition	Non-Standard	☒	☒	☒
201530	190331003	Type 2 diabetes mellitus with hyperosmolar coma	Condition	Standard	☒	☒	☒

図4 2型糖尿病の概念を網羅的に指定したケース (部分掲載)

Concept Set Expression	Included Concepts 18	Included Source Codes	Export
------------------------	----------------------	-----------------------	--------

Name: T2DM

Show 25 entries Search:

Showing 1 to 1 of 1 entries Previous 1 Next

Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption	Exclude	Descendants	Mapped
201826	44054006	Type 2 diabetes mellitus	Condition	Standard	☐	☒	☒

Classification Non-Standard Standard

図5 2型糖尿病の概念をSNOMEDの階層関係を利用したケース

「2型糖尿病」の概念について、一例では99種類の概念を個別に指定している（図4）。他の例では、SNOMEDの「Type 2 diabetes mellitus」とその下位概念を対象とすることで簡潔に指定するものもある（この場合は下位概念として18種類の病名が対応している。）

（図5）。個別指定か、階層関係を利用した一括指定のいずれかを利用するかは、分析者がどのような粒度で見ていきたいかに依る。RxNormやSNOMEDにおける概念の階層関係を利用した概念の指定が分析目的にそぐわない場合は、分析目的に合致するような概念の階層関係を独自に定義して、その階層関係を利用した概念セットを利用することも可能である。ただし、その独自に定義した概念セット、あるいは概念の階層関係を公開して共有

しなければ、研究者独自の抽出条件設定によるバイアスが生じる可能性がある。

以上のように、例では各々に対応する概念のセットとして、「metformin」「T2DM-PBA」「Pioglitazone」の3つの概念セットを定義している。

概念セットを定義した後は、GUIでコホートを定義する条件を組み立てていく（図6）。Initial EventのCriteriaとしてCondition Era, Condition Occurrence, Death, Device Exposure, Dose, Drug, Drug Exposure, Measurement, Observation等の基準を追加することができる。例では、最初にDrug Exposureの条件を「Add Initial Event」ボタンをして追加し、metformin概念セット（緑色ボタンより選択）を選ぶことでメトフォルミンを対

ATLAS

Home Data Sources Vocabulary Concept Sets Cohort Definitions Incidence Rates Profiles Estimation Prediction Jobs Configuration Feedback

Cohort #1747567

P-Comparator-new

Definition Concept Sets Generation Reporting Export

P-comparator

Initial Event Cohort

People having any of the following:

+ Add Initial Event

a drug exposure of metformin

+ Add criteria attribute...

Delete Criteria

for the first time in the person's history

with age Greater Than 40

with continuous observation of at least 180 days before and 0 days after event index date

Limit initial events to: earliest event per person.

Initial event inclusion criteria: From among the initial events, include:

having all of the following criteria:

+ Add criteria to group...

with at least 1 using all occurrences of:

a condition occurrence of T2DM-PBA

+ Add criteria attribute...

Delete Criteria

starting between All days Before and 0 days After event index date and ending any time.

restrict to the same visit occurrence

and with exactly 0 using all occurrences of:

a drug exposure of Pioglitazone

+ Add criteria attribute...

Delete Criteria

starting between All days Before and All days After event index date and ending any time.

restrict to the same visit occurrence

Limit cohort of initial events to: earliest event per person.

Remove initial event inclusion criteria

Additional Qualifying Inclusion Criteria

New qualifying inclusion criteria Please select a qualifying inclusion criteria to edit.

Limit qualifying cohort to: earliest event per person.

Cohort Exit Criteria

Cohort exit criteria based on a fixed time period relative to initial event start or end date:

A cohort end date is derived from adding a number of days to be offset from the specified initial event date. If an offset is added to the initial event start date, all cohort episodes will have the same fixed duration (subject to further censoring from other cohort exit criteria). If an offset is added to the initial event end date, persons in the

図6 コホートの定義画面

象にする。さらに最初の被爆，年齢条件の指定を"Add criteria attribute"ボタンをおして追加する。以下，イベントや条件を指定するためのテンプレートを利用してコホートを定義する条件式を完成させていく。CDMと統制用語集によってデータ構造とデータ内容が標準化され，曖昧性が排除されるため，コホートの抽出ロジックはアルゴリズム的に一意なものを生成することが可能になる。従って，自然文章によるコホートの定義からコホートを抽出するロジックを実装する際に実装者の解釈の違いや開発環境の差異によって発生するバイアスは排除される。

VI. 今後のとりくみについて

eMERGEやOHDSIの取り組みと我が国でのCDMに関する取り組みを俯瞰した上で，今後の取り組みの方向性について提案したい。主に医療機関内でのデータ分析のために構築されているDWHは大規模病院での導入にとどまり，広く普及していない。DWH構築にかかる工数と品質保証の負担が大きいためである。データソースとしてSS-MIXは有力であるが，未だSS-MIXに収録されていない医療情報がある。従って電子カルテシステムや部門システムからデータを取り出してDWHに格納する作業は当面必要である。医療情報システムおよびDWHの情報モデルと統制用語が統一されていないので，医療情報システムとDWH間のETLプログラム開発は双方のシステムに習熟している人材への属人性が高い。しかしながら，そのような人材の確保が困難である事情もDWH普及の障壁の要因の一つである。そして，DWHベンダーごとに，DWHについて独自の情報モデルを採用しており，ETLプロセスがブラックボックスになっているため，ユーザーが利用している医療情報システム内のデータ項目とDWH内のデータ項目との対応関係が不明である。結果，ユーザにとって必要なデータを探し出す作業の手間が発生したり，発生源データとDWHに収

録されたデータの整合性の検証も余儀なくされたりする。DWHの構造が異なるために，分析プロセスの移植を強いられたり，Business Intelligenceツールによるデータアクセスの特性を配慮したキューブ・非正規化のノウハウを他のユーザと共有できず，知見の共有・検討・改善のPDCAを円滑に回せない。これらの問題を解決するために，一般社団法人SDMコンソーシアムが中心となり，多次元データモデルの設計手法にもとづいたSemantic Data Model (SDM) というCDMを開発し，DWHベンダーやユーザに対する教育，データの検証，ETLプロセスの開発・コンサルティング等の活動を行っている[23]。このような状態にある我が国において，海外で公表されているCDMを一足飛びに実装することは困難であろう。理想論を言えば，openEHRのarchetype[24]のような多くの疾患概念を可及的に細粒度，網羅性をもって収録することを目標とした標準情報モデルにRWDをマッピングしてから，さらにCDMに変換することを検討すべきと考えるが，我が国ではそのようなイニシアチブを取る土壌が整っていない。医療情報学の実地面では，有限の期間に達成可能な方法論を検討することも重要である。OMOP CDMは国内で使われていない標準統制用語集に依存しており，国内で利用されている標準統制用語集とのマッピングが存在しないこと，CDMの英語で書かれた仕様書を理解し実装できる人材が不足している。そこで，このSDMをゲートウェイDWHとして位置づけ，国内におけるDWHの普及やデータ品質の向上に関する活動と，国内外のCDMの取り組みとのハーモナイズ活動との責任分解点として設定することを検討している（図7）。すなわち，国内の独自の事情を踏まえた，医療情報システムやSS-MIXからのETLおよびデータ品質の担保の取り組みはDWHベンダーも参加している一般社団法人SDMコンソーシアムが中心になって取り組んでビジネスモデルとして発展させる一方，研究者側はSDMを下地とし，CDM間マッピングとフェノタイプ

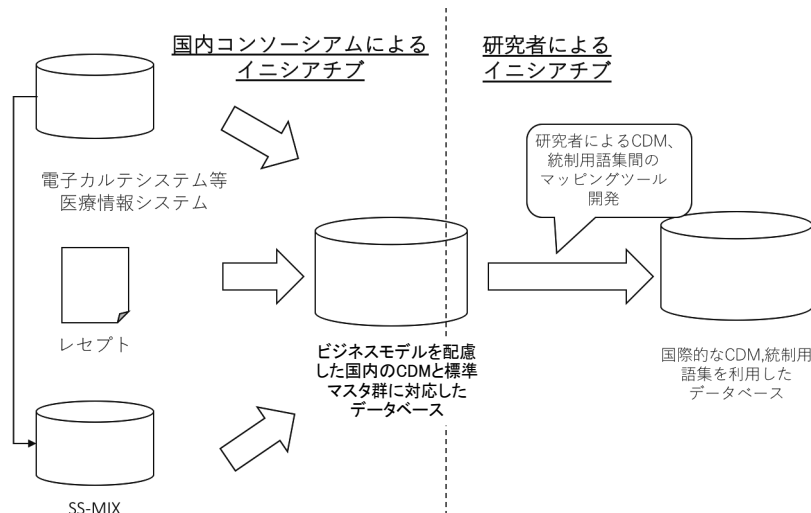


図7 2つのイニシアチブによる国内のデータソースから国際的なCDMへの橋渡しの構想

グ手法を開発して公開することで、RWDを臨床研究に貢献させる環境作りに必要なデータソースの確保の労力を削減できるという双方向の協力関係が築けるのではないかと考えている。

VII. 結語

データヘルスの課題として、悉皆的に収集されたRWDからエビデンスを導出できるような取り組みが求められている。データベース設計に関する課題とデータソースを巡るバイアス問題を提示した。これらの課題を踏まえて、RWDから観察研究データベースを構築する際の課題を提示した。本稿では、標準化されたデータベースとコホートを適切に定義する方法論について焦点をあて、先行研究としてeMERGEプロジェクトにおけるPheKBというフェノタイピング手法を公開するリポジトリと、OHDSIプロジェクトにおけるOMOP CDMによるデータベースの構造の標準化と標準統制用語集の取り組みについて紹介した。今後の我が国における取り組みの一案として、国内事情を踏まえたDWHの取り組みと国内外のCDMと統制用語集のマッピングの取り組みとの間に責任分界点を設定することを提言した。今後の課題として、国内標準統制用語集群とRxNorm, SNOMEDにて定義されている概念間の対応について調査し、また概念間階層関係を利用した概念セットの定義が国内のデータに対しても適切に適用された上で、海外のphenotypingの先行事例を移植あるいは相互交換性を維持できるかについて追究していきたい。

利益相反

本論文に関して開示すべき利益相反はない。

引用文献

- [1] de Lusignan S, Crawford L, Munro N. Creating and using real-world evidence to answer questions about clinical effectiveness. *Journal of innovation in health informatics*. 2015;22(3):368.
- [2] 医療情報標準化推進協議会 (HELICS協議会). 「医療情報標準化指針」一覧. 2018. <http://helics.umin.ac.jp/helicsStdList.html> (accessed 2018-04-28)
- [3] Kimura M, Nakayasu K, Ohshima Y, Fujita N, Nakashima N, Jozaki H, et al. SS-MIX: a ministry project to promote standardized healthcare information exchange. *Methods of information in medicine*. 2011;50(2):131-139.
- [4] 石田博, 研究開発代表者. 日本医療研究開発機構補助金臨床研究等ICT基盤構築研究事業「既存の診療情報と一体的に運用可能な症例登録システムの構築とアウトカム指標等の分析・利活用に関する研究」平成28年度研究成果報告書.
- [5] Sugiyama T, Miyo K, Tsujimoto T, Kominami R, Ohtsu H, Ohsugi M, et al. Design of and rationale for the Japan Diabetes compREhensive database project based on an Advanced electronic Medical record System (J-DREAMS). *Diabetology International*. 2017;8(4):375-382.
- [6] Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the impact of database heterogeneity on observational study results. *American journal of epidemiology*. 2013;178(4):645-651.
- [7] Kido T, Muramatsu K, Yatera K, Asakawa T, Otsubo H, Kubo T, et al. Efficacy of early sivelestat administration on acute lung injury and acute respiratory distress syndrome. *Respirology*. 2017;22(4):708-713.
- [8] Kishimoto M, Yamana H, Inoue S, Noda T, Myojin T, Matsui H, et al. Sivelestat sodium and mortality in pneumonia patients requiring mechanical ventilation: propensity score analysis of a Japanese nationwide database. *Journal of anesthesia*. 2017;31(3):405-412.
- [9] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*. 2013;21(2):221-230.
- [10] Health NIO. Big Data to Knowledge. 2017. <https://commonfund.nih.gov/bd2k> (accessed 2018-04-28)
- [11] Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*. 2014;21(6):957-958.
- [12] Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*. 2013;15(10):761.
- [13] Khoury MJ, Gwinn M, Ioannidis JP. The emergence of translational epidemiology: from scientific discovery to population health impact. *American journal of epidemiology*. 2010;172(5):517-524.
- [14] Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*. 2016;23(6):1046-1052.
- [15] Michalik DE, Taylor BW, Panepinto JA. Identification and validation of a sickle cell disease cohort within electronic health records. *Academic pediatrics*. 2017;17(3):283-287.
- [16] Upadhyaya SG, Murphree DH, Ngufor CG, Knight AM,

- Cronk DJ, Cima RR, et al. Automated Diabetes Case Identification Using Electronic Health Record Data at a Tertiary Care Facility. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*. 2017;1(1):100-110.
- [17] 香川璃奈, 河添悦昌, 井田有亮, 篠原恵美子, 今井健, 和彦大. SS-MIX2標準化ストレージを用いた2型糖尿病症例のEHR Phenotypingアルゴリズムの開発と評価. *医療情報学*. 2014;Supl.35:672-675.
- [18] Kimura E, Ishihara K. Virtual file system on NoSQL for processing high volumes of HL7 messages. *Studies in Health Technology and Informatics*. 2015; 210:687-691.
- [19] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in health technology and informatics*. 2015;216:574.
- [20] Informatics OHDSa. OHDSI 'Areas of Focus'. 2018. <https://www.ohdsi.org/who-we-are/areas-of-focus/> (accessed 2018-04-28)
- [21] Reich C, Belenkaya R, Natarajan K, Blacketer C. OMOP Common Data Model v5.3 Specifications. 2017. <https://github.com/OHDSI/CommonDataModel/wiki> (accessed 2018-04-28)
- [22] Observatioal Health Data Sciences and Informatics. ATLAS. 2017. <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:atlas> (accessed 2018-04-28)
- [23] 紀ノ定保臣, 山口永真, 英夫鈴. データの二次利用を支援するDWHの構築と運用をめざして. *IT VISION*. 2018;37:46-47.
- [24] Garde S, Knaup P, Hovenga EJ, Heard S. Towards Semantic Interoperability for Electronic Health Records--Domain Knowledge Governance for open EHR Archetypes. *Methods of information in medicine*. 2007;46(3):332-343.