IQVIA ソリューションズ ジャパン合同会社

SIP 第２期「スマートバイオ産業・農業基盤技術」における バイオデータ連携・利活用に係る成果の普及業務

1. はじめに

1.1 調査と報告書の目的

戦略的イノベーション創造プログラム（SIP）第２期（2018年～2022年）における「スマートバイオ産業・農業基盤技術」課題（以下「課題」という。）では、その活動の一部として、内閣府のバイオ戦略及び「バイオデータ連携・利活用に関するガイドライン中間まとめ（以下「中間まとめ）という。」に基づき、課題内で生み出された多種多様なバイオデータの連携・利活用の実践を図り、またその実践の中で得られた成果や課題、改善に向けた取組を成果としてとりまとめた。この成果は、「バイオデータ連携・利活用に向けたガイドブック（以下「ガイドブック」という。）」として公開されるに至っている。

他方、課題では、国内フードチェーンを構成する幅広い関係者のデータ連携・利活用の基盤となる情報システムの開発・社会実装にも取り組んできた。そのためバイオデータ連携・利活用に係る課題の成果の普及を図ることは、課題において開発・社会実装したバイオデータ連携・利活用基盤の活用の拡大や新たな価値創出を促すことが期待され、課題の成果の価値最大化につながるものである。

バイオデータに関連する分野においては、これまでに複数のデータベース構築の取組がなされ、実際に成果を産んでいるものも多い。一方で諸外国と比較した場合、それらデータベースの質と量は我が国のバイオ産業の育成に対して必ずしも十分とは言えず、国家あるいは地域間における競争優位とまでに育っているとは言い難い状況である。そこで、画期的な打開策となりうる方法として、データベース間の“連携”による利活用の促進を提案したい。即ち、個々のバイオデータを適切に連携する基盤の構築により、上記の競争優位足るバイオデータの質と量を確保しようとするものである。

本事業では、バイオデータ連携・利活用に係る課題の成果をバイオデータの連携・利活用の可能性のある者を対象に広く普及させることを目的とし、特に海外において様々な取組が先行している健康・医療領域の事例を丁寧に紹介しながら、そのエッセンスを抽出することにより国内における幅広いバイオデータ連携・利活用に資する提言を作成する。認定バイオコミュニティを始めとする、国内関係者におかれては、ガイドブック

と共に本調査報告書を十分に参考とされ、バイオデータの連携・利活用による価値創出を推進頂ければ幸いである。

1.2　調査報告書の要旨

バイオデータ利活用のための、データ基盤のあり方は概念として大きく 2 つに分けられる。１つ目の方法は、現在まで主流となっている、複数のデータ発生元からデータを転送、統合し、中央においてビッグデータベースを構築する方法（以下、「集合型」という。）である。この例としてはレセプト情報・特定健診等情報データベース（NDB）、介護保険総合データベース、国保データベース（KDB）などが挙げられる。バイオデータの利活用が将来のバイオ関連産業の競争力の源泉になるとの認識の下、世界では広範なバイオデータの囲い込みの動きが進んでおり、我が国としても、バイオデータ利活用の取組を進めていくことが重要な課題であることはいうまでもない。一方で、諸外国と比較した時に、我が国は集合型データの囲い込み競争に十分な優位を持つことができているとはいえない状況であること、集合型データの維持、管理、発展、そして利活用においては多大な費用および工数、情報セキュリティならびにプライバシー保護の課題が大きいことから、今後の発展に向けた方向性を「集合型データ」の構築ならび囲い込みに依存していくことに対する限界も指摘されている。

上記の「集合型データ」利活用における課題は国内に限ったものではない。国際間競争対象となる諸外国も同様の状況となりつつある。即ち、集合型ビッグデータ構築方向性の行き詰まりである。これに対し、実際に諸外国においては、先行して集合型に頼らない基盤の構築が進められている。これがデータ基盤の概念の 2 つ目である、「連合型(Federation)」と呼ばれる分散型データ連携・利活用のための仕組みである。本調査報告書第 3 章においてはリアルワールドデータ（以下、RWD という。）の汎用データモデル (Common Data Model)をこの連合型のデータ連携・利活用を進めるための重要な手段として解説する。また、本調査報告書第 4 章においては、実際の欧米ならびにアジア地域における先進事例とその成果ならびに関連技術に関して詳述する。上述の集合型データ活用の課題を踏まえ、我が国がバイオデータ利活用の「勝ち筋」を見出すためには、連合型データ連携で国際競争力を構築することが戦略的に不可欠であり、本報告書においては、全体を通してこの「連合型」を大きな戦略的キーワードとして解説することとする。

導入として上記「集合型」と「連合型」の概念を機会学習におけるフレームワークを用いて簡潔に説明する[図 1]。「集合型」は従来から用いられている方法であり、データ発生元各所からデータを中央に転送し、データ記載内容のチェック、規格統一などのクレンジング作業を通じてビッグデータベースを作成、そのデータを材料として機会学習をモデリングする方式である。一方、「連合型」は、データ発生元それぞれにおいて、それぞれのデータを材料として機械学習を行った“結果”を中央で統合する。さらに、その統合学習結果をデータ発生元にフィードバックするサイクルを繰り返すことで精度を高めていく方式である。この方式ではデータそのものを転送、統合する必要がないため、中央におけるデータの維持管理コスト（リソース）の平準化が可能であり、加えてデータ発生元における秘密情報やプライバシー情報の保護にも適している。この連合学習の国内バイオ産業における実例として、創薬における取組を紹介する。AMED 事業である産学連携による次世代創薬 AI 開発（DAIIA）は、産学連携を通じて AI 技術を活用した創薬支援基盤の構築を目的としており、この取組における化合物プロファイル予測 AI に連合学習が用いられている。創薬化合物情報は企業競争力に直結する秘匿性の高いものであり、企業外に持ち出すことができないデータについては、それぞれの企業内で AI 学習を実行し、AI モデルパラメーターのみを交換することで、複数組織が協調して AI モデルを構築することが可能な仕組みが構築されている(*)

連合学習 - Federated Learning

学習結果（モデルパラメータ）を統合することで、データの秘匿性とプライバシーに配慮

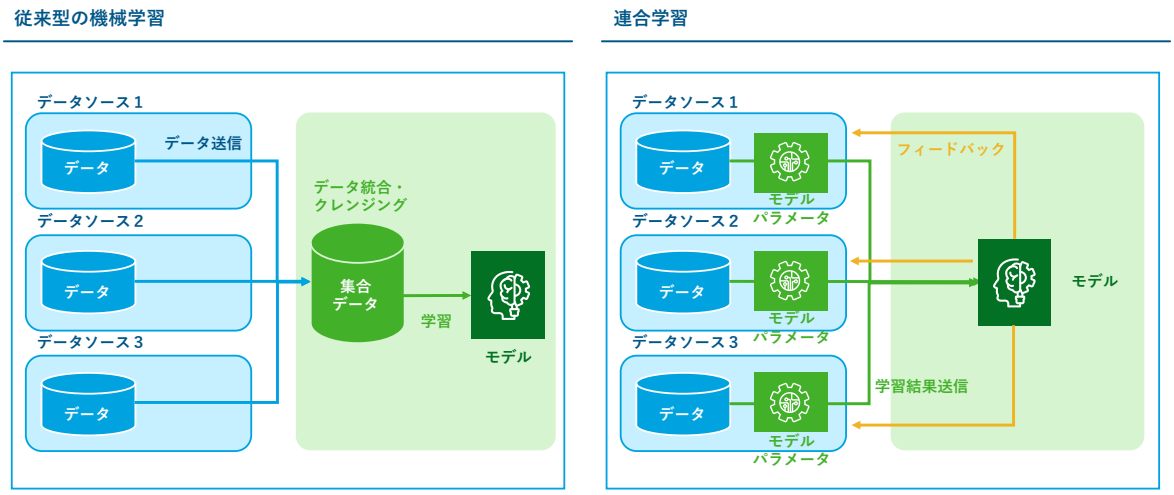


図 1. 連合学習 – Federated Learning

連合学習の概念はバイオデータの連携・利活用にも応用可能である[図 2]。上述のとおり、実際に RWD の利活用における連合型データ解析は諸外国において先行して広がっており、今後我が国も先進事例を参照しながら戦略的に構築に取り組むべきモデルである。

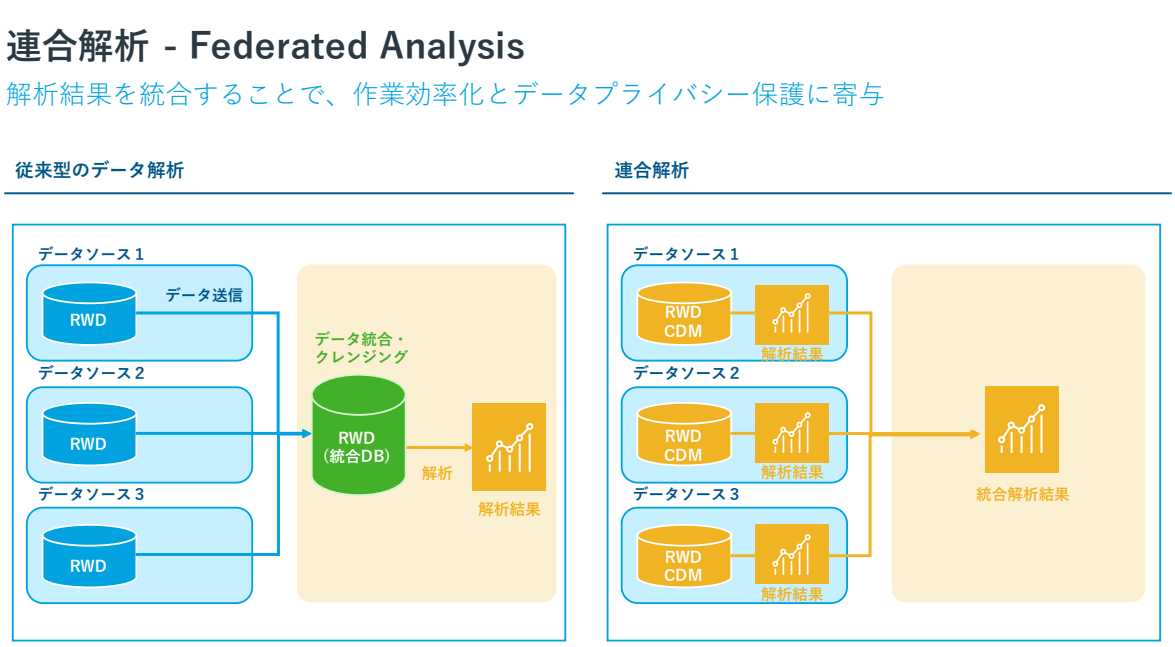


図 2.　連合解析　– Federated Analysis

詳細な説明は後章に展開していくこととするが、バイオデータの連携・利活用関係者においては、本報告書により「連合型」の概念ならびに実践的な構築、活用方法を理解頂き、国内におけるバイオデータ連携・利活用を推進頂くことができれば幸いである。

[参照]
*産学連携による次世代創薬 AI 開発（DAIIA）:
https://www.amed.go.jp/program/list/11/02/001_02-04.html

2. 健康・医療領域におけるデータ連携・利活用の先進事例

2.1　健康・医療領域におけるデータ連携・利活用の種類

本章では、後章での事例ならびに概念を整理する前提として、健康・医療領域におけるバイオデータ連携・利活用のための概念について整理する。具体的には、データ活用の前提となるデジタルトランスフォーメーション（以下、DX という。）の理解、健康・医療領域におけるデータ活用のフレームワーク、それらを踏まえた国内の現状と課題について整理する。

2.1.1 DX の概念と米国における電子カルテ普及政策からの学び

まず、コロナ禍以降重要性に関する声が飛躍的に増加した DX について解説する。誤解されがちであるが、DX は書類、データ、プロセス等などが単に「電子化」されることを指すものではない。デジタル技術を前提として(デジタル・ネイティブで)これまで以上の価値創造と提供ができるように、活動、仕組み、製品などを戦略的・構造的に転換することが DX の本質である[図 1]。

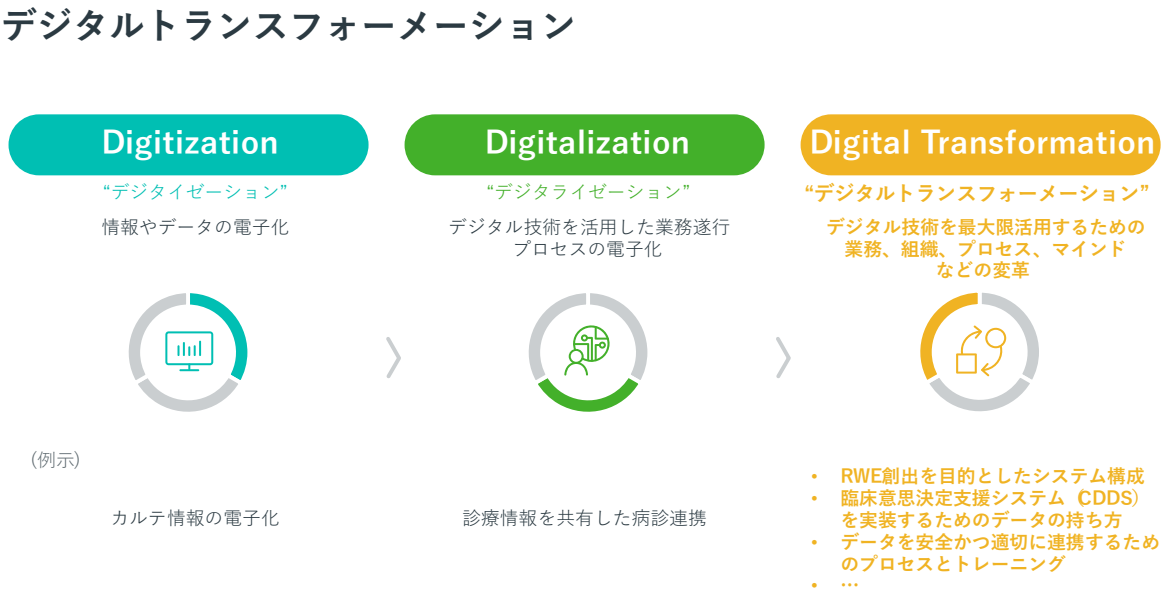


図 1.　デジタルトランスフォーメーション

地域医療情報連携を例に取れば、カルテ情報の電子化は Digitization（情報の電子化）、電子化されたカルテ情報とインターネットによる伝送を用いた病院、診療所間の情報連携は Digitalization（プロセスの電子化）、地域における病院・診療所・介護施設間のタイムリーな情報連携、患者・家族のスマートフォンへの自動的な診療情報保存（パーソナルヘルスレコード）、情報の統計解析による医療資源の最適化などを目的とした（デジタル・ネイティブな）システムの構築と運用、またそのための適切な体制整備、制度変更、インフラ整備などを推進していくことが DX である。

電子化はあくまで DX の前提条件でしかなく、DX の視点（電子化されたデータをどいかにして活用していくか）を持ちながら電子化を進めていくことが肝要である。ここでは米国における電子カルテシステムの浸透を学習の題材としたい(*)。米国における電子カルテ普及率は、病院全体で 85% 以上、200 床以上の規模においてはほぼ 100%とされる。令和 2 年度の国内における普及率は一般病院で 57.2%（400 床以上は 91.2%）と報告されている(**)。このように、現在では米国の方が電子カルテの普及率が高い状況ではあるが、2000 年代前半までは米国は電子カルテシステムの普及においては欧州と比較すると大きく遅れをとっていた。この流れを変える契機となったのが、2009 年に制定された HITECH 法 (The Health Information Technology for Economic and Clinical Health Act)である。米国政府は、HITECH 法の下に電子カルテシステムの導入奨励金として 170 億ドル規模の予算を投下し、普及を強力に後押ししてきた。これだけであれば“電子化”の推進政策ということになるが、バイオデータの連携・利活用（すなわち DX）の観点で特筆すべきは、電子化されたデータの活用推進に関する政策を合わせて進めてきた点である。米国政府は電子カルテ普及のインセンティブおよびペナルティ要件として、「意義ある使用 (Meaningful Use)」基準を設定し、運用してきた。2018 年以降は更なる EMR の導入/利用促進のため、Medicare からのインセンティブおよびペナルティ幅を拡大し、例えば電子カルテシステムからの情報提供を行った場合、2017 年以降の保険料を最大 4~9%増額する一方で、情報提供を行わなかった場合には 2017 年以降の保険料が最大 4~9%減額する規定 (MACRA 法: The Medicare Access and CHIP Reauthorization Act)を適用している。このように、データの連携・利活用を見すえたうえで戦略的な投資がなされていることはおおいに参考すべきであり、第 4 章紹介する諸外国の分散型データ連携ネットワークもまた、活用の出口を見据えたうえで大き

な戦略投資がなされていることを念頭に置きたい。さらに、米国の事例においては、実運用に関する検討を官民共同のコンソーシアムで進めてきた点も、データ活用において実効性の高い仕組みを構築する観点から重要なポイントと言える。

[参照]
*諸外国における医療情報の標準化動向調査：
https://www.mhlw.go.jp/content/10808000/000685914.pdf
**電子カルテシステム等の普及状況の推移：
https://www.mhlw.go.jp/content/10800000/000938782.pdf

2.1.2 健康・医療領域におけるデータ活用フレームワーク

続いて、健康・医療領域におけるデータ活用をフレームワークとして整理する。フレームワークはデータの種類と活用方法による 2 x 2 のマトリックスとして表される[図 2]。

- **データの種類**：「個人」のデータと「集団」のデータ
- **活用方法**：「個人」のための活用と「集団」のための活用

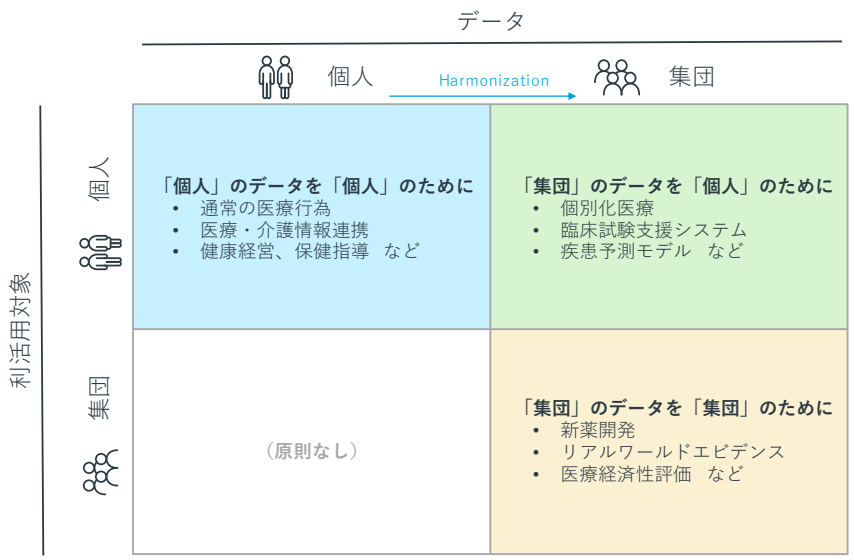


図 2. バイオデータと利活用の種類　概念図

データならび活用先が「個人」を想定したものであるか、「集団」を想定したものであるかは前提として大別すべき概念である。

「個人」単位の医療情報を「個人」のために活用するのは通常の医療行為そのものでもあり、個々の医療や健康のために行われるサービスもここに該当する。例えば、地域医療情報連携は「個人」のデータを効率的・効果的に連携・利活用することによって医療の質を高めるとともに医療リソースを最適化する取組であり、地方自治体や健康保険組合などによる保険指導などのサービスもこの領域に該当する。これらの領域において

は、そもそも個人情報の活用を前提とする領域となるため、サービス受益側はデータの利用同意（コンセントマネジメント）が、サービス提供側はデータの管理と効率的な活用（ダッシュボード、データマネジメントなどによるサービス提供者と適切な提供内容の特定）が一義的な関心事となる。この領域の海外事例は、この後の「2.2 海外におけるデータ連携・利活用の先進事例」の章にて、米国および英国の先進的な事例を取り上げ解説する。

「集団」のデータ、即ちビッグデータを「集団」のために活用している代表例が、新薬や新規治療の開発とリアルワールドにおける安全性検証、費用対効果などのリアルワールドエビデンス（以下、RWE という。）の構築である[図 3]。RWD 自体はデータであり、RWE となることによって解析結果に伴う示唆が生まれる。しかしながら、RWE 創出までの工程は決して平坦ではなく、むしろ解析の前工程に 8 割以上の工数と時間が必要となることが多い。この前工程をデータの Harmonization と呼び、この工程においては、データのクレンジング（記載内容の確認、記載項目との整合性、外れ値の除去など）、解析フォーマットへのマッピング、用語集の作成と統一などが含まれる。効率的な RWE 創出のためには、目的に応じた RWD があることは前提として、何よりもこの Harmonization にかかる労力を削減できるかが大きな要素である[図 XX]。RWE の事例は、この後の「2.3 ビジネス価値創出志向型事例」ならびに「2.4 学術価値創出志向型事例」の章にて具体的に解説する。

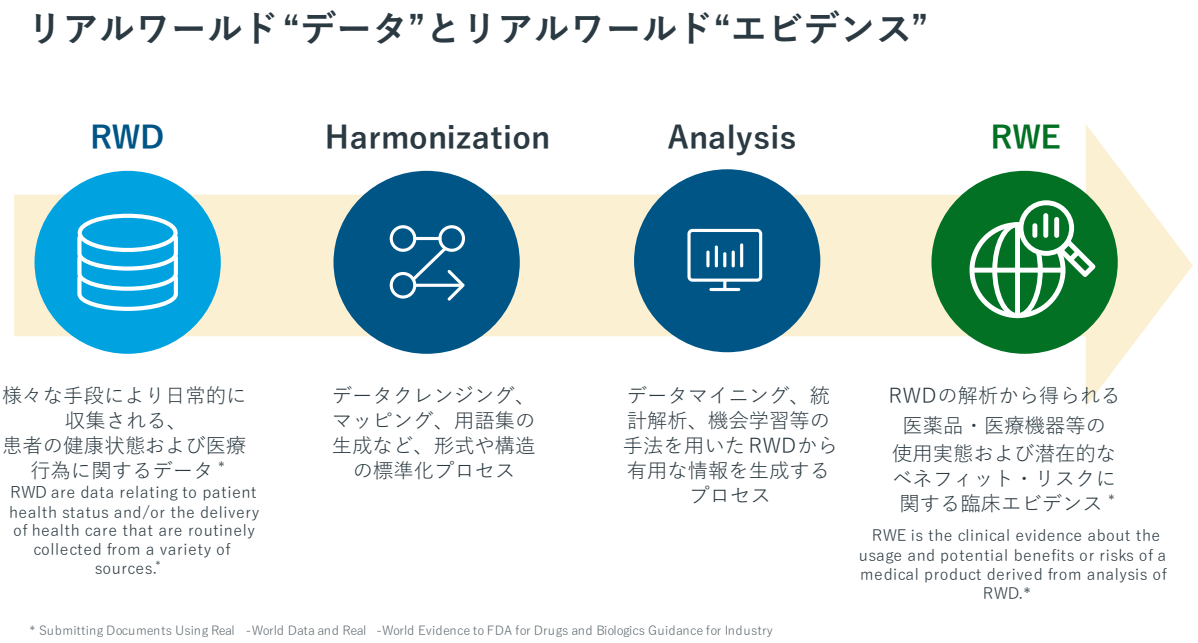


図 3. リアルワールド“データ”とリアルワールド“エビデンス”

そして「集団」のデータを「個」のために活用する最たる概念が個別化医療である。治験、臨床試験を通じたバイオマーカーに基づく医療・投薬の選択という意味の個別化医療も当然あるが、ここではより「個」に特化した活用として、統計解析モデルやアルゴリズムの適用をとりあげたい。健康診断結果、レセプトデータ、検査値などの RWD を用いた疾患や健康または医療行為における予測ならびに提案型のソリューションがこの数年間で加速度的に開発が進んでおり、特に臨床意思決定支援システム（Clinical Decision Support System: CDSS）は諸外国において急速に医療現場のオペレーションを変革しつつある。国内においても企業の健康経営や特定保健指導に現場において、RWD をベースとした疾患発症リスク予測モデルを用いた行動変容の取組など、「集団」の RWD を基にしたアルゴリズムを「個」への適応する取組が社会実装される事例が増えつつある。

2.2 海外におけるデータ連携・利活用の先進事例

前章で触れた「個人」単位の医療情報を個人の医療サービスの質の向上を目的に活用しているケースに関して、海外の先進例を紹介したい。

一つ目の事例は米国ニューヨーク州保険局がスポンサーとなり医療データのプラットフォームを構築・運用している Healthix の例である。Healthix は全米の自治体の中でも数多く構築・運用されている Regional Health Information Organization (RHIO)の中でも極めてデータ規模が大きく、サービス利用者にとっての利便性の高さでも成功している事例として知られる。最大の特徴は患者の医療連携ネットワークとしての機能が充実し、賛同するステークホルダーが得る恩恵が大きいことだと言える。海外有識者によると、データの研究目的での活用も進んでいる。

二つ目の事例では、英国の公的医療保険サービス NHS(National Health Service)が保有する地域医療ネットワークとしてロンドン市を中心とした地域で展開されている Coordinate my Care (CMC) について紹介する。CMC は、救急搬送または終末期のケアの為に、患者個人が事前に希望する内容を登録し、同意を得た医療提供者が閲覧できるというサービスである。普及の大きな要因としてかかりつけ医が患者のケアプラン登録を支援している事が大きく、結果として患者中心医療の提供だけではなく、医療提供者側の負担軽減という業務効率化にも影響を与えている。

2.2.1 Healthix, ニューヨーク州保険局の取り組み例

概要：

Healthix は、米国内最大の公衆衛生情報交換機関 Health Information Exchange (HIE) であり、ニューヨーク市、ロングアイランドを含むニューヨーク州南部地域を対象にサービスを提供している非営利法人で、ニューヨーク州保険局がスポンサーとなっている。Healthix では 5 つの RHIO が統合された巨大な医療連携ネットワークであり、2024 年 2 月時点、8,000 以上の医療施設から 2,000 万人以上の患者に関するデータを保有している(*)。

Healthix は、患者個人に対してより適切な医療サービスを提供することを第一の目的としている。各医療機関が患者ごとの治療の受診歴、処方歴、検査結果、アレルギー情報などを Healthix に提供し、患者情報がひとつのプラットフォームで確認できるようになることで、救急時の適格な診断とより良い医療サービスの提供に役立てられる仕組みとなっている。

米国ではニューヨーク州の例以外にも、各州や地域単位で Regional Health Information Organization (RHIO)の検討が進み複数実行されている。背景には、複数の医療情報システムが乱立することの弊害として、システムを跨いでの患者の医療情報の共有や開示が阻害される事で、適切な医療サービスの提供を困難としていた長い歴史がある。

利用方法：

患者登録は本人の同意を必須とし、医療機関を含めた登録機関では患者に対し同意を拒否する方法も説明することを義務付けられている。医療機関側では患者の個人データを閲覧する為には患者の同意を得る必要がある。

Healthix に加盟している医療機関（病院、介護施設、開業医、訪問ケアなど）から医療サービスを受けた際、病態、投薬歴、検査結果や治療歴などを含む個人の医療情報は自動的に更新される。医療機関がこのような患者の情報にアクセスできる事により、治療歴を完全に把握した上でより適切な医療サービスが提供できる。

Patient Notice（患者通知）- 施設側では、Healthix に加盟している事を患者に通知する事が義務付けられている。また患者通知は、保護された個人の健康情報が Healthix にアップロードされた事を患者本人に通知される。施設からは同意の拒否についても方法が説明される。

医療施設を含む加盟機関については、ウェブサイト上の　participant list のページで公開され、全ての機関の住所、ウェブサイト情報などに患者がアクセスできるようになっている。リストには病院、救急医療サービスなど医療機関のみならず、薬局、在宅ケア、地域支援機関や検査ラボまで含まれている。医療機関の種類とデータ提供範囲の 2 つの条件から検索が可能で、検索結果のリストでは各施設が提供している患者データの種類が確認できる。

図 1. Healthix 参加機関検索結果

Health Systems and Hospitals

Facility	Parent	Patient Name	Gender	Date of Birth	Race	Ethnicity	Preferred Language	Smoking Status	Vital Signs	Allergies	Medications	Problem List	Lab Tests	Lab Results	Care Plans	Procedures	Care Team	Immunizations
Brookdale Hospital - FS FS000001286	ONEBKYH	●	●	●	●	●	●	●	●	●	●	●	●	●	x	●	●	●
Calvary Homecare Hospice FS000007805	CALVARY	●	●	●	x	x	x	x	x	x	x	x	x	x	x	x	●	x
Calvary Hospital - Bronx FS000001175	CALVARY	●	●	●	x	x	x	x	x	x	x	x	x	x	x	x	●	x
Calvary Hospital - Brooklyn FS000010223	CALVARY	●	●	●	x	x	x	x	x	x	x	x	x	x	x	x	●	x
David H. Koch Center For Cancer Care at MSK FS000010355	MSK	●	●	●	●	●	●	●	●	●	●	●	●	●	x	●	●	●
Flushing Hospital - Parsons Ave FS000001628	MEDISYS	●	●	●	●	●	●	x	●	●	●	●	●	●	x	x	●	●
GLEN COVE HOSPITAL at 101 St Andrews Lane FS000000490	NORTHWELL	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Good Samaritan Hospital FS000000925	CHSLI	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Good Shepherd Hospice FS000004473	CHSLI	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Harlem Hospital Center at 506 Lenox Avenue FS000001445	NYC H + H	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

図 2. Healthix 参加医療機関ごとのデータ提供項目

COVID-19 パンデミック禍での貢献：

Healthix が市民の医療サービス向上に大きく貢献した事例として、COVID-19 のパンデミックの中、救急搬送された患者の情報を医療従事者側で容易に確認できたことが挙げられる。2020 年、緊急事態宣言中の医療施設での口頭および書面による同意について、NY 週保険局が患者同意書の免除を許可し、COVID-19 検査結果の通知を医療従事者に用意に送信できるようになった。

戦略的パートナーシップ事例　*My Health Record NY*：

Healthix のもう一つの特徴として、患者側のユーザビリティ向上の為に様々なパートナー企業と戦略的に提携を重ねていることが挙げられる。2023 年 12 月に認証システム企業 CLEAR 社との提携によって提供開始された *My Health Record NY*　のサービスがある。

My Health Record NY　では CLEAR 社が保有する本人認証技術によりシームレスな本人確認が可能となり、健康情報への安全に、かつ、簡単にアクセスが可能となる。NY 州全体の医療システムと医療機関感の相互運用が実現し、患者は顔写真（自撮り写真）で本人確認を行うだけで自分の医療データにアクセスし、集約された記録を閲覧で

きるようになった。医療従事者はこれまで時間を費やし手作業で確認していたプロセスから解放され、より患者ケアに時間を割くことができる。

　My Health record NY では、初めてのユーザーは政府発行の ID と自撮り写真を使用し、CLEAR で本人確認が必要。CLEAR 再登録メンバーも同様の手続きを要求される。

[参照]

*https://healthix.org/what-we-do/healthix-services/

2.2.2 Coordinate My Care (CMC)

概要

Coordinate My Care (CMC) は、英国の公的医療サービスである NHS（National Health Service）が提供するサービスであり、患者の救急搬送時に誰もが同じプラットフォームから患者情報へアクセスできる仕組みである。医療提供者は、患者が事前に登録した情報にアクセスすることによって、緊急時にどのような治療を望み、誰に連絡をしてほしいかを確認することができる。地域としてはロンドン市内と周辺地域を対象に提供されている。

CMC の起源は 2007 年に設立された非営利活動法人で、終末期医療の専門病院であるロイヤル・マースデン病院の医師ジュリア・リリー氏が中心となり、“緊急時に誰もがアクセス可能なターミナルケアの標準化”を理念に、救急ケアや関連するガン患者に対する終末治療を包括的に提供する取り組みを開始した。救急医療時の課題として患者についての重要な情報が不足していること、患者自身にとっても準備が足りていないことなお、医療提供の体制に改善の余地があった。現在は NHS がサービス提供母体となり、患者の意志に沿った治療を施すことを目標とするサービスを関連する医療チームがケアプランを共有している。

CMC では、かかりつけ医（GP：General Practitioner）のみならず、救急サービス、NHS111（救急電話）、医療機関、時間外 GP 診療所、療養所、ホスピス、ソーシャルワーカーなどと連携し、以下の患者個人情報が共有される。

- 患者基本情報（生年月日、性別、住所、診察治療記録、医療関係者の連絡先など）
- ケアプラン（治療時の希望など）
- 家族や知人との関係性／人間関係（家族の連絡など）
- 本人の意思（治療方針、看取りの場所、サービス内容、医療費の範囲、延命措置の必要性など）

現在、CMC に登録した患者の数は公表されていないものの、2021 年に公開された活動報告レポート（*）によると患者のケアプランの登録件数は 2010 年のサービス開始以降 14 万件超、そのうち 2019 年以降に MyCMC（CMC に自分のケアプランの要望を

登録するサービス）で登録された件数は約 3,400 件に上る。2021 年 10 月単月データでは、約 1,600 件のケアプラン登録されており、緊急医療提供時に閲覧された登録済プランは約 8,800 である。10 月単月での新規登録プラン数は約 450 件。患者側は自ら CMC に希望するケアプランを登録し、いつでも変更することが可能であり、また患者によって同意を確認できた医療関係者がそれを閲覧できる。

2021 年直近 1 年間のデータでは、CMC のケアプランの 5 割はかかりつけ医が登録を主導している。これは英国ではかかりつけ医制度（NHS に登録することで、無料でかかりつけ医からの医療サービスを受け、必要に応じ専門病院を紹介され診察を受ける仕組み）を採用しているためで、患者個人の意思とはいえ、プラン登録にはかかりつけ医の影響・役割が大きいことが分かる。なお同データによると、終末期の患者が登録した「最期を迎える場所」については、登録患者の 78%は自分が要望していた場所で最期を迎えることが出来たと報告されている。

このように CMC は国の医療行政の一部を担うサービスとなっている。活用成功の要因としては、英国では医療サービスの特徴としてかかりつけ医（GP）に力を入れていることは知られている通りであるが、既に GP 間での電子カルテの普及やその連携が発達しており、国民や社会に活用するという議論が進んでいるということが CMC 普及を推進したと考えられる。

二つの海外事例から地域医療ネットワークの成功要因を考察する。

- **ネットワーク活用による経済性向上：**ニューヨーク州 Healthix では患者情報の共有、医療提供者側の業務効率改善から、地域の疾病の予測や重症化予防まで発展しており、ネットワークの利活用による経済性が向上していること。
- **医療サービス向上の可視化：**Healthix では、Covid-19 拡大期のような有事の際に個人の医療情報へのアクセスを簡素化した例のように、医療サービス利用上の恩恵が明確に可視化されていること。
- **チームによる患者情報共有を可能とするシステム：**ロンドンの CMC では、患者の同意を得た医療提供者が、同一のネットワークにアクセスすることが可能となっていること。
- **かかりつけ医の主導：**CMC という地域医療情報ネットワークはかかりつけ医（GP）との連携の上に実現していること。

[参照]

* CMC 活動報告レポート：

https://createsend.com/t/j-FAA339C99577404E2540EF23F30FEDED

2.2.3. 日本国内での地域医療連携ネットワークの課題

一方、日本国内においても 2000 年代後半から地域医療連携への取り組みは促進されているものの、未だに地域連携ネットワークの運用そのものについての課題が多く存在する。(*) 地域医療介護総合確保基金及び地域医療再生基金を活用して構築したネットワークのシステムについて運用がない、または低調であるという報告もある。また日本医師会総合政策研究機構のワーキングペーパー（**）から具体的な課題点を探ると、自次世代医療基盤法への認知が低い、HL7 FHIR の理解浸透が浅い、など、様々な点で医療 DX への推進に大きな障壁があることがうかがえる。

日本では地域医療ネットワークのデータプラットフォーム構築において米国や英国と比較して遅れを取っている。ただ、厚生労働省の「医療 DX 令和ビジョン 2030」(***) では全国医療情報プラットフォームの構築を進めており、具体的には医療機関同士などでのスムーズなデータ交換や共有を推進するため、HL7 FHIR を交換規格とし、交換する標準的なデータの項目及び電子的な仕様を定め、それらの仕様を国として標準規格化を進めるとしている。標準化の推進により地域医療ネットワーク活用推進が期待される。

[参照]

*厚生労働省，地域医療情報連携ネットワークの現状について
https://www.mhlw.go.jp/content/10800000/000683765.pdf
**日医総研ワーキングペーパーICT を利用した全国地域医療情報 連携ネットワークの概況（2022 年度版）
https://www.jmari.med.or.jp/wp-content/uploads/2023/09/WP475.pdf
***厚生労働省，第 1 回「医療 DX 令和ビジョン 2030」厚生労働省推進チーム資料について
https://www.mhlw.go.jp/stf/newpage_28128.html

2.3 ビジネス価値創出志向型事例

ヘルスケア関連情報データベースにビジネスとしての価値を見出し、最も活用しているのは製薬企業といえるだろう。製薬企業では、アンメットメディカルニーズを考慮した医薬品の創薬や開発にデータベース情報を活用しているだけではなく、臨床研究開発や承認申請時における、上市後の市販後調査による安全性の確保など、医薬品開発から販売後までのあらゆるビジネスステージにおいて活用している。これらの活用実態の中から、最も活用頻度の高い、上市後の市販後調査の事例と承認申請の活用事例について紹介する。さらに、近年米国を中心に近年注目が集まっている患者団体と協働したデータベース活用実態についても概要を紹介する。

2.3.1　市販後調査への活用事例

製薬企業が自身のビジネスとして最も活用しているのが、市販後調査である。医薬品の上市後も、市販後調査を実施し、流通医薬品の安全性を確保することが求められる。しかしながら、従来のサンプル調査では、発現率の低い副作用をとらえるには n 数が不十分であるという課題を抱えていた。そこで、米国 FDA は 2009 年にパイロットプロジェクトとして Sentinel プロジェクト（詳細は 3.2 RWD における Common Data Model (CDM)を参照）を開始し、発現率の低い稀な副作用を大規模データベースによって検知しようという世界初の取り組みが行われた。同プロジェクトは年を追うごとに拡充され、蓄積されたデータベースを活用した安全性評価の結果が、従来の研究調査と同等であるという研究成果も論文化(*)もされた。他にも、Sentinel の導入により「センチネルシステムに由来するリアルワールドデータ及びエビデンス（RWD/RWE）の使用により、5 種類の医薬品に対する 9 つの潜在的な安全上の問題に関して市販後調査の必要性がなくなり、市販後の安全性評価がより迅速かつ効果的となった」という報告(**)もある。プロジェクト開始から 10 年以上経過した現在では、市販後調査における安全性評価の一般的手法として広く活用されている。

日本国内でも平成 22 年（2010 年）から、米国を参考にデータベースを活用することで医薬品の安全性を高める検討が開始された。日本における「市販直後調査」とは，新医薬品の販売開始後（効能・効果の追加時等は承認後） 6 ヶ月間、診療において当該医薬品の適正使用を促し，必要な副作用等に関する情報を迅速に把握するために、医薬品リスク管理として製造販売業者が実施するもので、「医薬品，医薬部外品，化粧品及び医療機器の製造販売後安全管理の基準に関する省令（平成 16 年厚生労働省令第 135 号）」にて定められた調査である。新医薬品の承認審査の過程において市販直後調査が必要であると判断された場合、個々の医薬品の承認の条件として付される。市販直後調査は，使用成績調査，製造販売後臨床試験等のように，症例の登録を行い、予め決められた事項について実施する調査ではなく、適正使用情報の収集・提供等の活動の一環として行われる調査のため、いくつかの限界を抱えていた。例えば、すべての副作用が自発的に報告されるわけではいとう報告バイアスの限界や、医薬品を投与された分母がわからないために副作用の発現率やリスクが算出できないという限界、さらに製薬企業側

としても、稀な副作用リスクを特定するための全例調査も含めた情報収集は実施負担が大きいという課題も抱えていた。

そこで日本の当局である PMDA（医薬品医療機器総合機構）が中心となり MID-NET という医療情報データベースを構築した。MID-NET(***)は、国内のいくつかの医療機関が保有する電子カルテ（オーダリング、検査結果等を含む）、レセプト（保険診療の請求明細書）及び DPC（入院費用の包括評価制度）の各種データの利用が可能な医療情報データベースで、データの信頼性確保するために、GPSP に準拠するよう管理運営されている。2024 年 3 月現在の MID-NET の利活用状況(****)は、製薬会社など製造販売会社によるものが 15 件、PMDA によるものが 23 件、その他大学などのアカデミアによるものが 2 件の合計 40 件に留まっている。製造販売会社による利活用 15 件のうち 14 件は製造販売後調査として活用されており、適応拡大など新たな承認事例獲得等への利活用が課題となっている。

[参照]
*米国 Sentinel データベース活用結果と従来調査の結果の同等性に関する論文
Gagne J. J. et al., Clin. Pharmacol. Ther., 100, 558-564(2016).

**米国における市販後調査が不要になった事例
https://www.jpma.or.jp/information/evaluation/results/allotment/lofurc0000005ln4-att/fdas_rwep.pdf#page=103

***MID-NET に関する情報
https://www.pmda.go.jp/safety/mid-net/0001.html

****MID-NET の利活用状況
https://www.pmda.go.jp/safety/mid-net/0010.html

2.3.2　承認申請への活用事例

米国では市販後調査だけではなく承認申請にも医療データベースを活用している。2022 年 9 月に日本製薬工業協会がまとめた「製薬企業における健康医療データの利活用に関する期待と課題」(*)調査報告書によると、製薬企業側では、市販後調査である PMS においては医療データベースを活用できているが、研究開発といったエリアでは利活用が進んでおらず、研究開発にも活用できるデータ基盤の構築と環境整備の必要性を訴えている。

図 1.　製薬企業におけるデータ利活用状況と課題

一方、米国では市販後調査を超えた、研究開発における活用も進んでおり、同報告書の中では米国におけるパルボシクリブの男性乳がん適応追加について、臨床試験の代替として医療データベースを評価した事例が紹介されている。

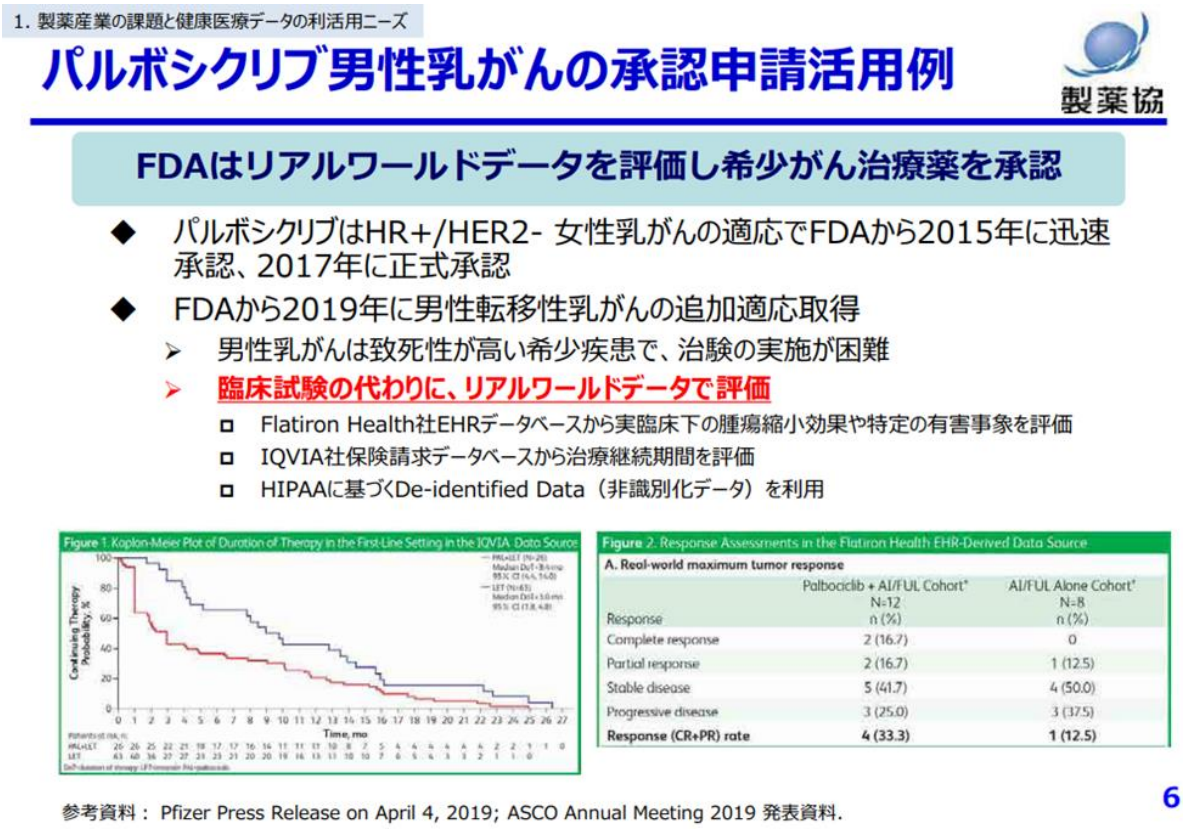


図 2. 市販後調査を超えた研究開発への活用事例

この他にも米国の事例として、アベルマブのメルケル細胞がん承認申請事例において、医療データベースを外部対照群として利用し評価された事例やエヌトレクチニブの ROS1 陽性非小細胞肺がん申請において外部対照群として利活用に挑戦した事例が紹介されている。このように、臨床症例を集めることそのものが難しいような稀少事例において、医療データベースの利活用が期待されている。今後は日本においても市販後調査を超えた領域への展開を期待したい。

[参照]

*「製薬企業における健康医療データの利活用に関する期待と課題」調査報告書
https://www8.cao.go.jp/kisei-kaikaku/kisei/meeting/wg/2201_03medical/220922/medical09_0102.pdf

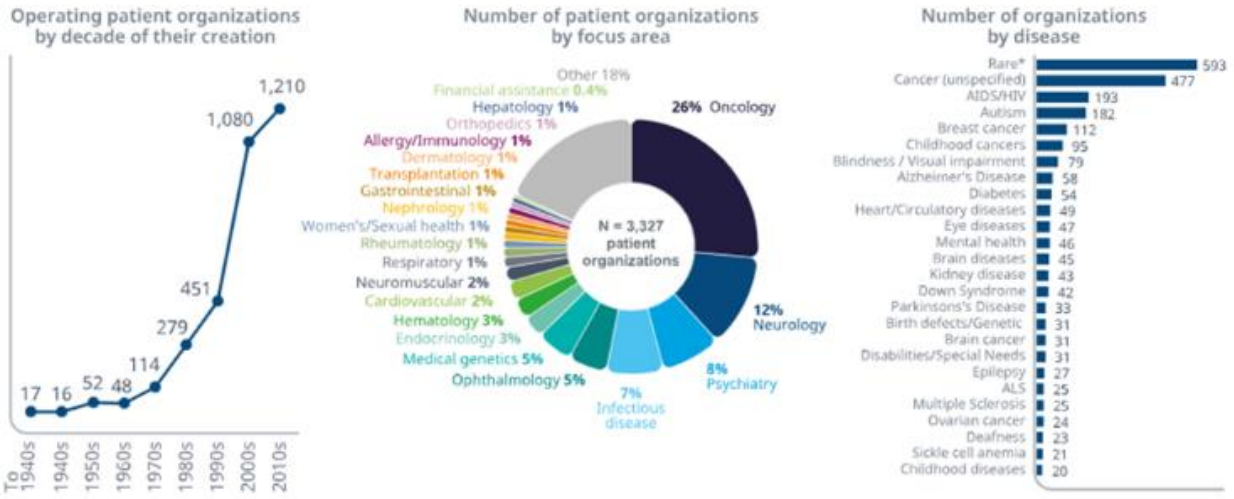
2.3.3 患者団体との共同事例

ビジネス価値創出志向型の事例として最後に、患者団体との協働した取り組みを紹介する。米国における活動的な患者団体の数は 3000 以上にも上っており、患者の健康サポートだけではなく、製薬企業や医療機関の研究にも協力を行っている。患者団体は病気をより深く理解するための患者登録の開発や治療法を見つけるための資金提供など、その役割が近年拡大を続けている。過去 15 年間で患者団体とライフサイエンス企業との間での取引件数は公表されているものだけでも 700 件にも及び、総額 24 億ドルの取引があったと推定される(*)。また患者団体そのものの総収益も、過去 5 年間で 625 億ドルを超えており、助成金やコミュニティプログラムなどの支援に活用されている。

患者団体には様々なタイプがあるが、今回は、ライフサイエンス企業との連携で成功した事例を紹介するとともに、今後の改善に向けた課題についても言及する。

患者団体は、オンコロジーや小児疾患や先生性疾患に重きを置いた活動を行っている団体が多い。米国における 3327 の患者団体のうち、18%にあたる 593 の団体が稀少疾患に重点をおいている。現在運営されている患者団体の 3 分の 1 は過去 10 年間に設立されたばかりのもので、近年その動きが活発化している。この背景には、2000 年代以降に患者団体が自分の病気のことやその治療法、また医薬品のベネフィットとリスクおよび副作用についてインターネットを積極的に活用し始めたことに起因している。加えて 2010 年代以降はソーシャルメディアプラットフォームの発展も相まって、ブログや患者コミュニティで自らの体験を他の人と共有する活動が活発になった。その結果、医療従事者や企業などの関係者から、ライブ感のある新情報として生の患者体験の重要性が重要視されるようになってきた。

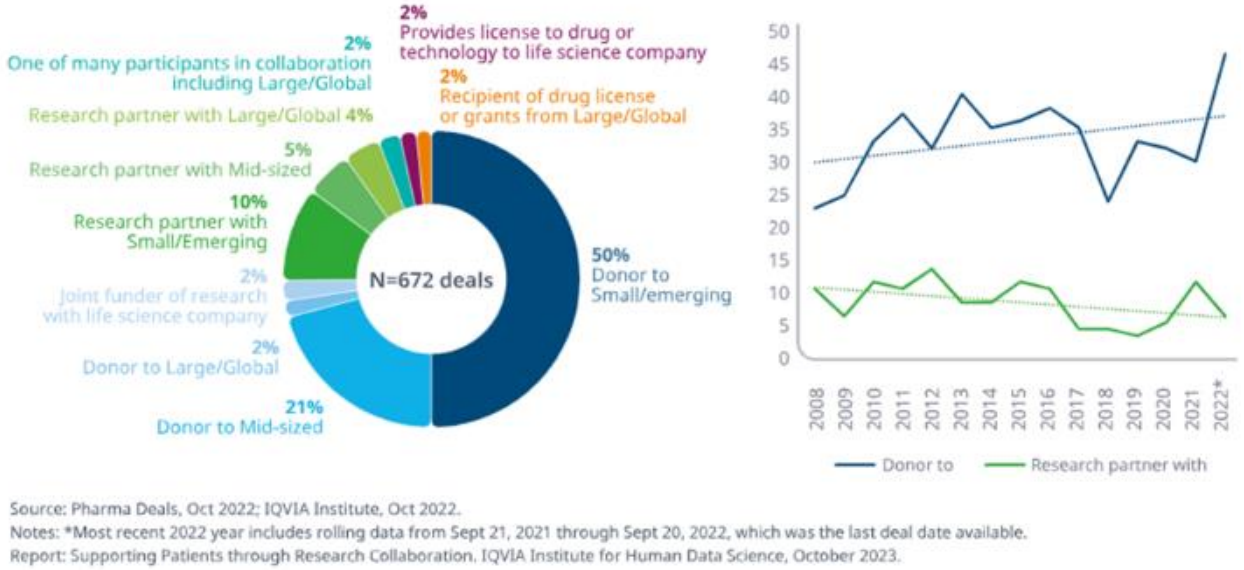
患者単体の全取引の 4 分の 3 が企業による助成金や知的財産の提供などで占められており、共同研究は現在減少傾向にある。



Source: Form 990 data from IRS Statistics of Income (SOI) program, years 2015–2021.
Notes: Left chart: The creation of the organization was determined using the ruling date, which is the date the organization received its tax-exempt status from the IRS. Select dates were overridden based on website information for those with ruling dates earlier than 1950. Charts includes any patient organization filing 2015–2021. ALS = amyotrophic lateral sclerosis.
Report: Supporting Patients through Research Collaboration. IQVIA Institute for Human Data Science, October 2023.

図 3. 米国における患者団体の実態と注力疾患領域

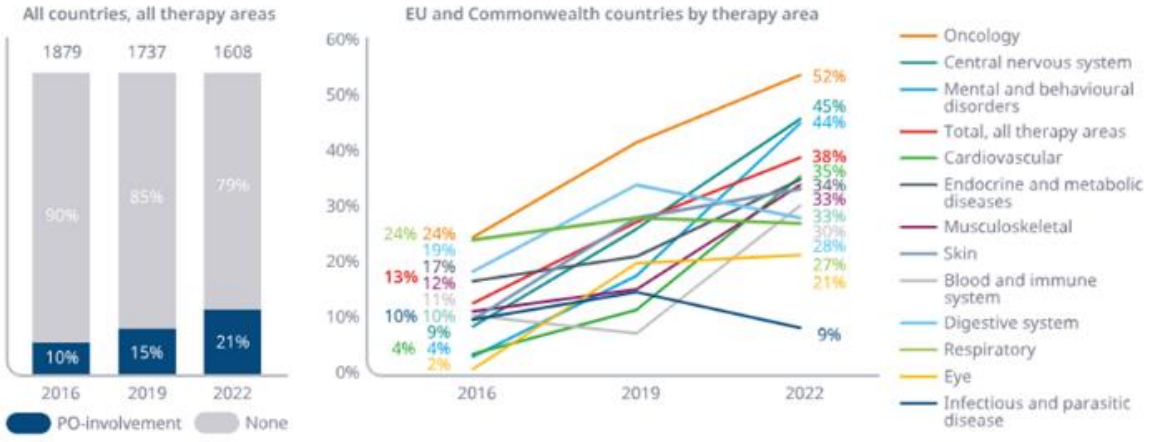
特に患者団体の資金源における企業との取引のうち 3 分の 2 は、中小企業やベンチャー企業によるものである。患者団体は、アンメットニーズとなっている疾患分野における複雑な研究ツールに関して有用な支援をしており、研究開発初期におけるトランスレーショナル研究を迅速化するために、「ツールボックス」というものを構築し、アッセイライブラリーや生体サンプル、バイオマーカーなども共有している。これらの患者団体により構築されたデータが企業における細胞レベルや動物研究レベル、遺伝子変異解析などに活用されている。



Source: Pharma Deals, Oct 2022; IQVIA Institute, Oct 2022.
Notes: *Most recent 2022 year includes rolling data from Sept 21, 2021 through Sept 20, 2022, which was the last deal date available.
Report: Supporting Patients through Research Collaboration. IQVIA Institute for Human Data Science, October 2023.

図 4. 患者団体とライフサイエンス企業の企業規模別取引

また、患者団体は、全世界の技術評価 HTA の 21%の情報も提供しており、EU やイギリスなどの一部の地域では最大 52%も寄与している。イギリスには、従来から新しい治療法に対して、費用の面からその妥当性を評価する医療技術評価 HTA の仕組みがある。その中の判断項目の一つとして、新薬の服用が生活の質や機能にどう影響したのか、またアウトカムや寿命を改善したのかという定性的な評価点について、患者や介護者の視点を取り入れている。NICE、CADTH、SMC などの機関は、ほぼすべての HTA 申請に患者の意見を提出することを求めている。



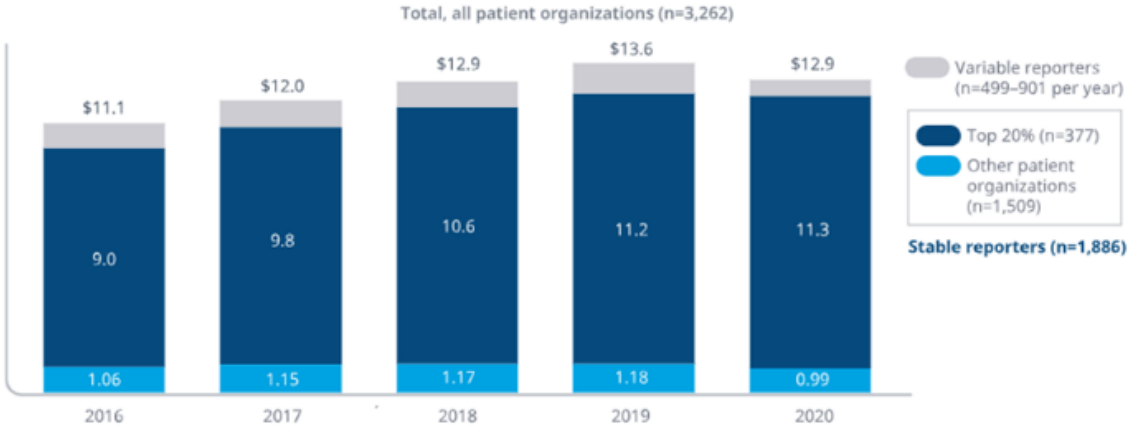
Source: IQVIA HTA Accelerator, Aug 2023.

Notes: Includes single drug and multiple drug assessments only. Excludes non-submissions, new formulations, suspended and cancelled assessments. Includes completed, published and draft assessments. EU and Commonwealth agencies included her are CADTH, Canadian Agency for Drugs and Technologies in Health; G-BA, Federal Joint Committee (Gemeinsamer Bundesausschuss); HAS, French National Authority for Health (Haute Autorite de sante); HTA = Health technology assessment; INESSS and pCODR in Canada; IQWiG Germany; PBAC in Australia; NICE, National Institute for Health and Care Excellence; SMC, Scottish Medicines Consortium; UK, United Kingdom; ZIN, Zorginstituut Nederland.

Report: Supporting Patients through Research Collaboration, IQVIA Institute for Human Data Science, October 2023.

図 5.　患者団体とライフサイエンス企業の企業規模別取引

患者団体の収益に注目してみると、過去 5 年間の患者団体の収益は 625 億ドルを超えている。



Source: Form 990 data from IRS Statistics of Income (SOI) program, years filed 2016–2021.

Notes: Stable reporter data includes only patient organizations filing continuously through the period 2016–2020 (n=1,886). Top 20% are stable reporters with the top 20% of revenue in 2020, i.e., $2.69 million and above. Variable reporters number vary by year from n=499 to n=901. * Five-year growth is calculated over 2015 values (not shown).

Report: Supporting Patients through Research Collaboration, IQVIA Institute for Human Data Science, October 2023.

図 6.　患者団体の収益構造

患者団体は患者コミュニティに既にベネフィットをもたらしているが、今後は、新しい治療法の探索加速のために最新のテクノロジーを活用することが求められている。医療そのものはデータやテクノロジーとの結びつきがより一層深まっていくのに対して、患者団体も患者中心の医療を進めるためのよりよい機会を活用しようとしている。患者団体にとっては、レジストリや医療データを産業界と協力してプラットフォームを構築することで、データと患者の実体験をつなぐ新しい発見やこれまでにないインサイトをもたらしてきている。患者団体に加え、研究者、ライフサイエンス企業、規制当局のそれぞれのステークホルダーが一体となってデータを活用することで、より画期的かつ革新的な治療法の探索へとつながっていくだろう。

Source: IQVIA Institute for Human Data Science. Empowering patient-driven research to improve patient outcomes: the essential role of patient organizations – key themes and takeaways from the IQVIA Institute Patient Advocacy Summit held on December 1, 2021. Feb 2022. Report: Supporting Patients through Research Collaboration. IQVIA Institute for Human Data Science, October 2023.

図 7. 患者団体の今後の可能性

[参照]

*IQVIA INSTITUTE “Supporting Patients through Research Collaboration INTERACTIONS BETWEEN PATIENT ORGANIZATIONS AND LIFE SCIENCES COMPANIES”
https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/supporting-patients-through-research-collaboration/iqvia-institute---supporting-patients-through-research-collaboration-10-23-forweb.pdf

2.4 学術価値創出志向型事例

学術的価値創出事例として最も代表的かつ大規模なものとして、LEGEND(Large-scale Evidence Generation and Evaluation across a Network of Databases)イニシアティブがあげられる。LEGEND イニシアティブでは OHDSI（The Observational Health Data Sciences and Informatics 略称オデッセイ）という国際データベースネットワーク内を活用し、同データベース内にある数億件にも及ぶ患者記録をベースにハイレベルな観察研究を行っている。代表例としては、高血圧や糖尿病、うつ病や COVID-19 の治療効果などの研究が展開されている。それらの研究成果は、Lancet や JAMA などの有力紙に掲載され、治療時の意思決定エビデンスとして活用されている。

OHDSI について簡単に紹介する（詳細は、次章 3. リアルワールドデータ(RWD)の利活用に向けた取組の中の 3.3 OMOP CDM と OHDSI を参照されたい）。OHDSI は様々なステークホルダーが関与した学際的なコラボレーションプログラムで、すべてオープンソースで公開されている。コロンビア大学に中央調整センターが設置されている。研究者と観察健康データベースの国際ネットワークは全世界に広がっており、最も活用度の高い研究用共同ネットワークとなっている。OHDSI のミッションは、よりよい健康上の意思決定とケアを促進するエビデンスをコミュニティが協力して生成できるようにすることで、同組織のビジョンは観察研究により健康と疾病を包括的に理解することを掲げている。OHDSI の下には様々な疾患単位、地域単位のサブグループが設定されており、巨大データベースを研究者や臨床家が使いやすいような組織構造が整えられている。

以下に OHDSI を活用した代表的な LEGEND イニシアティブ研究(*)を 3 つ（高血圧・糖尿病・COVID-19）紹介する。

2.4.1 高血圧治療薬に関するランダム化研究

まずは大規模データベースを活用した研究として最も有名な高血圧治療薬に関する論文を紹介する。この研究成果は医学的に評価の高い LANCET(**)と JAMA(***)に掲載された。


Articles

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, Ruijun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, Patrick B Ryan

Summary

Background Uncertainty remains about the optimal monotherapy for hypertension, with current guidelines recommending any primary agent among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium channel blockers, in the absence of comorbid indications. Randomised trials have not further refined this choice.

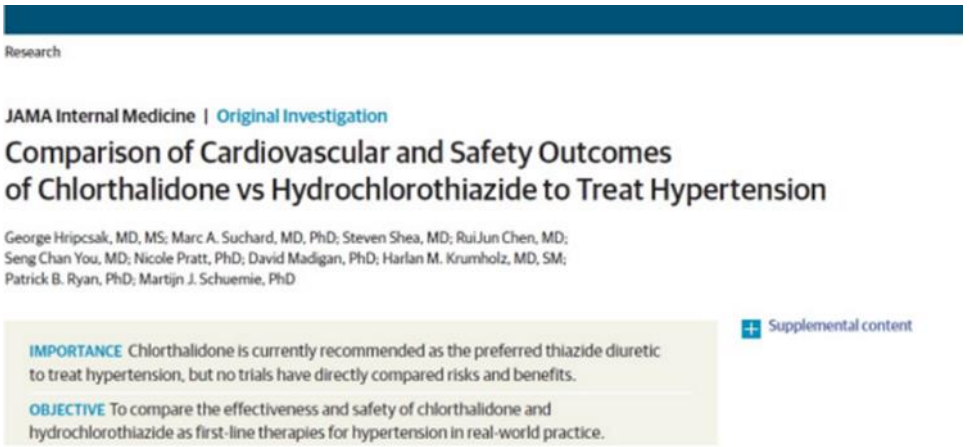
Published Online October 24, 2019 https://doi.org/10.1016/S0140-6736(19)32317-7

図 1. LANCET 掲載論文

LANCET に掲載された論文(**)は Marc A Suchard らによる” Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis”「 第一選択の降圧薬クラスの有効性と安全性の包括的な比較：体系的、多国間、大規模分析」である。高血圧症に対する単剤療法における第一選択薬の有効性と安全性を評価するために、4900 万人分の患者データが活用された。データベースとしては 6 つの請求系データベースと 3 つの電子医療記録データベースを組み合わせて実施された。主要なアウトカム指標として急性心筋梗塞、心不全による入院、脳卒中の 3 つを設定し、他にも二次的な有効性 6 項目、安全性アウトカムに関する 46 項目について相対リスクが比較された。この結果、ほとんどの項目でクラス間の差異は見られなかったが、サイアジドまたはサイアジド様利尿薬は、アンジオテンシン変換酵素阻害薬よりも優れた一次有効性と安全性を示したことが明ら

かにされた。これにより、高血圧症の単剤療法を開始するための薬剤クラス間として最適なガイドラインが大規模データで検証されることにつながった。

JAMA に掲載された論文(***)は、George Hripcsak らによる“Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension”「高血圧治療におけるクロルタリドンとヒドロクロロチアジドの心血管および安全性の結果の比較」である。



Research

JAMA Internal Medicine | Original Investigation

Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension

George Hripcsak, MD, MS; Marc A. Suchard, MD, PhD; Steven Shea, MD; RuiJun Chen, MD; Seng Chan You, MD; Nicole Pratt, PhD; David Madigan, PhD; Harlan M. Krumholz, MD, SM; Patrick B. Ryan, PhD; Martijn J. Schuemie, PhD

Supplemental content

IMPORTANCE Chlorthalidone is currently recommended as the preferred thiazide diuretic to treat hypertension, but no trials have directly compared risks and benefits.

OBJECTIVE To compare the effectiveness and safety of chlorthalidone and hydrochlorothiazide as first-line therapies for hypertension in real-world practice.

図 2. JAMA 掲載論文

高血圧治療薬クロルタリドンは、高血圧治療に推奨されるサイアザイド系利尿薬として推奨されているが、リスクと利益を直接比較した試験はなかったことから、実際の臨床現場での高血圧の第一選択療法としてのクロルタリドンとヒドロクロロチアジドの有効性と安全性を比較した研究である。データベースとしては米国内の 2 つの請求系データと 1 つの電子医療記録データが用いて、2001 年～2018 年までの降圧薬単独療法の初回使用者の外来および入院治療エピソードデータがある患者 73 万人のデータが解析に用いられた。主要アウトカムは LANCET 同じ急性心筋梗塞、心不全による入院、脳卒中、に加えて心臓突然死を含む複合心血管疾患アウトカムも評価対象とされ、安全性に関しては 51 項目が評価された。この結果、クロルタリドンの使用はヒドロクロロチアジドと比較した場合、心血管への有意なベネフィットは見られなかったが、腎臓および電解質異常のリスクの増加と関連していることが明らかとなった。これにより、初めて

使用する患者の高血圧治療にはヒドロクロロチアジドよりクロルタリドンを優先するという現在のガイドライン推奨を支持しないことが明らかになり、大規模データに基づいてガイドラインを検証するという取り組みができるようになった。

2.4.2 糖尿病治療薬治療プロトコルに関する研究

次に LEGEND イニシアティブで行われた糖尿病治療薬におけるセカンドラインプロトコルに関する研究事例(****)を紹介する。Rohan Khera らによって発表された”Comparative Effectiveness of Second-line Antihyperglycemic Agents for Cardiovascular Outcomes: A Large-scale, Multinational, Federated Analysis of the LEGEND-T2DM Study”「 第二選択の血糖降下薬の心血管疾患に対する有効性の比較：LEGEND-T2DM 研究の大規模な多国間連合分析」では、SGLT2 阻害剤 と GLP-1 が、2 型糖尿病患者における重大な心血管イベントを軽減するとされているが、それらの相互作用や他の二次血糖降下薬との相対的な有効性については明らかにされていなかった点に注目して行われた。1992 年から 2021 年にわたる約 30 年間分のデータを用いて、約 150 万人に近い糖尿病患者データを対象に、心筋梗塞、脳卒中、死亡の主要アウトカム指標とそれらに心不全入院を加えた 4 つのリスク指標が解析された。この結果、SGLT 阻害剤および GLP1 により同等の心血管リスクの低減が見出され、両薬剤は DPP4 阻害剤よりも効果的であることが明らかとなった。これにより、心血管イベントリスクを抱える 2 型糖尿病患者の第二選択薬には SGLT2 阻害剤と GLP1 が優先されるべきであることが示唆された。

2.4.3 COVID-19 関連に関する研究

OHDSI の Web サイト内(*****)には、COVID-19 の最新情報に関する専用ページが設けられており、OHDSI を活用して行われた COVID-19 関連の論文リンクが一覧となって掲載されている。2024 年 3 月現在、公表掲載された論文が 12 件、掲載前の論文が 7 件紹介されている。

最も古い公表論文は、COVID-19 が本格流行した年に当たる 2020 年の 8 月に、LANCET リウマチに掲載されている。Jennifer C E Lane らによって公表された論文タイトルは”Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study” 「関節リウマチの治療におけるヒドロキシクロロキン単独およびアジスロマイシンとの併用のリスク：多国籍後ろ向き研究」で、概要について簡単に紹介する。

関節リウマチの治療に一般的に使用されるヒドロキシクロロキンは、新型コロナウイルス感染症（COVID-19）肺炎患者の治療に緊急使用が認可されたが、有害事象の発生で否定的な評価となったことを受け、関節リウマチ患者の日常治療におけるヒドロキシクロロキンの使用に関連するリスクを判断するために、ヒドロキシクロロキンの単独およびアジスロマイシンとの併用の安全性に関する研究が実施された。データには、ドイツ、日本、オランダ、スペイン、英国、米国における 14 の請求データまたは電子医療記録のソースが用いられ、多国籍比較が行われた。1000 名近くのヒドロキシクロロキンの使用患者と併用患者 100～300 名規模に対して、30 日間追跡調査が行われた。その結果、ヒドロキシクロロキン治療は、関節リウマチ患者において短期的にはリスクの増加を示さないように思われるが、長期的には心血管死亡率の超過と関連しているということが示唆された。また、アジスロマイシンを追加すると、短期であっても心不全や心血管系死亡のリスクが増加することが明らかになり、COVID-19 流行下における、ヒドロキシクロロキン治療患者へのカウンセリングにおいて、利益とリスクのトレードオフを慎重に考慮するよいエビデンスとなった。

[参照]

*LEGEND イニシアティブ

https://www.ohdsi.org/wp-content/uploads/2021/10/LEGEND_OHDSI_Community_Spotlight.pdf

**LANCET 掲載の高血圧治療薬研究

https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)32317-7/abstract

***JAMA 掲載の高血圧治療薬に関する研究

https://pubmed.ncbi.nlm.nih.gov/32065600/

****糖尿病治療薬プロトコルに関する多国籍連合分析研究事例

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10871374/

*****OHDSI 内の COVID-19 特設ページ

https://www.ohdsi.org/covid-19-updates/

******関節リウマチの治療におけるヒドロキシクロロキン単独およびアジスロマイシンとの併用のリスク：多国籍後ろ向き研究

https://www.thelancet.com/journals/lanrhe/article/PIIS2665-9913(20)30276-9/fulltext

2.5 まとめ　国内における課題と示唆

バイオデータの連携・利活用を推進には DX が欠かせない。ただし、重要なことは、単に電子データを作っていくことでもなく、積みあがったデータの活用方法を考えることでもない。データの連携・利活用を前提とした、活動、仕組み、製品などを戦略的・構造的に構築、実装していくことが DX である。

バイオデータの活用は 2 つに大別される。1 つ目は「個人」データの活用であり、もうひとつは「集団」データの活用である。前者については、2.2 章で海外の先進事例として米国ならびに英国の医療情報連携の取組を紹介した。地域医療機関の間で医療情報を連携する地域医療連携ネットワークは、国内においても地域医療再生計画がスタートした 2011 年度頃を境として急増し、250 を超えるネットワークが構築されてきた。一方で、事実上 1 医療機関のための地連ＮＷにとどまっている事例や、自主財源がなく事業の継続性に疑問のあるネットワークも散見されるなど、課題が大きい状況である。そうした背景もあり、現在厚生労働省では医療 DX 令和ビジョン 2030 を掲げ、その中で全国医療情報プラットフォームの構築を進めている。プラットフォーム構築において重要な役割を果たすのが医療情報交換の国際標準規格である HL7 FHIR であり、HL7 FHIR については 3.2 章において詳述する。

2 つ目の活用方法である「集団」データの活用についても課題は大きい。日本製薬工業協会の調査によると、国内において約 500 件の疾患レジストリが存在し、医薬品開発においては市場性調査や患者リクルートなどへの活用が既に始まっているとされる。新規医薬品の開発ターゲットが市場規模の大きい生活習慣病領域から、難治性・希少疾患といったアンメットメディカルニーズ領域に変わりつつある中、これら疾患レジストリの活用には非常に期待が寄せられている一方で、2.3 章で紹介したような海外における RWD を用いた Historical Control 群の構築や薬事承認申請への活用は、国内においては極めて例外的である。解決策の方向性のひとつは、データベース間の連携によるデータの質と量の確保であり、そのためには独立のデータベース構築から複数データベース間の連携を視野に入れることが考えられる。冒頭 1.2 章で紹介した次世代創薬 AI 開発

（DAIIA）における連合学習の事例はまさにこの発想である。今後は幅広くバイオデータを連携し活用していくこと、それを可能とするために連合型を始めとする技術の検討と実装を進めていくことが強く求められる。

3　リアルワールドデータ(RWD)の利活用に向けた取組

3.1　RWD の種類と特徴

リアルワールドデータ（Real-World Data, 以下 RWD という。）は「様々な手段により日常的に収集される、健康状態および医療行為に関するデータ(*)」と定義される。健康・医療領域においては、特定行為および物質の有効性、安全性を確認するために臨床試験、臨床研究の形が取られるが、試験や研究における対象はその目的がゆえに条件が厳密にコントロールされた集団であることが通例であり、日常（Real-World）における挙動とは必ずしも一致しないことがある。ゆえに、RWD の有効な利活用はコントロール集団を対象とした試験や研究と同様に、健康・医療領域の発展において非常に意義が大きい。本項においては RWD の種類と特徴、ならびに利活用における課題について、以下、構造的に整理する。

[参照]

* Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products: Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products | FDA

3.1.1 RWD の種類

国内外において利活用が進んでいる RWD の代表例として、診療報酬明細書（レセプト）、DPC、電子カルテなどが挙げられる。これらはいわゆる”医療行為”を直接的な起点として発生するデータであり、日本薬剤疫学会による「日本における臨床疫学・薬剤疫学に応用可能なデータベース調査」では、医療機関ベース、保険者ベース、保険薬局ベース、その他に分類し、データベースの特徴が整理されている(*)]。加えて、テクノロジーの進化は RWD に「多様性」をもたらしている[図 1]。医療行為に限らない、日常的な健康・医療に関連する行動（例えば運動、食事、睡眠など）、消費に関連する行動（例えば食品、医薬品、健康食品、その他嗜好品などの購買）、ソーシャルメディア上で個人から発信される情報なども RWD として利活用が検討されている。関連し、個人のスマートフォン上での PHR (Personal Health Record)アプリケーションの利用は、これらのデータに「結合性」を持たせ、より網羅的なデータセットを構成している。さらに、ウェアラブルデバイス、IoT デバイス等の普及は、これまで“点”でしか生成されなかった生体計測データに「連続性」を与えつつある点は RWD を考える上で大きな進歩である。これまでの医療行為を起点とする RWD は年に 1 回の健康診断、四半期に 1 回の定期健診といった限られた時点でしか生成されなかったわけであるが、デバイスの普及によって日単位、分単位、秒単位で生体情報（連続生体量）が RWD として追加されていくことはビッグデータとしての情報量が指数関数的に増加していくこと意味するとともに、RWD 利活用を考えるうえでも新たな視点をもつ必要があることを強く示唆する。

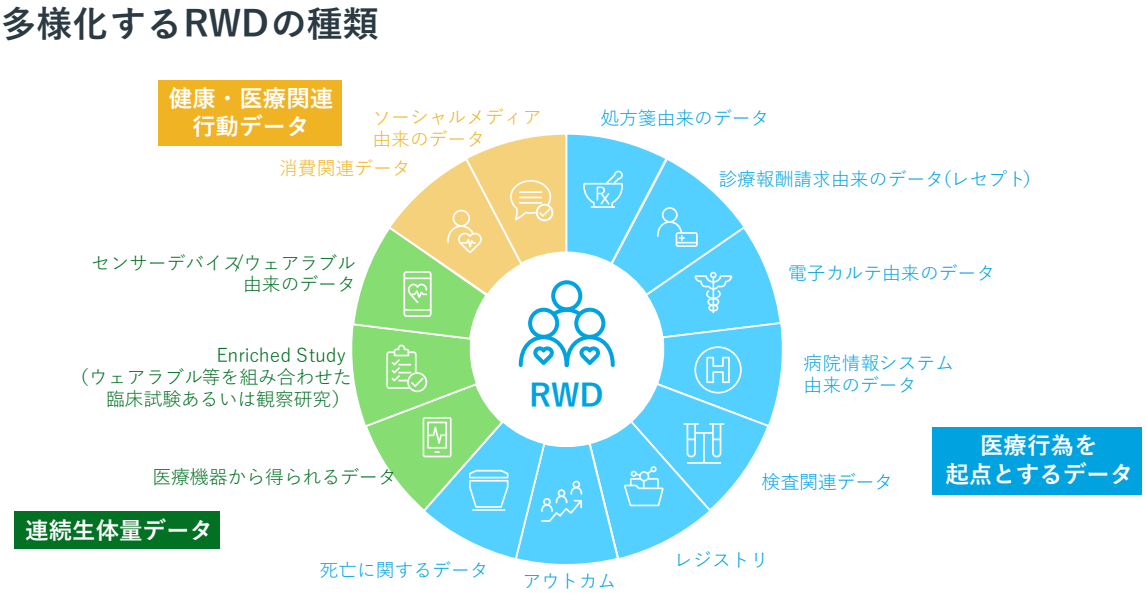


図 1. 多様化する RWD の種類

なお、上記は主に集合型のビッグデータベース（複数の発生元で収集されたデータを 1 か所で統合的に集めたもの）であるが、1.2 章の導入でまとめたように、プライバシー保護の観点などから、データを 1 か所に集めるのではなく、データそのものは発生元の各所に保持したまま目的に応じた統合解析をする、連合型と呼ばれるデータ利活用の基盤整備が世界各国で戦略的に進められてきている。

[参照]

*日本における臨床疫学・薬剤疫学に応用可能なデータベース調査：
https://www.jspe.jp/committee/020/0210/

3.1.2 利活用目的に応じて考慮すべきポイントと RWD の特徴

RWD はその利活用に応じてデータが選択されるべきである。本項では RWD を選択するにあたって検討すべきフレームワークを紹介するとともに、フレームワークに添って主だった RWD の特徴を整理する。

ＲＷＤを選択する際に考慮すべき主な観点は以下の通りである。

- **広さ**：解析目的に足るデータ量（患者数、属性分布など）
- **深さ**：解析目的に必要なデータ項目（特に“構造化”されている項目）
- **時間**：解析目的に適切な期間（時期、観察期間、即時性など）
- **利用プロセス**：解析するために必要な手続き、費用、制約条件など

広さ

RWD 利活用をイメージした時に第一義的に確認をすべき内容がデータの広さ、データ量である。解析目的に必要な N 数が確保できることが前提となるが、健康・医療領域においては、特にデータ背景の分布を含めて“広さ”を確保することが必要となる。例えば、ある生活習慣病の解析を考えた場合、単純な N 数は十分であったとしても、特定の地域、年齢、治療歴などに偏ったサンプル構成であれば、それは疾患を代表するデータにと見なせない可能性が高い。これは逆もまたしかりであり、例えば希少疾患かつ特定の治療歴のある集団の解析といった目的であれば求められる N 数は前述のケースよりも相当に少なくなることもある。

深さ

医薬品産業を中心に一定の RWD 利活用経験が蓄積されている昨今においても、常に課題となるのがデータの深さ、データ項目である。十分な N 数があっても、解析目的に必要な項目（データ）がなしには望む結果を得ることは難しい。例えば、医科レセプトは一般的に N 数が確保しやすく、疫学ならびに医療経済性分析などにおいて大変有用なデータである。一方で、医療アウトカムの分析を考慮した場合には、医科レセプトでは充足されていない、検査値の推移、画像所見、転帰、主訴などを含むデータソース（例えば、病院情報システム由来のデータ）を用いる必要がある。

RWD の深さは、構造化処理（2.1 に記載した、データのハーモナイゼーションの解説も参照）ならびにプライバシー保護の観点を考慮せねばならない。解析にあたっては、検査値はコードや単位の統一が必要となり、画像や記述においてはそのままの状態で統計解析を行うことができないため、構造化されたデータに変換する前工程が不可欠である。また、データの深さにはプライバシー保護の課題も避けられない。データ項目が増えれば増えるほど解像度の高い分析が可能になるわけであるが、平行して対象者の再特定リスクも上昇する。以上のように、解析に必要なデータの深さを担保しつつ十分な N 数を確保することは決して容易ではなく、ほぼ全ての研究者が壁に当たった経験をしていることだろう。自然言語処理技術、画像解析 AI 等による構造化技術と連合型ネットワークの組み合わせは、広さと深さの課題を解決しうる方法であり、4 章にて海外における先進事例ならびに技術について解説する。

時間

適切なタイミングのデータを適用できるかどうかも、RWD 利活用において不可欠の要素である。検討したいイベントの発生時期が含まれているか、解析に必要な追跡期間が満たされているかはその典型例であるが、先のコロナ禍で課題となったように即時性、即ちどれだけ最新のデータを分析し適切なアクションを検討することができるかという点も RWD を選択する上での重要な観点としてクローズアップされている。

利用プロセス

最後はプロセスの観点、平たくいうと、どれくらい簡単かつ安全にデータを利用できるかである。健康・医療領域のデータはその性質上最善の注意を払うことは大前提であるが、その中でもどのような申請を経てどれくらいの期間でデータが利用可能になるか、データ利用段階でどれくらい“きれいな”データとなっているか（解析までの前工程をどれだけ短縮できるか）、そして解析ツールの提供など、データ活用におけるサポート環境がどれくらい整っているかといった観点も RWD を選択するうえで重要な要素となる。

以上の観点を踏まえて、上記、日本薬剤疫学会による「日本における臨床疫学・薬剤疫学に応用可能なデータベース調査」にも掲載されている代表的な RWD の特徴を整理する。

処方箋データ

複数のデータベースがありかつデータ量を確保しやすい使いやすい RWD である一方、例えば疾患名の記載がないなど深さは限定的であり、利用用途は限定される。

広さ：複数のデータベースが整備されており、日本全国の院外処方の 2 割以上をカバーしているものなどもあり、データ量は充実している。

深さ：処方箋記載内容に留まるため、項目数は限定的である。

時間：更新頻度の速さや十分な蓄積期間のデータを提供しているものもある。

利用プロセス：民間データベースと解析に利用可能なソフトウェアなども充実している。

レセプト/DPC データ

複数のデータベースがありかつデータ量を確保しやすい使いやすい RWD であり、健康保険組合由来のレセプトデータは対象保険組合に属している限りデータの連続性も担保される。一方、臨床検査値や画像診断、テキスト情報などを含まないため、医療アウトカム分析用途は限定される。

広さ：NDB に加え、民間のデータベースも複数あり、社保、国保を含めデータ量は充実している。

深さ：診療報酬請求書に含まれる項目がカバーされており、疫学や医療経済性分析に用いることができる。

時間：更新頻度の速さや十分な蓄積期間のデータを提供しているものもある。

利用プロセス：民間データベースも充実しており、解析に利用可能なソフトウェアなども充実している。

電子カルテ/病院情報システムデータ

レセプトデータには含まれない情報を含み、医療アウトカム研究に用いることができる。一方で、処方箋、レセプトデータに比してデータ量は限定的で、医療機関が変わると連続性は担保されないことが多い。また、テキスト情報や画像情報は解析時において構造化処理が必要となる点、情報の深さが得られる反面、プライバシー保護の観点も課題となることがしばしばある。

広さ：特に患者数の少ない疾患や複数条件の設定が必要な場合はレセプトや DPC データと比較して十分な N 数を確保することが難しい場合がある。

深さ：アウトカムを含むデータを得ることが可能であるが、データクレンジングやテキスト、画像データの構造化処理が必要となるケースが多い。また、米国における臨床健康情報の相互運用を支持する標準用語集 SNOMED CT のような統一規格が国内にないことからも、解析に先立って Harmonization に大きな工数を要する。

時間：即時性のあるデータ活用や十分な観察期間のあるデータを利用することが難しい場合がある。また通常医療機関を跨いだ場合にはデータの連続性は失われる。

利用プロセス：民間データベースも提供されているが、深い情報を含むデータを活用する場合には、個人情報保護や倫理面などの課題から適切な利用のための申請プロセスや審査期間などがかかることが一般的である。

3.1.3 RWD 利活用における課題と注意点

RWD 利活用における課題と注意点について解説する。図 5 に示した RWD から RWE 創出までのプロセスに則り、ここでは大きく 4 つの課題に分けて整理する。

- **課題１：正規化**
- **課題２：構造化**
- **課題３：個人情報保護**
- **課題４：データ間の比較と統合**

課題１：正規化

RWD として利活用するためには、データが電子化されていることが必要条件となるが十分条件ではない。電子化は図 3 でも解説したデジタイゼーションの段階であり、DX 即ちデータ利活用を前提とした場合には、データが統計解析に適した形で正規化（用語、単位、規格などの統一）されている必要がある。電子カルテシステムの普及に関しては 2.1 章でも触れた通りであるが、現状の国内において電子化自体も依然として進展の余地はあるものの、DX の観点を踏まえ、より課題が大きいのはデータの正規化である。電子カルテシステム自体を見ても、ベンダーごと製品ごとに仕様は異なっており、さらに多くの場合、各医療施設に最適化された仕様にカスタマイズされている。また、大規模な病院では電子カルテシステム以外にも数 10 の情報システムが動いており、システム間の相互接続環境も含め、一律に標準化を目指すことは原則困難である。

電子カルテ情報の標準化については、具体的な対応として、Web サービスの技術を用いて医療情報を交換する際の国際標準規格である HL7FHIR の活用が進められており（HL7 FIHR の解説は 3.2 章に記載する）、「医療 DX 令和ビジョン 2030(*)」においては、HL7 FHIR 準拠の標準クラウドベースの電子カルテシステムの導入を 2030 年までに 100%とする目標が掲げられている。本取組により、診療情報連携などの「個人」のための「個人」データの活用は大きく進展することが想定できる一方で、「集団」のための「集団」データ解析である RWE 創出ならびにアルゴリズム構築などを考えた場合には、考慮すべき課題として「データの正規化」が残る。米国の電子カルテシステムにおいて標準的に使用されている用語集に SNOMED CT（Systematized Nomenclature of Medicine-Clinical Terms）がある。SNOMED CT は、病理疾病分類 SNOMED をベースに、英国病院用の医療用語集や MesH（米国国立医学図書館の医学用語集）などを統合して作成され、医学用語を統一的に扱い、表記を統一することを目的とした医療用語集であり、Stage 2 Meaningful Use において、問題リストの記録には SNOMED CT が必要とされている他、米国ヘルスケア情報技術標準パネルの相互運用性仕様でも必須とされている(**)]。用語の統一は RWD 利活用を効率的に行うために必須の要件であり、裏腹に国内における大きな課題でもある。この課題の解決策のひと

つが OMOP CDM を始めとする RWD の Common Data Model であり、3.2 章以降で詳細を解説する。

なお、上記は電子カルテならびに病院情報システム由来の RWD に焦点を当てたものであるが、国内においても利活用が進んでいるレセプトデータや DPC データについても少し触れておきたい。社会保険診療報酬支払基金によるレセプト請求形態別の請求状況を見ると、95%を超える診療報酬請求が電子レセプトにてなされている(***)。レセプトは基本的に請求形式として形式が定められている（データが正規化されている）ため、活用のためのハードルが相対的に低い。加えて、レセプト情報提供および研究に関する倫理指針や各種ガイドラインの整備も進んでいることもあり、これらがレセプトデータの活用が相対的進んでいる理由でもある。

課題２：構造化

電子カルテシステムあるいは病院情報システム由来のデータのテキスト情報には、診断や治療の判断と根拠、患者の症状詳細など、重要なアウトカムデータが含まれている。しかしながら、これらは非構造化情報であるため、統計解析を試みる場合には構造化処理が課題となる。例えば、「喫煙歴」がある患者の投薬情報を抽出する場合、テキストから「喫煙」という単語を拾っても、それが「喫煙歴があるのか、ないのか」、「患者自身の喫煙歴なのか（家族などの背景情報か）」、はたまた全く関係ない「ご家族に院内の喫煙について説明」といったメモなのかは単語を抽出するという動作だけでは判断ができず、テキスト情報には患者氏名、家族の連絡先、患者の職歴や趣味など、個人情報が埋もれていることが多いことも、適切な利活用に際してリスクが大きい。そのような観点から注目されているのが自然言語処理技術（Natural Language Processing: NLP）技術である。昨今では基礎的な医学用語集やオントロジーなどがそろいつつあることもあり、精力的に研究が進められている。自然言語処理技術については 4.2 章において詳細に解説する。

課題３：個人情報保護

個人情報保護もまた RWD 利活用における重要な視点である。RWD は個人識別符号または要配慮個人情報を含むケースがほとんどである。RWD の匿名化に対して「k 匿名化」あるいは「リスクベースアプローチによる非特定化手法」が一般的に用いられている。これらは個人について特定可能な「直接識別因子」を削除あるいは置換のうえで、組み合わせにより個人特定が可能な「間接識別因子」をデータの特性、利用者、前例などに基づき、許容可能なリスク閾値を設定し、そのリスク閾値に収まるまで間接識別因子の削除や加工を繰り返す手法である。リスクベースアプローチは、HITRUST (the Health Information Trust Alliance), IOM (Institute of Medicine), Canadian Council of Academies, European-based PhUSE などの国際団体および HIPAA (Health Insurance Portability and Accountability Act)法においても推奨されている[Ref]。一方で、k 匿名化に要する工数はしばしば課題となり、データによってはそもそも匿名化自体が不可能であるケースもある（遺伝子配列情報などはその典型例）。個人情報保護の観点では、いわゆる次世代医療基盤法による認定匿名加工医療情報作成事業者が定められている(****)ものの、これらのような課題感から匿名化および仮名化以外の方法として、学習結果、解析結果のみを統合する「連合型」が昨今注目されてきている。

課題４：データ間の比較と統合

課題 1 から課題 3 は主に各 RWD やデータセットを構築する際の課題であるが、それらをクリアしても残るのが、異なるデータ間の連携（比較、統合）である。本報告書の主題でもある連合型のデータ連携を可能にするためには、課題 1（正規化）からさらに踏み込んで、データ間の規格が統一されていることが必要となる。その際の有力な手法となるのが、3.2 章以降で解説する RWD CDM である。その際には、上述の電子カルテの標準化において医療情報交換の“国際標準規格”である HL7 FHIR の導入が検討されていることと同様に、利活用の幅を狭めないためにもガラパゴス化することなく、国際標準規格の導入を念頭に置くことが重要である。

[参照]

*医療 DX 令和ビジョン 2030

https://www.mhlw.go.jp/stf/shingi/other-isei_210261_00003.html

** SNOMED CT

https://www.nlm.nih.gov/healthit/snomedct/index.html

***社会保険診療報酬支払基金　レセプト請求形態別の請求状況

https://www.ssk.or.jp/tokeijoho/tokeijoho_rezept/index.html

****「次世代医療基盤法」とは

https://www8.cao.go.jp/iryou/gaiyou/pdf/seidonogaiyou.pdf

3.1.4 RWD 利活用における医療情報システムの課題

電子カルテシステムの多くは、各々の医療現場において最適なオペレーションができるようカスタマイズされている。また、レジストリやバイオバンクは、当然のことながら各々の研究目的を達成するために、最適なフォーマットに則った構造でデータが収集、蓄積されている。このように、既存の医療情報システムはそもそもの設計思想として、施設間、医療従事者間で共有することや比較すること、あるいは複数のデータソースを連携して解析することを前提としていない。そのため、医療情報のリアルタイムでの共有や連携はもちろんのこと、統合解析を試みようとした場合でも、解析の前準備としてのデータクレンジングに多大な時間とコストが必要となるのは当然である。さらには、複数システムにより複雑な接続体系で構成されている医療機関内システムから、利活用に応じたデータセットを生成するためにはデータウェアハウスの構築などが必要となること、その上でもそもそも入力がないデータや誤った箇所に入力されている場合にはデータの欠損、脱落が生じることもあり、総じて課題は大きい[図 2]。

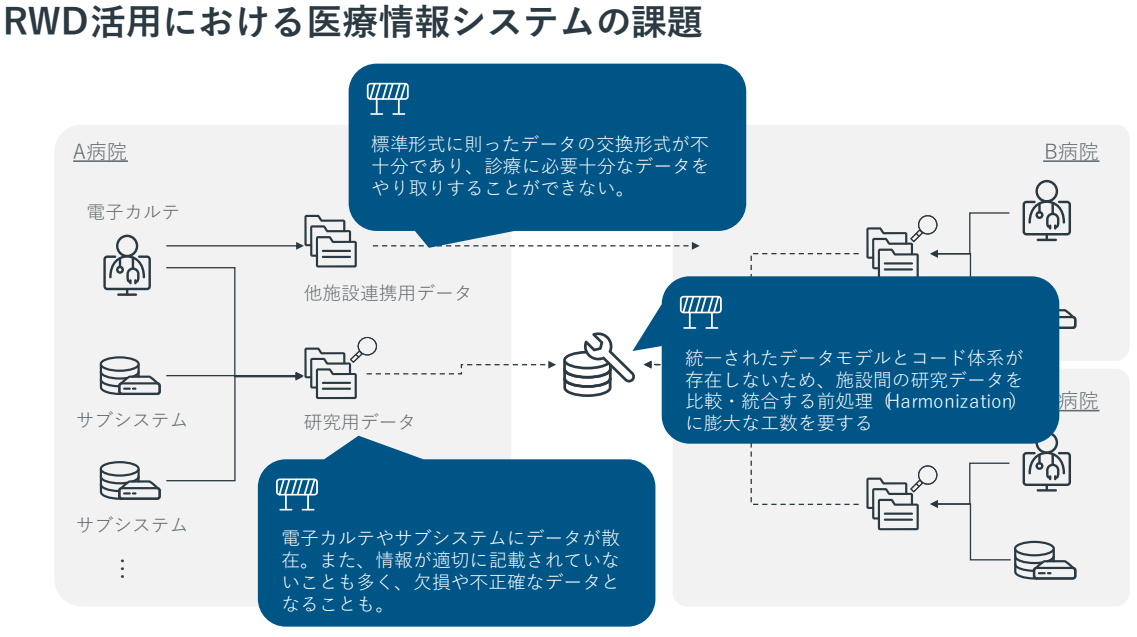


図 2. RWD 活用における医療情報システムの課題

これらの課題解決のためには、大きく 3 つの方向性が考えられる。1 つ目は医療情報交換の標準規格の整備と採用、2 つ目は RWD 解析に有用な標準データ規格とレポジトリの整備、3 つ目はデータ入力段階における情報の正規化と構造化である[図 3]。

課題解決の方向性

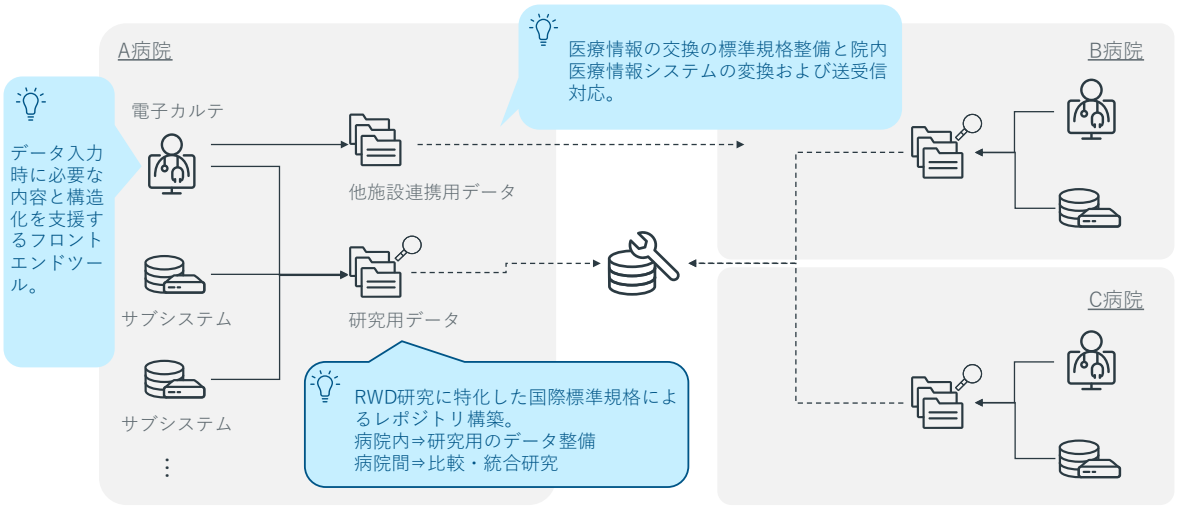


図 3. 課題解決の方向性

次章においては、主にこの 2 つ目「RWD 解析に有用な標準データ規格」について詳述するが、1 つ目の HL7 FHIR をはじめとする医療情報交換規格、3 つ目のデータ入力のテンプレートシステムについても合わせて解説する。

3.2 RWD における Common Data Model（CDM）

3.2.1 RWD CDM とは

リアルワールドデータには用途や出自の異なる様々なデータが含まれる。通常データセットあるいは保有施設（診療録等であればであれば医療機関等）により、データ形式や入力ルール、用語（用語コード）が異なるのは通例である。そのためデータ分析には、データの前処理やクリーニングと呼ばれる作業が必要となる。前処理は単一データセットの分析でも必要であるが、分析対象が複数のデータセットあるいは施設を跨いだ分析となる場合にはより煩雑な処理作業が要求される。例えば、我が国における診療報酬明細書（レセプト）データでは審査支払機関（社会保険診療報酬支払基金又は国民健康保険団体連合会）への提出フォーマット(*)や診療行為・医薬品・傷病名といったデータ項目ごとのマスター(**)が定められている。そのため、大規模のデータでも比較的分析しやすいといえる。一方で、疾患レジストリや電子カルテに基づくデータベースの場合、統合して解析する際には統一されたフォーマットがないかあるいはあっても細かいデータ形式や用語に違いが出てくるため、分析前に要求される作業は非常に煩雑になる。

このような課題解決の方法として、Common Data Model（CDM）を採用することの有用性が検証されている。CDM を採用することで、データの前処理に係る労力を削減できるだけでなく、同じ CDM を採用しているデータセットに対しては同じ手法で解析が行えるため、解析自体の労力も削減することができる。また、データセットや施設を跨いだ連合型のデータ利活用の実現のためには CDM でデータフォーマットが統一されていることが必須である。CDM 採用のデメリットとしては、データ変換に一定のコストがかかること、データ変換時にどうしても情報の一部が脱落することが挙げられる。しかしながら、すべてのデータ種において一次利用の段階からデータのフォーマットをそろえておくことは、そもそも非現実的であり、そもそもトランザクショナルなデータベースと分析用データベースでは要件が対立するため、分析用には別のデータベースが必要となることが多いとされる(***)。そこで、情報の一部を犠牲にしてでもリアルワ

ールドデータの利活用のための目的に応じたデータベースを用意しておこうという考えに基づき CDM を採用することには合理性がある。

[参照]

*社会保険診療報酬支払基金
https://www.ssk.or.jp/smph/yoshiki/yoshiki_01_h30i.html#cms04
**診療情報提供サービス（厚生労働省）
https://shinryohoshu.mhlw.go.jp/shinryohoshu/searchMenu/
*** Murphy SN. Data warehousing for clinical research. In:Liu L, Tamer OM, eds. Encyclopedia of database systems Springer, 2009, ISBN: 978-0-387-49616-0.

3.2.2 医療情報交換規格、入力テンプレートシステムと RWD CDM

3.1.4 章で言及したように、RWD 活用を進めるためには医療情報システム上の課題解決が不可欠である。今後課題解決策のひとつである RWD CDM に焦点を充てて解説を展開していくものの、本章では HL7 FHIR や SS-MIX2 に代表される医療情報交換規格ならびに、システムへの情報入力段階を制御する入力テンプレートシステムの紹介と RWD CDM との違いを整理することとする。

医療情報交換規格

医療情報の分野においては、システムの可換性、医療機関間での情報交換の必要性、診療情報の効果的な利活用の観点から、標準化を行っていく必要がある(*)。厚生労働省では、診療情報の形式・交換・蓄積の規格や用語のマスター等保健医療分野において必要な標準規格を厚生労働標準規格として認め、普及を図っている(**) [図 4]。

出所：第 4 回健康・医療・介護情報利活用検討会、第 3 回医療等情報利活用ＷＧ及び第 2 回健診等情報利活用ＷＧ　資料（厚生労働省）
https://www.mhlw.go.jp/stf/newpage_14239.html

図 4. 厚生労働省標準規格

医療情報交換のための標準規約として、HL7（Health Level Seven）がある。HL7 は医療情報システム間の ISO-OSI 第 7 層アプリケーション層に由来しており、患者管理、オーダー、照会、財務、検査報告、マスタファイル、情報管理、予約、患者紹介、患者ケア、ラボラトリオートメーション、アプリケーション管理、人事管理などの情報交換を取り扱う(***)。HL7 V2 メッセージ標準は病院運営、物流、診療プロセスなどのメッセージをサポートする標準規格であるが、HL7 V2.5 は後述する SSMIX2 標準化ストレージにおける各種診療情報の標準格納形式として定義されており、我が国において比較的広く用いられている(****)[医療情報第 6 版、医療情報システム編、pp. 344-345]。

ここ数年、新たな医療情報の標準規格として海外において先行して普及が始まっている HL7 FHIR（Fast Healthcare Interoperability Resources）が注目されている。HL7 V2 や V3 はあくまでデータをどう表現するかの構造を指定したものであるが、HL7 FHIR では施設内のみならず施設間でのやり取りや実装を重視した規格となっており、病院情報システムやその他のアプリケーションでどうデータのやり取りを実現していくかを想定している。FHIR では現在の Web サービスで主流のアーキテクチャーで主流となっている RESTful API を採用しており、PHR（Personal Health Record）などの現状の医療情報システムの外側のアプリケーションとの連携もより柔軟に行える。

諸外国においては特に新規分野において積極的に活用することで、FHIR 規格への転換を政策的に誘導することが試みられている。例えば米国においては、国の情報システムや官民連携イニシアティブによる実装ガイド及びサンドボックスの構築・展開（CMS Blue Button 2.0 等）やインセンティブ施策（Meaningful use における認定電子カルテへの補助金等）により普及を促進している(*****)。

一方で、データをどう格納、蓄積していくかという観点からは診療及びデータの二次利用を視野に入れた取り組みとして SS-MIX2(******)があり、広く普及している。SS-MIX2 は HL7 V2.5 メッセージ（処方・検査情報等）を格納する標準化ストレージと、標準化されていない情報を格納する拡張ストレージからなる。

このように、医療情報標準規格はあくまでも診療における一次利用を中核的な目的としながら如何にデータを効率的に運用、保存、交換するかという観点から設計されている。すなわち、2.1 節で提示したフレームワークにおける「個人」のデータを「個人」のためにいかに有効に管理するかということに主眼が置かれている。医療介護現場で医療情報標準規格が浸透していけば、例えば電子カルテと検査システム等の病院での異なるシステム間の情報連携がスムーズになり診療業務が効率化され、地域における医療連携において診療情報提供書のやりとりを含む情報交換が円滑になることにより、医療介護連携や地域包括ケアの実現に貢献する。また、特に HL7 FHIR のような情報の交換に力点が置かれた規格の浸透により、個人で医療情報を管理し様々なアプリケーションで活用することが可能となってくる。ただ、リアルワールドエビデンスの創出や AI/ML のような技術を活用した個別化医療の実現など、「集団」のデータを活用していくうえではこれだけでは十分ではない。

入力テンプレートシステム

複数の施設からデータを収集することにより、より多くデータを得られることになるが、施設が異なれば利用しているシステムが異なりことからデータ形式や構造の違いが生じる。入力テンプレートシステムは、必要とする項目を盛り込んだ情報入力テンプレートを作成し、電子カルテ等の情報システムにテンプレートに添った情報が入力されることにより、構造化されたデータを蓄積するためのツールである。

国内ナショナルセンターで研究、構築が進められている入力テンプレートに関連する取組に JASPHER (Japan Standard Platform for Electronic Health Records)が挙げられる。JASPEHR プロジェクトは電子カルテに蓄積された電子的診療情報を効率的かつ効果的に集約して、本邦における全国規模での診療情報データベースを構築し、欧米並みに共用できる研究リソースシステムを構築することを目指すものであり、その中でテンプレートシステムの構築と運用が行われている。標準規格に基づいた“共通のテンプレートマスタ”を配布し、自動的に各社のテンプレートシステム形式に変換、入力されたデータを収集するという、一連の仕組みの検討を行っており、標準規格としては HL7 International を中心に策定された FHIR が採用されている(*******)。

米国の医療情報スタートアップ企業である Flatiron Health 社は当時いわゆるユニコーン企業としても注目を集めた入力テンプレートを用いたオンコロジー特化型の電子カルテシステムを展開する民間企業である。同社は、オンコロジーに特化した医療情報記録と解析ツールを組み合わせた OncoCloud という仕組みを作り上げた、RWD 活用を目的とした基盤整備から実際の活用までの仕組みを作り上げた好事例である。実際に、2.3.2 章で取り上げたように、米国において Flatiron 社の電子カルテを通じて構築された RWD を用いた薬事承認も行われている。また、国内においても、このような情報入力段階におけるシステムを用いて解析を見据えた構造化・正規化情報を構築しようという動きは広がりつつある。京都大学と NTT による産学連携企業である新医療リアルワールドデータ研究機構株式会社（PRiME-R）社は、プルダウンメニュー形式で、リストから標準化された医療情報を選択することで、データを構造化した形で電子カルテシステムに入力し、構造化データベース化を可能にするシステムである Cyber Oncology を医療機関向けに提供している。

この様に入力テンプレートシステムは、必要な情報を入力段階から構造化して取得するための有用な手段であるものの、一方で弱点は、入力テンプレートの構築、入力そのものにかかる工数が必要となる点ならびに、入力データが蓄積されるまでに期間を要すること（前向きに情報集めるための工数と時間がかかること）であり、医療情報交換規格ならびに RWD CDM を平行し、既に存在している医療情報の活用を含めた設計を目的に応じて適切にしていくことが重要である。

[参照]

*医療情報の標準化とは（一般社団法人医療情報標準化推進協議会）

http://helics.umin.ac.jp/aboutHelics.html

**医療分野の情報化の推進について（厚生労働省）

https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/johoka/index.html

***HL7 とは（日本 HL7 協会）

https://www.hl7.jp/whatis/

****医療情報 第 6 版 医療情報システム編、pp. 344-345

*****HL7 FHIR に関する調査研究の報告書（厚生労働省）

https://www.mhlw.go.jp/stf/newpage_15747.html

******SS-MIX2 とは？（SS-MIX 普及推進コンソーシアム）

http://www.ss-mix.org/cons/ssmix2_about.html

*******JASPEHR プロジェクト概要

https://cmii.ncgm.go.jp/project_jaspehr.html

3.2.3 RWD CDM の種類と特徴

RWD CDM の特徴を比較する上での観点として下記を提示する。

- データ構造：データを格納するテーブルの構造
- データ入力規則：各テーブル、カラムにデータを入力する上でのルール
- 用語指定：各テーブルに格納される用語集・コードの指定（例：医薬品コード）

データ格納の観点からはデータ構造の指定があれば事足りるが、より効率的にデータ分析・利活用を行おうとするとデータの入力規則や用語の指定まで行われていることが望ましい。大規模なデータを活用した臨床研究を実施する際には、まず共通のプロトコルを作成し、解析プログラムが用意される。データ構造が共通であれば何らかの処理は可能となるが、中に入っているデータの入力規則がそろっていなければ意味のある結果を導き出すことができない（例：性別を｛女、男、その他｝と入力するか、｛女性、男性、不明｝と入力するか）。また、用語の指定がなければ概念のまとまりをどの階層でとらえるかが異なってくるため、これもまた意味のある結果を導き出すことが難しい（例：心疾患>>不整脈>>心房細動のどの階層で意味を指定するか）上に、用語コードが統一されていなければ同じ内容のデータとして捉えることも困難である。このように、分析を目的として考えた場合は上記の 3 点をすべてクリアしていることが望ましい。ただし、当然ながらデータマネジメントの観点からはデータ処理コストが増えることになる。本節では、海外特に米国において代表的な RWD CDM について、運営主体及びその目的も併せて概説する。

i2b2（Informatics for Integrating Biology and the Bedside）

■目的および概要

i2b2 は、臨床データのデータウェアハウス構築と分析のためのオープンソースのソフトウェア/プラットフォームであり、RWD CDM である i2b2 CDM を定義している(*,**)。特にゲノムに関するトランスレーショナルリサーチを円滑に進めることを念頭に置いており、医療ヘルスケアおよび基礎研究のデータを統合、標準化して分析・共有し、精密医療の発展のための連携を可能にすることをミッションとする。また、研究者に診療記録や臨床研究に関するデータをゲノムデータにも適応するかたちで研究に活用できるようにするツールを提供することをビジョンとしている。

i2b2 のプラットフォームには 2 段階の活用方法があり、1 段階目としては、研究者コミュニティにおいて研究対象としたい患者群を、プライバシーを保護した形で同定する。2 段階目としては、IRB の承認のもと研究プロジェクトごとにデータマートをつくり、より詳細なデータセットの作成ならびに研究に沿った前処理を行うこともできる。この過程で、診療録データと、CRF（case report form）由来の情報や GWAS（Genome-Wide Association Study）などの研究用に作成されたデータを統合することも可能である。

■運営組織

i2b2 は非営利組織である The i2b2 tranSMART Foundation により運営されている。同財団は 2017 年に i2b2 と tranSMART Foundations がそれぞれのプラットフォームと共に統合されてできたもので、米国国立衛生研究所（NIH）により資金拠出された臨床試験のプラットフォーム（i2b2）およびトランスレーショナルリサーチのためのソフトウェアを管理運用している。

i2b2 は NIH（National Institute of Health）により資金拠出されている National Center for Biomedical Computing (NCBC) のイニシアティブにより支援されている 7 つのプロジェクトのうちの一つである(***)。Harvard Medical School により開発され、Brigham and Women’s Hospital と Massachusetts General Hospital が中心となって設立されたボストンの Partners HealthCare System（現 Mass General

Brigham(****)）を中心とした取り組みであるが、現在は世界の 250 以上の研究機関で活用されている。NCBC は NIH Common Fund のもとでバイオ医学研究におけるコンピュータアーキテクチャーを整備するための取り組みである。

一方の tranSMART データマネジメントシステムは、もともと 2009 年に Johnson & Johnson と Recombinant Data Corporation の研究者によりが開発された。tranSMART Foundation は 2013 年に設立された官民パートナーシップであり、米国と EU の研究者の連携に端を発しており、the University of Michigan、Imperial College London 及びライフサイエンスイノベーションのためのコンソーシアムである the Pistoia Alliance(*****)[https://www.pistoiaalliance.org/]、が設立パートナーとなっている。現在では 100 以上の企業、非営利団体、アカデミア、患者団体、政府組織が tranSMART のコミュニティに参加している。tranSMART では eTRIKS（データカタログ）と TraIT（トランスレーショナルリサーチのための IT インフラ）という EU のトランスレーショナルリサーチの取り組みで開発されている 2 つのデータマネジメント/分析ツールと連携して tranSMART プラットフォームの改良を図っている。

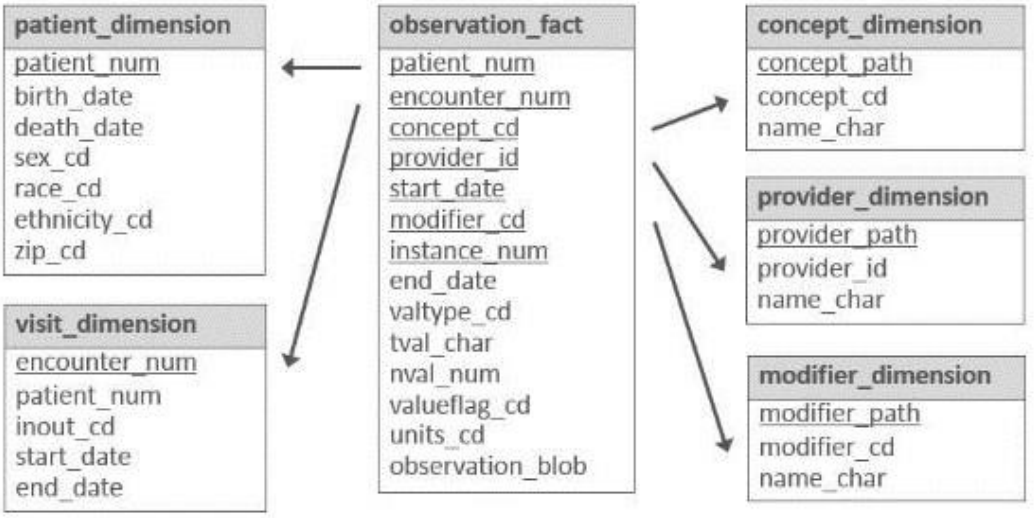
■データモデル概要

ゲノムデータを含めた情報の統合と研究の推進を図るうえで重要な役割を果たすのが、i2b2 Common Data Model (CDM)である。i2b2 CDM は 2004 年に初めて開発された、シンプルで柔軟性の高いデータモデルであり、i2b2/tranSMART ツールでの活用を前提としている。スタースキーマ(******)というシンプルな構造と高い検索性を持つデータ構造により構成されている。i2b2 CDM においては、 診断や処方ごとにテーブルを分けず、ある患者の情報は単一の fact テーブルに格納されている。その中では施設やデータセットにより異なるコード（用語集）を使っており、オントロジー設定により調整している。このため、参加施設は自施設内で用いているローカルコードをそのまま使え、共通のコードにマッピングする必要がない。ゆえに異なるタイプのデータセットに対してある程度汎用的なクエリやツールを提供することが可能となる。さらに、新しいタイプのデータ種もオントロジーを拡張するだけで比較的容易に追加することができる。例えば COVID-19 のように新しい診断や検査が出てきた際にもデータベースそのものではなくオントロジーの解釈設定をアップデートすればよいため、迅速に対応するこ

とができる点もメリットである。現在世界約 200 の施設が i2b2CDM に準拠したデータベースを用意している。

i2b2 において fact table は ONSERVATION_FACT テーブルであり、diagnoses, procedures, medications, laboratory test results などある患者に関するすべての観察項目はこのテーブルに格納されている（これに対して、後述の OMOP CDM などデータ種に応じてテーブルを分けている CDM もある）。付随的な情報は ONSERVATION_FACT テーブルに関連付けられた各 dimension テーブルに入っている[図 8]。

i2b2 CDMにおけるスタースキーマの構成



出所：i2b2 Community Wiki
https://community.i2b2.org/wiki/display/BUN/2.+Quick+Start+Guide

図 8. i2b2 CDM におけるスター・スキーマの構成

例えば CONCEPT_DIMENSION テーブルには concept_path として薬剤のコードを階層構造として格納することができ、ユーザーは i2b2/tranSMART のツールを介して任意の階層で薬剤を指定し、該当する症例を抽出するといったことが可能である[図 9]。このように、各施設はデータそのもの変換は不要なものの、研究プロジェクトにおいて定義されたオントロジーへの対応をメタデータテーブルで設定する必要がある。

CONCEPT_DIMENTIONテーブルにおける階層と定義づけ

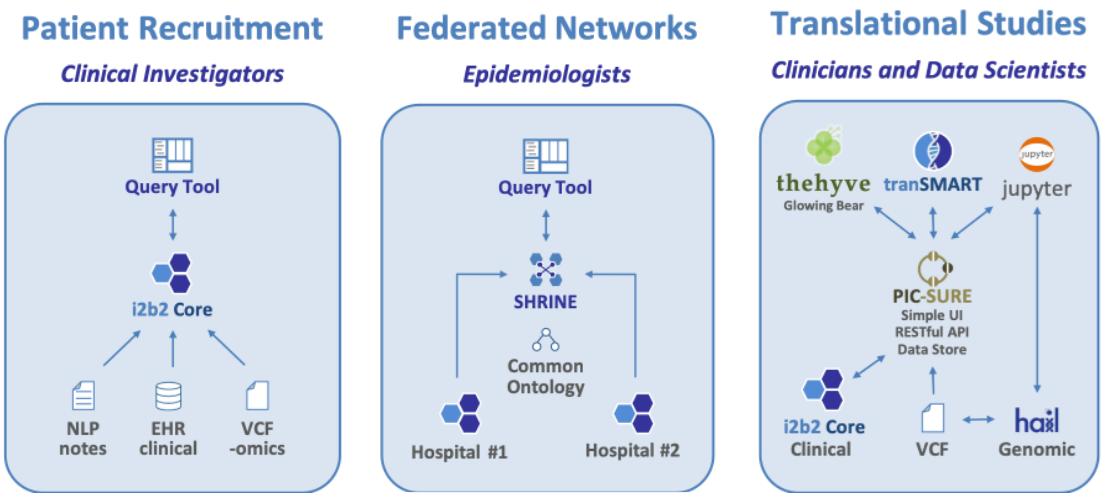
concept_path	concept_cd	name_char
\Med\		Medications
\Med\anti-infectives\		Anti-Infectives
\Med\anti-infectives\ampicillin\	NDC:60429002340	Ampicillin
\Med\anti-infectives\Bactrim\	NDC:00003013850	Bactrim
\Med\anti-infectives\penicillin\	NDC:00002032902	Penicillin

出所：i2b2 Community Wiki
https://community.i2b2.org/wiki/display/BUN/2.+Quick+Start+Guide

図 9. CONCEPT_DIMENTION テーブルにおける階層と定義づけ

■ソフトウェア・アーキテクチャ

i2b2 はデータベースモデル、アプリケーションレイヤー、API を含む臨床試験のプラットフォームであり[図 9]、統合されたデータは匿名化されている。クエリツールにより臨床試験に適格な患者コホートを探索・同定することが主要なユースケースで、疾患に関する情報と遺伝子プロファイルに関する情報を調査することが可能である。



出所：i2b2 tranSMART Foundationホームページ
https://i2b2transmart.org/software/

図 10. i2b2 で提供されるソフトウェア

SHIRINE（Shared Health Research Information Network）という Web ベースのツールにより構成される連合ネットワークにおいては、参加施設ごとに特定のコホートの患者数を確認することができる。これにより疫学等研究者は単一施設でなくより大きなネットワークの中でコホートを作成することができる。

tranSMART はトランスレーショナルリサーチのために開発されたツールである。データの探索、分析、可視化および ETL 機能を含む。ユーザーは transSMART 内でデータの探索や分析を行えるが、Jupyter 等を通して API（図 9 の PIC-SURE）からデータにアクセスすることも可能である。

■活動内容

・community wiki というサイト(*******)が情報共有の中心となっており、最新のソフトウェアに関する情報やコミュニティでどのような活動が行われているかを確認し、質問のやりとりをすることができる。

・ The Accrual to Clinical Trials (ACT)プロジェクトにおいて、National Clinical and Translational Science Award (CTSA) Consortium の施設間で連合型ネットワークを形成し、国として優先度の高い治験への患者組み入れを増やす取り組みを行っている。まずは SHRINE により、ネットワーク全体の患者数を検索し、次のフェーズで各施設において i2b2 で個別患者のレビューができるようにしている。

・i2b2 ソフトウェアでは現在 OMOP の要請にこたえられるような対応を行っている。現状のスター・スキーマのデータモデルでは、一つの fact テーブルに複数の属性を表すテーブルが紐づけられているが、OMOP ではドメインで識別される複数のテーブルの集合としてデータを持っている（procedures, condition, drug, measurement, observation）。最新の i2b2 v1.8.0 では OMOP CDM との互換性を担保している。これは i2b2 が適切なオントロジーの設定と OMOP のテーブル群を複数の fact テーブルとして扱うことにより可能となる。

■課題

データテーブルは定義されているが、参加施設および個々のデータセットで使う用語が統一されていない（入れる値の自由度が高い）ため、メタデータテーブルの設定が必要である。

[参照]

*i2b2 ホームページ

https://www.i2b2.org/

** Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. doi: 10.1136/jamia.2009.000893. PMID: 20190053; PMCID: PMC3000779.

***National Center for Integrative Biomedical Informatics
https://www.ncibi.org/ncbcs.html
****Mass General Brigham
https://www.massgeneralbrigham.org/en/about
*****the Pistoia Alliance
https://www.pistoiaalliance.org/
******スター・スキーマとスノーフレーク・スキーマ（IBM）
https://www.ibm.com/docs/ja/db2/9.7?topic=SSEPGG_9.7.0/com.ibm.db2.abx.cub.doc/abx-c-cube-starsnowschemas.htm
******* i2b2 Community Wiki
https://community.i2b2.org/wiki/

Sentinel

■目的および概要

RWD における Common Data Model (CDM)の事例として Sentinel Initiative を紹介する。Sentinel Initiative は 2008 年に米国の FDA が始めた取り組みで、医薬品及びワクチン、医療機器の市販後安全性調査を実施するために、既存のデータベースを活用した産官学共同プロジェクトである。

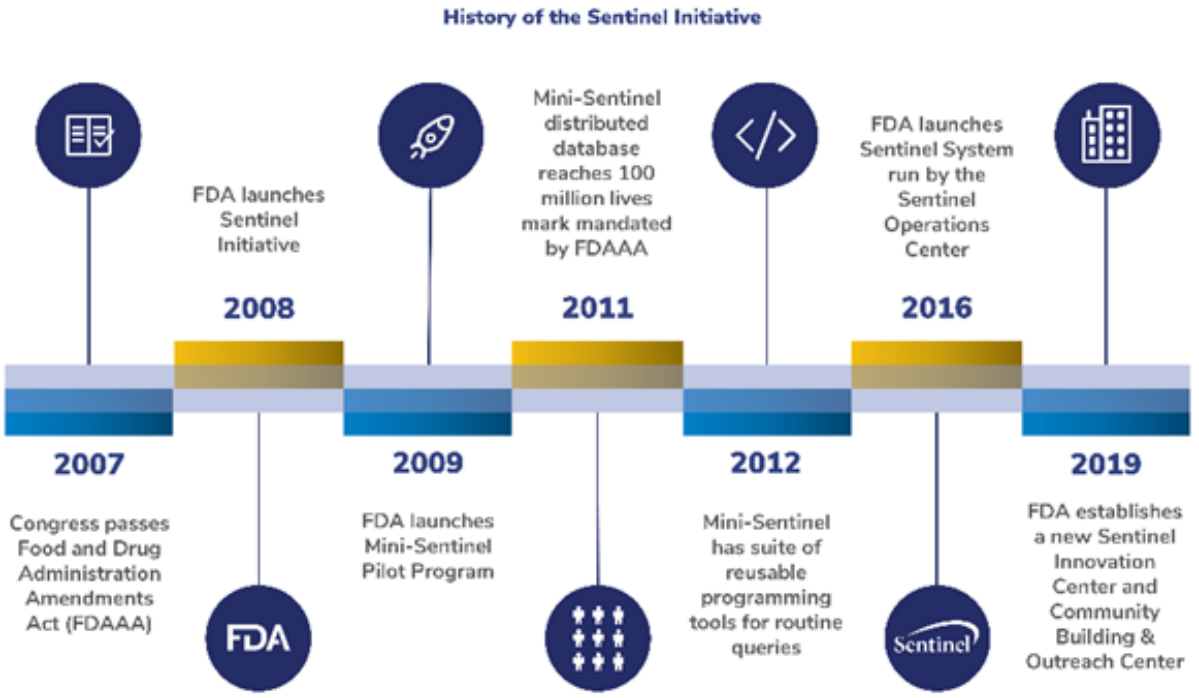


図 11.　Sentinel イニシアティブの歴史

■運営組織

Sentinel initiative の組織は次の 3 つのセンターから構成されており、互いに独立して運営されている。

① Community Building and Outreach Center (CBOC)

デロイトと IQVIA、Nova Research が主体となって、レギュラトリーサイエンスにおけるアイディアをリードし、データの二次利用を促進している組織。

② Sentinel Innovation Center (IC)

Harvard Pilgrim Health Care Institute が主体となり、Sentinel に新しいデータソースを活用できる手法を開発している組織。同組織は、電子カルテから情報抽出して構造化する手法を開発し RWD のエビデンス機能を強化している。

③ Sentinel Operations Center (SOC)

Harvard Pilgrim が主体となり、疫学や統計学、データサイエンスなどの各学会をサポートし、CDM の強化と進化を推進している。

■データモデル概要

安全性監視データベースシステムをベースに、電子カルテデータからも情報を抽出して構造化する手法を開発したことで、より詳細な 1000 万人以上の有用なデータベースへと進化を重ねている。

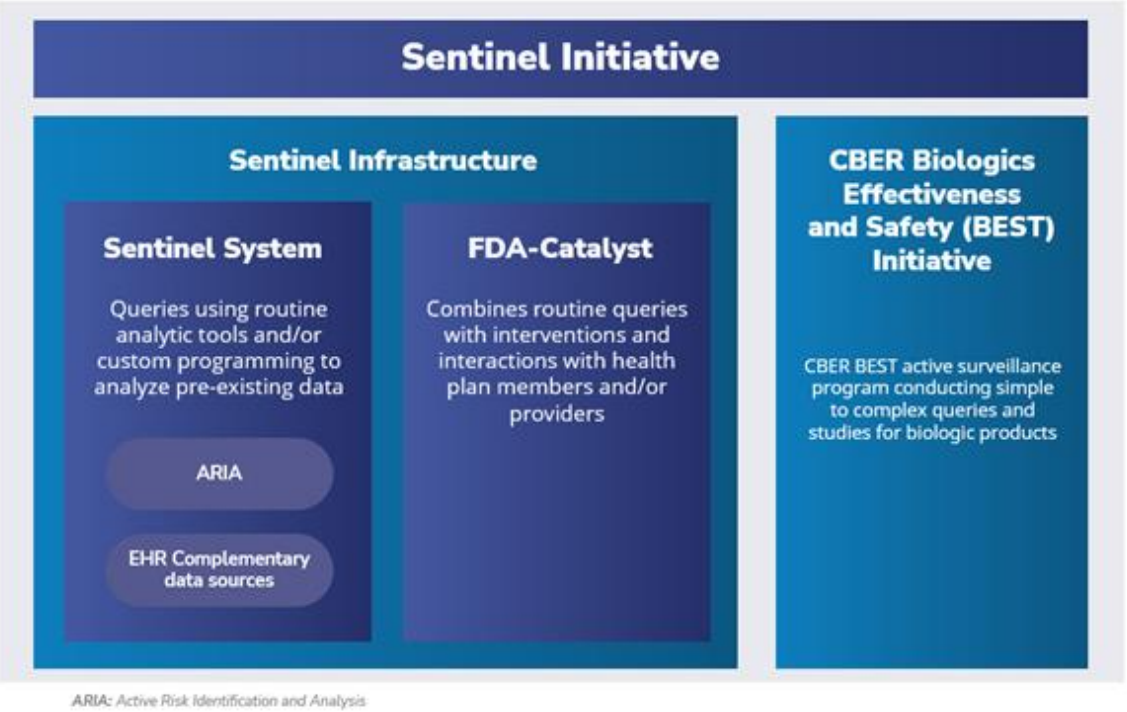


図 12.　Sentinel イニシアティブの主要 3 機能

Sentinel イニシアティブは、①Sentinel システム機能、②FDA-Catalyst をサポートする機能、③CBER 生物製剤の有効性と安全性(BEST)を評価する機能の 3 つから構成されている。その中でも主要な機能となる①と②について詳しく述べる。①Sentinel システム機能は FDA からの承認済み医薬品および医療機器に関する質問をサポートしている。各医療機関におけるデータベースは分散させたまま、プログラムを作成することで、統計的に医療費請求情報と電子カルテの関係のパターンを分析している。特に 2016 年から運用が開始された ARIA と呼ばれる市販後安全性監視システムは、Sentinel 方式の CDM で規定された医療データと、この規定化されたデータを分析するためのツールから構成されており、ルーチン分析やカスタム分析を実行するクエリ処理が、それぞれの医療機関にある既存データベースを分散させた状態のままで行うことができる。②FDA-Catalyst 機能は患者と医療従事者とのやりとりデータを提供し、Sentinel システム内のデータと結合させることで、データを補完している。

■データモデルのしくみ

Sentinel 分散データベースのしくみ

　各医療機関が保持する請求関連情報と電子カルテデータデータに対して既定のプログラムを実行することで、匿名化された分析データを Sentinel 中央オペレーションセンターへ送信することで、患者のプライバシーが保たれたまま、分析に必要なデータを収集統合することが可能となっている。これにより、FDA が安全性を評価する際に、稀な有害事象でも検知できる大規模なデータベースが構築されている。

プログラム実行による匿名化処理：

　医療機関は、各自のローカルデータベース内で、Sentinel Common Data Model 形式に変換することで、患者情報がマスクされるような仕組みになっている。これらの処理はルーチンクエリとなっているため、必要な研究ごとにプログラムを作成する手間がかからない点が特徴的だ。

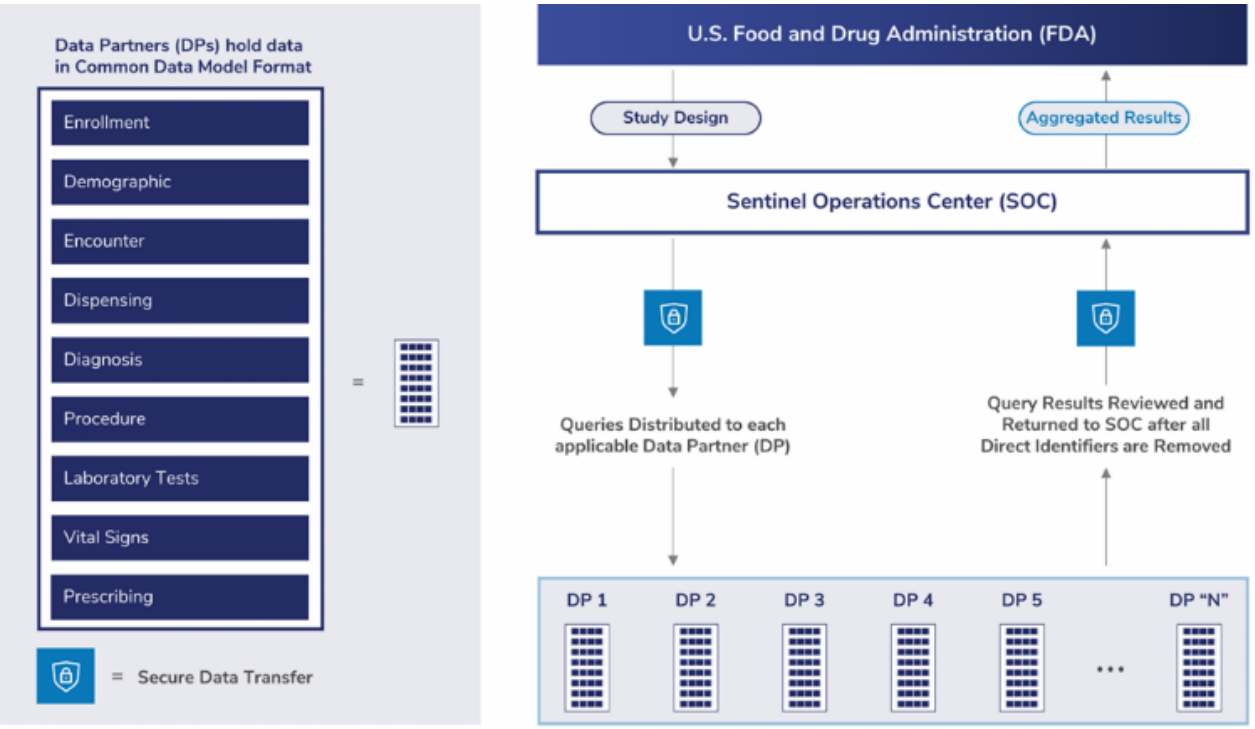


図 13.　Sentinel 分散データベースのしくみ

Sentinel システムに含まれるデータ内容：

患者に関する医療機関での診療行為情報や保険会社との請求情報をベースに、患者がどこの医療機関を受診していたとしても、診断や手技手法（Procedure）、処方箋情報などを得られることが特徴となっている。これらに加え、ワクチンや死亡、がん、その他の薬剤使用や健康アウトカムに関する情報に関する他のデータベースとも連携を図り、分析強化が目指されている。

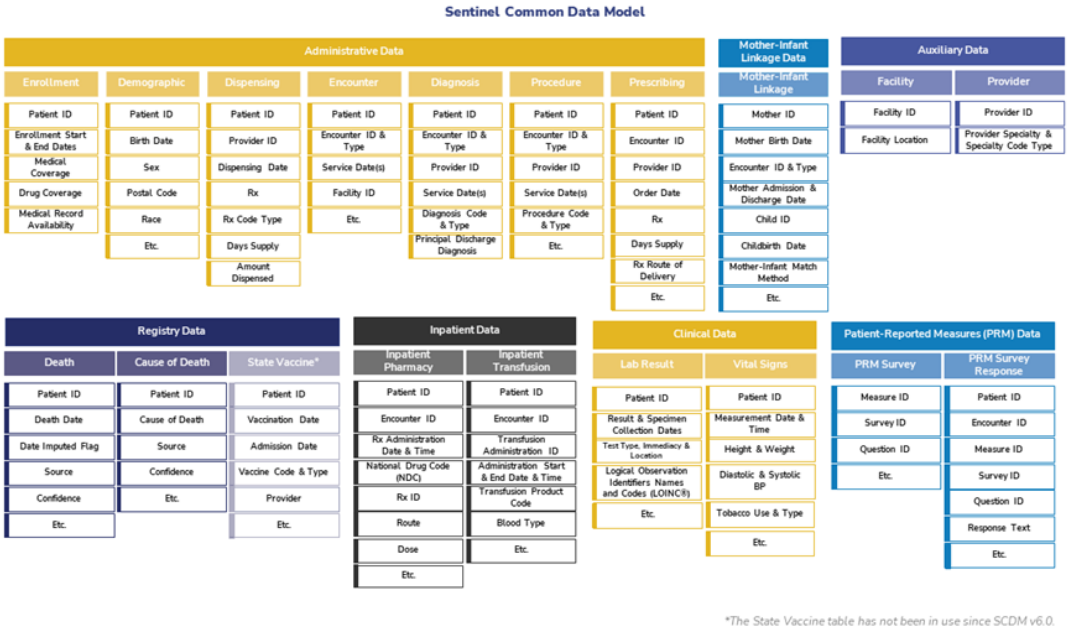


図 14. Sentinel Common Data Model に含まれるデータ

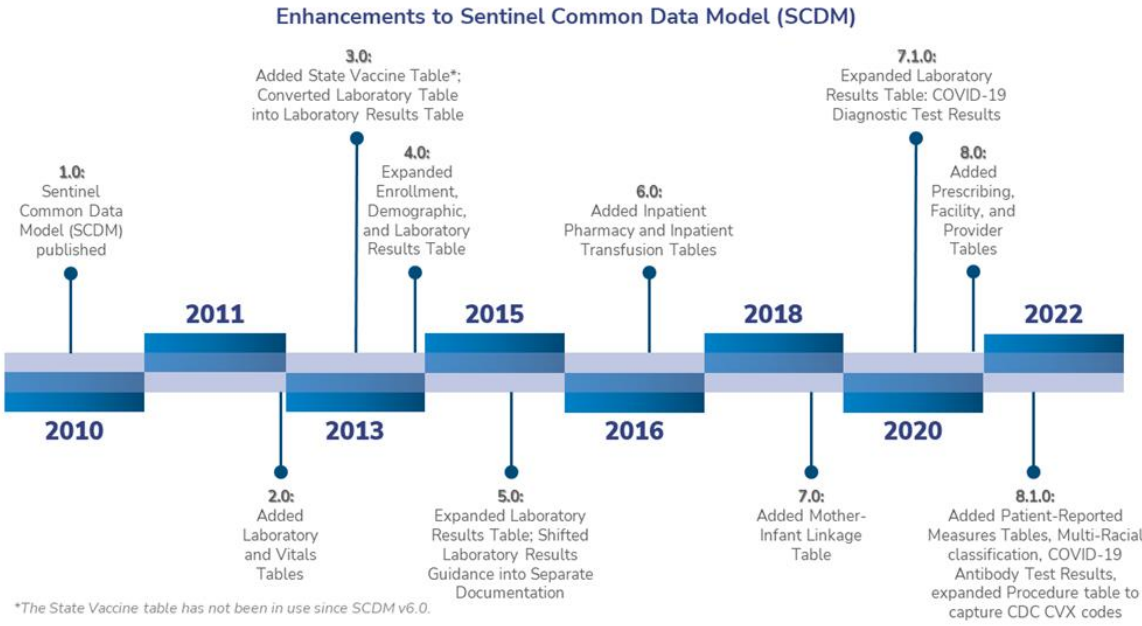


図 15. Sentinel Common Data Model の機能強化の歴史

データ統合手順と品質保証

SOC と各データ提供元が共同で、Sentinel の CDM を的確に運用するためのガイドラインを作成し、抽出や変換作業にばらつきがないようにしている。

まず、データ提供者は、各自のローカルシステム（保険請求システム、電子カルテ等）からデータを抽出する。次に、データ提供者は、Sentinel の CDM の形式に従って変換する。最後に、変換データを SDD に送信する。品質保証は以下の 3 レベルの検証を行っている。

レベル１：完全性と有効性の自動検証…変数や値、テーブル全体の項目の欠落を検証

レベル２：データの意味：テーブル間のデータの一貫性を自動で評価

レベル３：前回抽出データとの比較チェック

レベル 1 とレベル 2 はデータ提供者側で自動プログラムによって行われ、レベル 3 は SOC が実施する。これらすべてをクリアすると Sentinel 分析データとしての使用が承認される。このような検証プロセスを踏むことで、データ提供者側の負担を軽減した上で、長期にわたるデータの安定性と一貫性が担保されている。

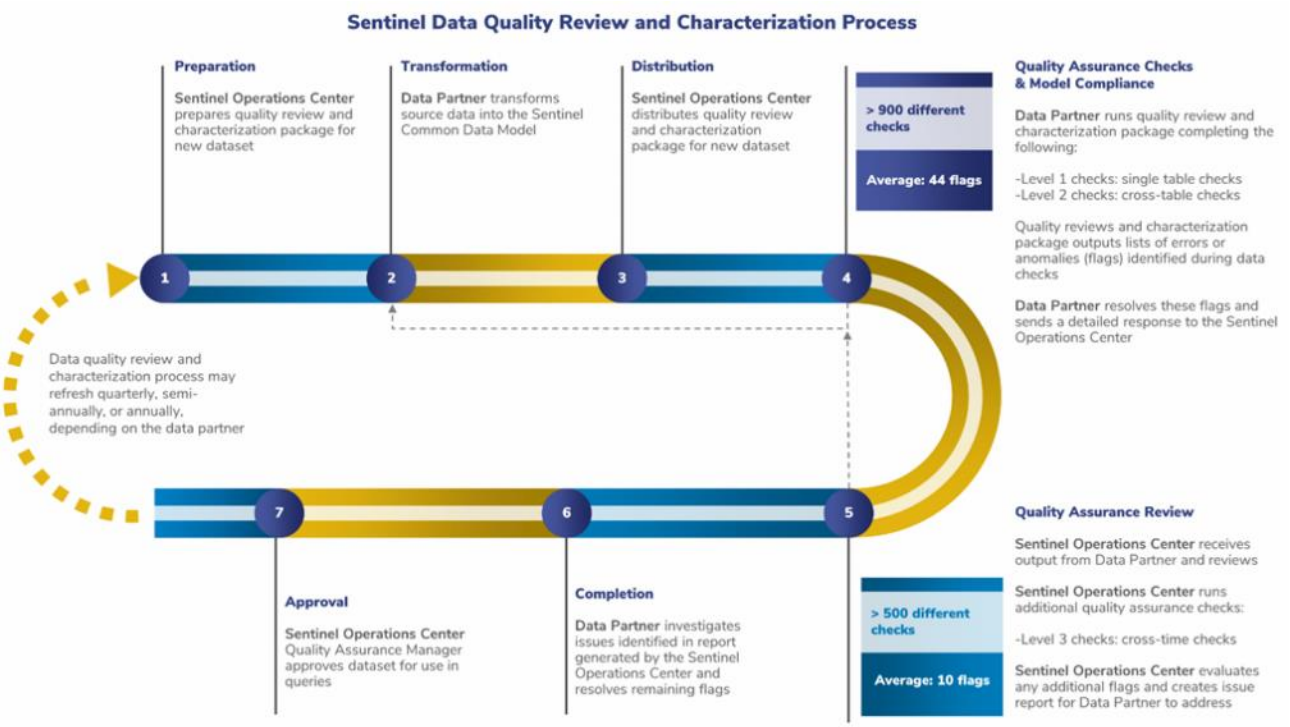


図 16. Sentinel データの質的保証プロセス

■Sentinel システムの普及

Sentinel 分散データベースは 2000 年～2023 年までに合計 4 億 6330 万人の患者 ID が登録されている。保険プランの移動による重複を考慮しても、少なくとも 3 億 4200 万人のデータベースとなっている。

データベース内には、1 億 1290 万人の新規データ、延べ 11 億人年のデータ、197 億件の処方箋データ、202 億円件の個別診療機会（Unique Medical Encounter）、6730 万人の検査結果に加え、800 万件の母子出産データが含まれている。

Sentinel 分散データベースに登録されているメンバー数の推移は下図 17 に示す通り。2019 年に会員数が大幅に減少したのは、メディケイドがデータ終了日を反映したことによるものである。

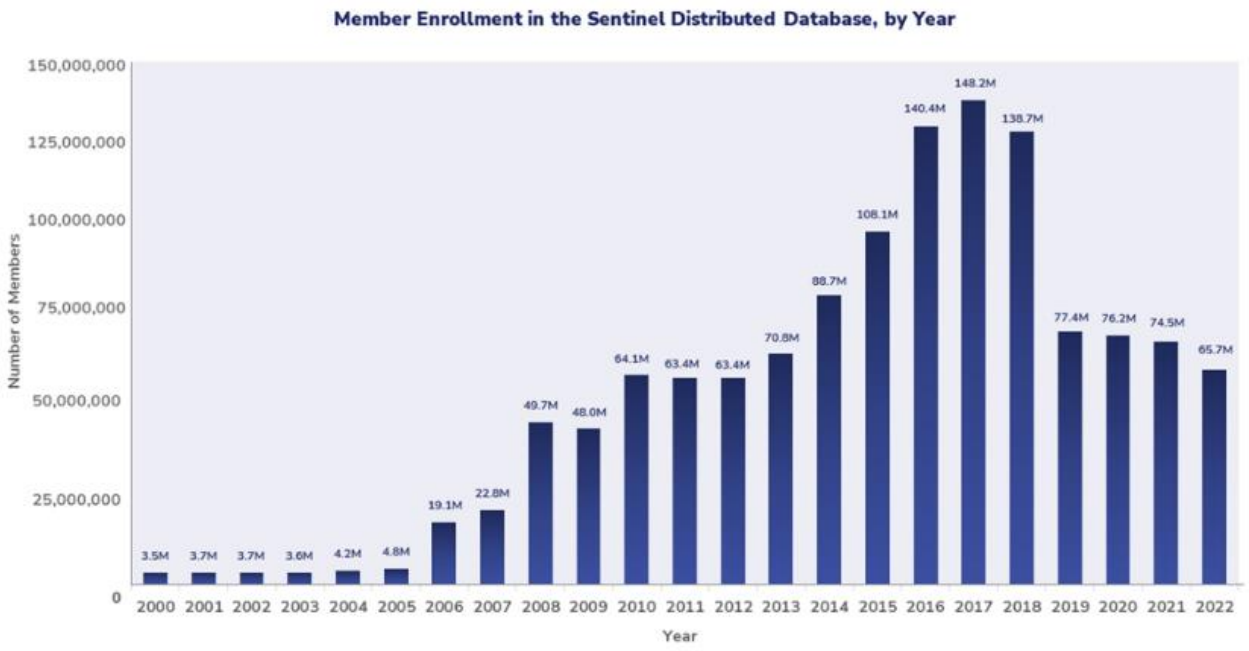


図 17. Sentinel 分散データベースにおける会員数の年次推移

継続登録期間の分布を示したもの。診療と医薬品の双方がカバーされている全 4 億 1490 万人のうち、6 カ月未満が最も多い 3 億 3970 万人で、5 年以上の登録期間を保持する会員も 6200 万人となっている。

図 18. Sentinel 分散データベースにおける累積登録機関別会員数

米国内における地域のカバレッジはやや東部の比重が高くなっているが、性別や年齢によるばらつきは小さい。

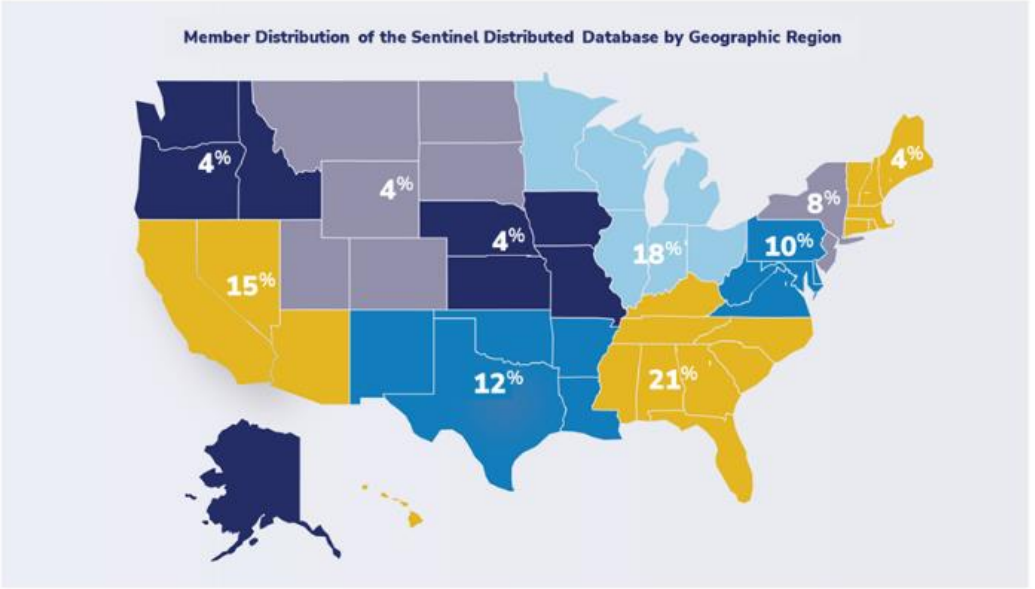


図 19. Sentinel 分散データベース会員の地理的属性

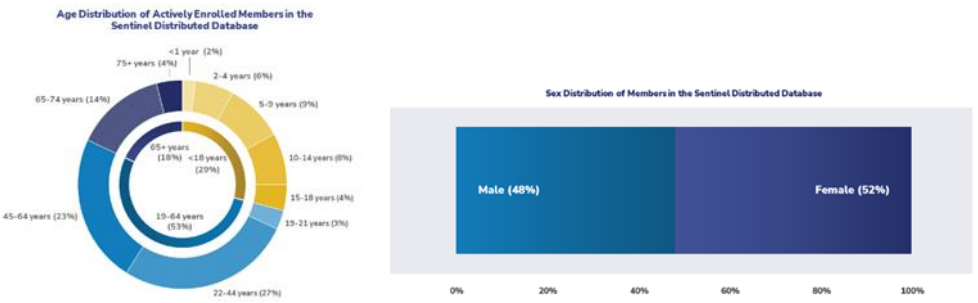


図 20.　Sentinel 分散データベース会員の性年齢別属性

■分析活用事例

分析活用事例として、簡単に 2 例紹介する。1 つ目は COVID19 重症度アルゴリズムのチャート妥当性評価（2024 年 1 月 19 日分析中）で、この研究では統合デリバリーシステム（IDS）における医療記録レビューを使用して、Sentinel システム内で定義づけられている COVID19 疾患の重症度アルゴリズムに対して陽性的中率を用いて評価検証された。2 つ目は 2023 年 7 月に公表された論文で、2020 年 2 月～2021 年 9 月における、米国で COVD-19 に感染して入院した小児が受けた治療とケアに関して、電子カルテ情報から抽出した Sentinel 分散データシステムを利用して行われた後ろ向き記述研究の論文である。慢性肺疾患や先天性心疾患や糖尿病、肥満など重症化に関わる合併症リスクを特定した研究で、小児における COVID-19 の診療実態への理解が深まった。

■課題

CDM としてはデータのテーブル構造は標準化しているが、意味内容の（semantic な）標準化にまで踏み込んでいるわけではない。また標準化の対象としているデータ項目は医薬品の市販後安全性調査という目的に適っている一方で限定的であるともいえる。

PCORnet

■目的および概要

PCORnet は、研究をより安価でより早くより良いものにするために構築されたデータベースのネットワークである。膨大な健康関連データや患者自身のインサイト、専門家の知識などが国主導で長年にわたって蓄積統合されてきた。データを標準化することで、肥満やオピオイド濫用などの公衆衛生上の課題に対する複数施設にわたる大規模な研究もサポートしており「ネットワーク中のネットワーク」と呼ばれている。

■データモデル概要

データソースと研究支援

6600 万件を超える電子カルテ情報や、保険者支払い情報、患者報告データなどから構成される健康情報が、全米で年間 3000 万人分以上蓄積されている。これらのデータは、 8 つの大規模な臨床研究ネットワーク（CRN）の分散型研究ネットワーク モデルを介してアクセスすることができ、患者中心の研究を推進している。これら 8 つの大規模施設（CRN）には、全米の主要な臨床研究者が在籍しており、その集合的な知識と経験をデータネットワーク利用者に提供する支援も行っている。

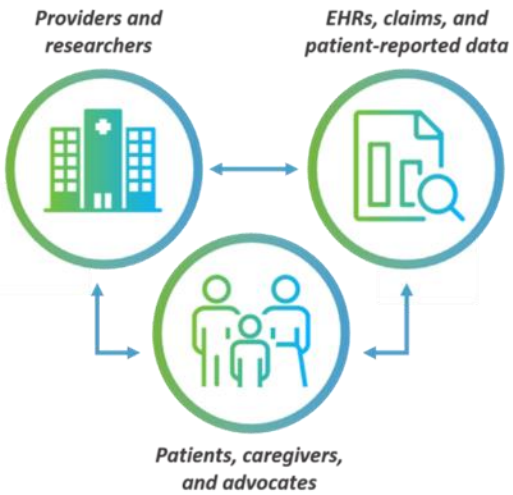


図 21. PCORnet の構成データソース

図 22. 全米における PCORnet の 8 つの研究支援センター

■標準化データと標準化モデルのしくみ

大規模データを有効活用するには標準化が必須である。PCORnet ではデータをネットワークシステム内で共通の 単一言語に標準化することで、より迅速なインサイトの発掘に貢献している。

研究用に即時利用可能なデータ
死亡日、診断、医薬品オーダー、請求書情報、検査情報、人口統計、手技手法
これらのデータは臨床研究から入手可能なデータで、PCORnet の共通データネットワーク内で既に標準化されているため、研究にすぐに使用することが可能。

利用可能だが開発途中のデータ
地理コード情報、健康に関する社会的決定要因、がん登録情報、生検情報、患者報告アウトカム、遺伝子結果、自然言語をベースにした概念用語、患者創出データ
これらのデータは一部の臨床研究で入手可能なデータネットワークのため、PCORnet の共通データモデルと一緒に利用するには、追加作業が必要となるデータ群である

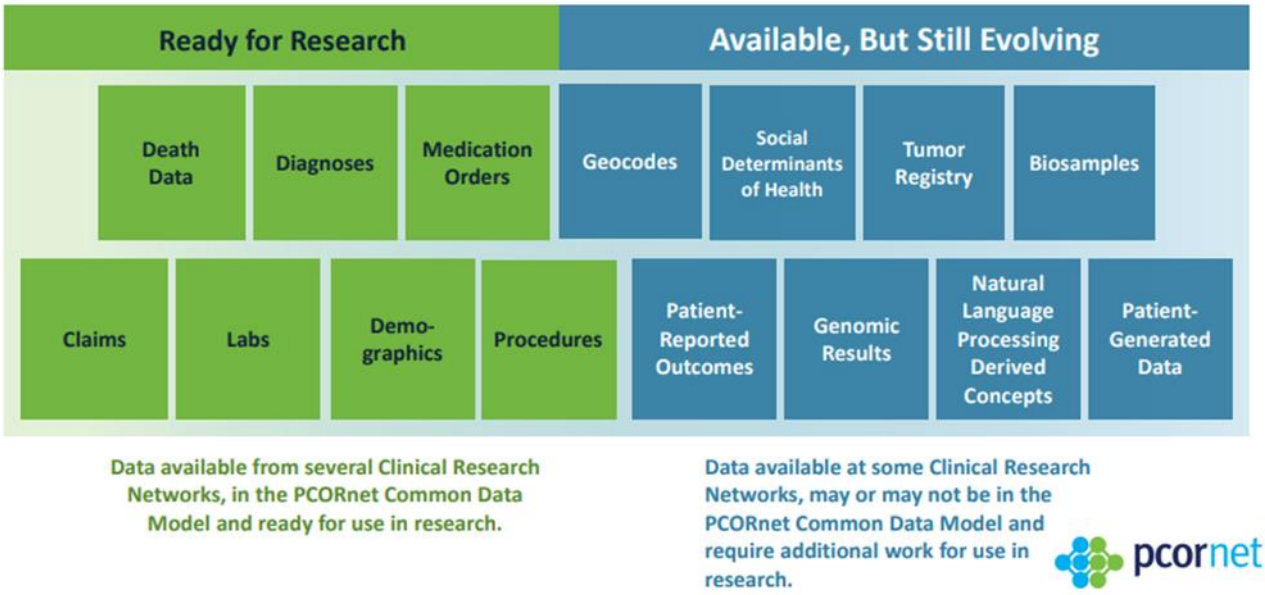


図 23. PCORnet における即時利用と追加作業が必要なデータ種別

2024 年 3 月現在標準化モデルの最新バージョンは Version6.1 で、以下の図 24 にあるテーブルから構成されている。

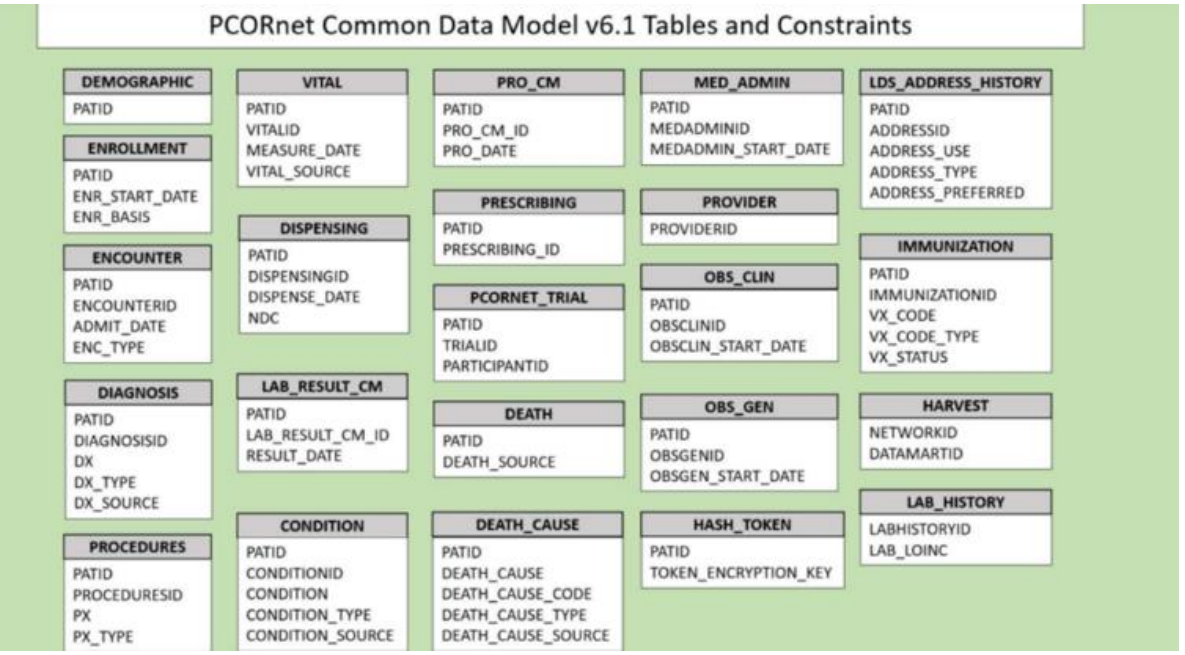


図 24. PCORnet Version6.1 におけるテーブル構成

データクオリティは 2 段階プロセスで厳密にレビューされている。

第一段階では、PCORnet のデータ標準構成と表現に準拠しているかどうかを調べるデータの適合性、診断コードの整合を含む完全性、値が論理的に意味がることを保証するための妥当性、更新の間にデータが消失しないようにするための永続性の 4 つが調査される。

第二段階では、特手の研究目的に対するデータ品質のレベルを検証する。PCORnet 調整センターは、アクセス可能なデータから、特定の研究や当該母集団内における主な変数に対して品質上の懸念を同定しており、研究活動の透明性担保に寄与している。

PCORnet のセキュリティの特徴は、HIPAA 準拠のファイアーウォールで保護されたそれぞれのネットワーク内にデータが留まっており、統合された単一のデータプールやデータウェハースにデータが集められていない点にある。クエリとその応答は安全な分散

型リサーチネットワーククエリーポータル Distributed Research Network Query Portal を介して実施され、役割に応じたアクセス管理や監査など堅個なガバナンスが敷かれているため、必要最小限の情報のみが共有される。クエリはデータに送信されるが、研究者に戻ってくるのはデータそのものではなく分析結果（回答）のみになる。

■普及と利用方法

PCORnet を利用したい場合は、以下のステップで行う。まずステップ 1 では、データ要望者が PCORnet の「Front Door」から質問を送る。次にステップ 2 では、調整センターが質問の内容を精査し、データ要望者に次のステップを案内する。ステップ 3 では、調整センターが要望をクエリに変換し、クラウド上にある安全性の担保されたポータルを介して、ネットワークパートナーへ送信する。最後のステップ 4 ではネットワークパートナーが調整センターからデータ要望者を通じて送られたデータに返信を行う。

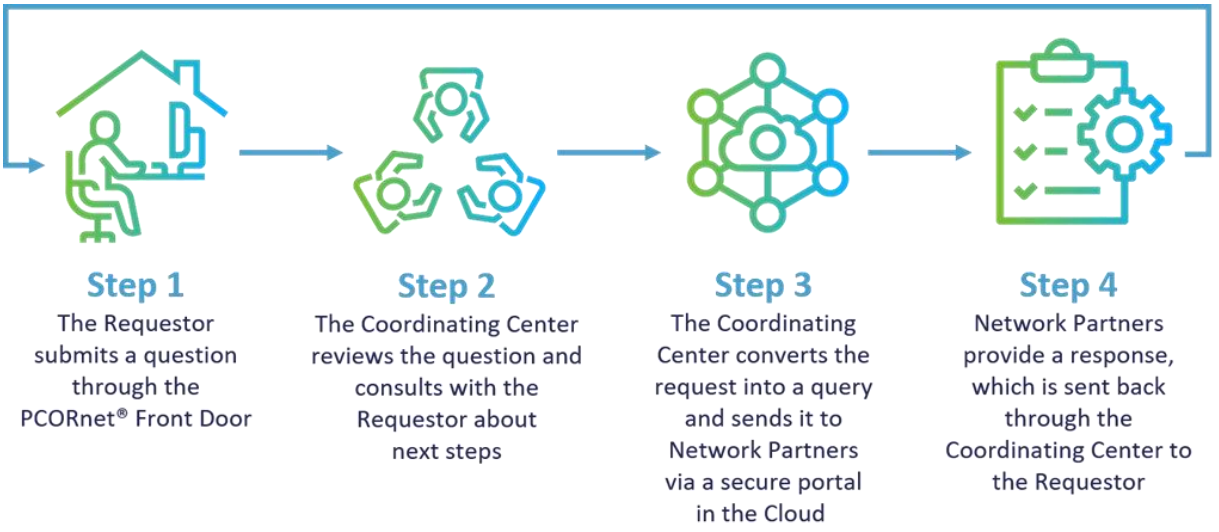


図 25. PCORnet 利用までの手順

PCORnet のデータ年間 3000 万人分で、図 26 にあるように、全米分布図を見ても地域に偏りがないデータであることが分かる。

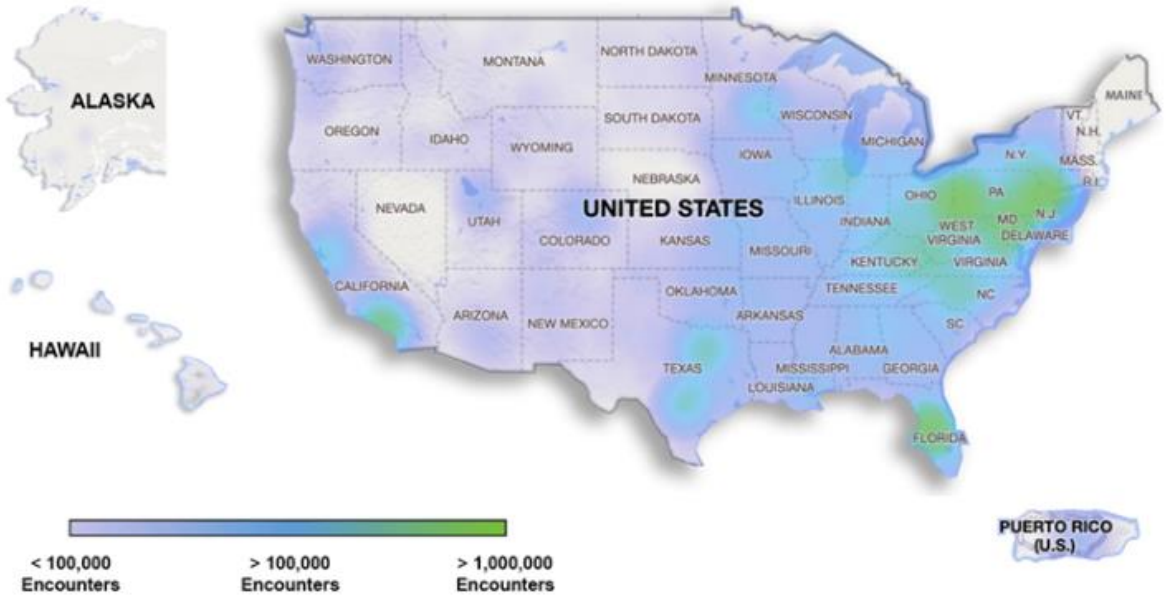


図 26. PCORnet の全米におけるデータ分布図

	Feasibility queries	Identify patient populations for trial recruitment	Coordinated data curation	Connection to patients and providers
PCORnet	●	●	●	●
The ACT Network	●	●	●	●
TriNetX	●	●	●	●
OPTUM	●	●	●	●
OHDSI	●	●	●	●
Sentinel	●	●	●	●

● = capabilities ● = limited capabilities ● = lacks capability

pcornet

図 27. 他のデータベースとの比較表

■分析活用事例

PCORnet はリアルワールドデータ研究や効果比較研究、集団健康調査、医療システムの研究などに活用されている。以下に直近の事例を 3 つ紹介する。

① 予防効果に関する研究

後期高齢者における高脂血症治療薬のベネフィット評価

研究課題：一般的な心血管予防薬の服用は 75 歳以上の認知症を予防に有用か？

研究概要：全米 100 施設における 75 歳以上 2 万人の患者データ、電子カルテ情報、メディケアデータ、電話調査、処方薬調査を用いた実践的なデザインで実施された。

研究への貢献：PCORnet の専門家知識によりニューヨークタイムズや AARP 出版、地方のシニアガイド紙を含む、迅速な患者登録に貢献。PCORnet は研究参加が障壁となるような母集団に対して、実践研究を効果的に支援することが可能となった。

② 適応の評価

アスピリン投与に関する患者中心の長期的な効果評価研究

研究課題：1）効果と出血リスク最小化に最適バランスとなるアスピリン用量は？

2）PCORnet を用いて、患者関与のドライバーとなる臨床試験モデルの回答を見つけることができるか？

研究概要：実践的な臨床試験で、40 施設 15000 人超の心疾患患者が 38 カ月間参加し、2021 年 5 月に New England Journal of Medicine に論文掲載された。

研究への貢献：登録患者への四半期ごとにニュースレターを発行することで 493 人の参加者がパーソナルヒストリーを共有することに貢献した。他にも、学習教材を改訂することで、より深い理解と研究におけるコミュニケーションの活性化や、臨床医の関与による参加率の向上にも貢献した。

③COVID-19 インサイトに関する迅速研究（HERO：Healthcare Worker Exposure Response ＆Outcome プラットフォームの活用）

研究課題：新型コロナウイルス感染症に関する未知の質問やヘルスケア関連機関で働く人への影響などの言及に関して、PCORnet は迅速なコミュニティグループの創出が可能か？

研究概要：ヘルスケア機関で働く 3 万人以上が HERO に登録し、COVID-19 に対するヒドロキシクロロキン群とプラセボ群とのランダム化研究が実施された。新型コロナウイルス感染症ワクチン接種者 20000 人以上の 2 年間にわたる安全性に関する調査報告。

研究への貢献：PCORnet は、コミュニティを迅速に構築し、患者が待つことができない状況下でも迅速にインサイトを届けるのに役立った。

■課題

標準化の対象としているデータ項目が限定的である点、用語の規定が一部の項目に限られている点等があげられる。

3.3 OHDSI と OMOP CDM

前述のようにいくつかの標準データモデルが開発、活用されている。これらはデータの構造や入力規則までは定めているものの、中に入る用語のコード体系までは指定していない。そのような中、データモデルに入る用語（ボキャブラリー）に関してルールを規定することでデータの取扱いや解析を効率的に行えるようにしたデータモデルとして、近年 OMOP CDM に注目が集まっている。本節では OMOP CDM について解説する。

■目的および概要

OMOP CDM は、リアルワールドデータ研究やデータベース調査を容易にするための標準データモデルである*。もともと 2009 年に始まった米国 NIH（National Institutes of Health）、FDA（Food and Drug Administration）および米国研究製薬工業協会(Pharmaceutical Research and Manufactures of America、PhRMA)によるプロジェクトである OMOP（Observational Medical Outcomes Partnership）においてデータベースによる医薬品の安全性調査等を容易にする試みが発端である。その後 EU におけるデータでの検証も経て、OMOP CDM、ボキャブラリー（用語集）、リアルワールドデータ解析の方法論が成果物として得られた。現在はこれをオープンサイエンスコミュニティである OHDSI（The Observational Health Data Sciences and Informatics）が引き継いで発展させており、世界 70 か国以上に広がる取り組みとなっている。なお、成果物はオープンソースとなっている。

■運営組織

前述のように OMOP の取り組みは OHDSI にオープンサイエンスの取り組みとして引き継がれ、2015 年に最初の OHDSI シンポジウムが開催された。OHDSI は観察研究による健康と疾病の包括的な理解が可能となる世界の実現をビジョンとしており、Innovation/Rproducibility/Community/Collaboration/Openness/Beneficience の 6 つをバリューとしている。世界中から Collaborator と呼ばれる協力者がボランティアで参

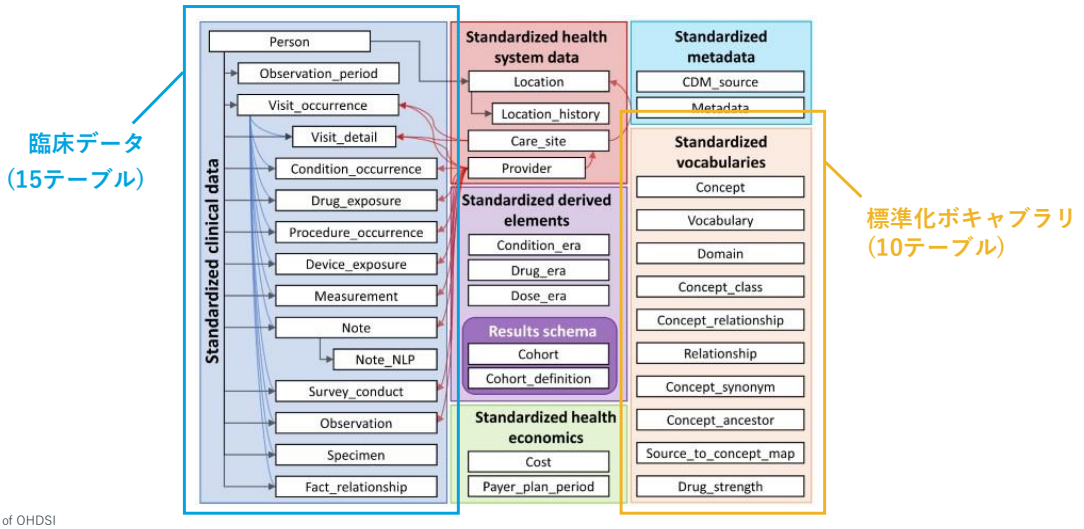
加しており、ベストプラクティスを共有しながらデータ標準、ソフトウェア、方法論研究に取り組んでいる。

■データモデル概要

前述のようにデータは診療報酬請求、診療、研究など多様な一次利用の目的で発生しているため、格納フォーマットやある一つの概念を指定する方法がデータセットによって変わってくる。OMOP CDM は観察研究の質・効率・透明性を向上するためにデータの構造[図 X]、内容、および解析を標準化している。データの標準化の対象は大きく二つあり、データ構造（データモデル）と表現内容（ボキャブラリー）である。

中でも重要な役割を果たすのが、OHDSI 標準ボキャブラリーである。これにより階層化・標準化された医学用語を標準データモデルの各領域に格納し、種々の標準化された転用可能な分析手法の活用が可能となる。標準ボキャブラリーとして指定されていないものも含め登録されているすべての用語には OMOP concept id が定められており[図 X]、標準ボキャブラリーとの対応関係が定められている。OMOP CDM においてはこの標準ボキャブラリーに基づいた concept id を活用することで、データの扱いや解析の効率がボキャブラリーを定めない場合と比べ飛躍的に向上している。

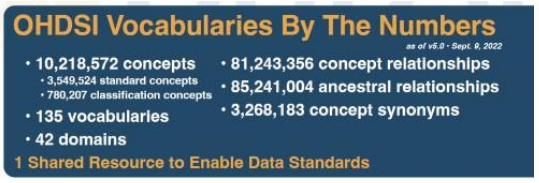
OMOP CDM テーブル構成（ver.6.0）



（出所）The Book of OHDSI
https://ohdsi.github.io/TheBookOfOhdsi/

図 28. OMOP CDM テーブル構成（ver.6.0）

OHDSI標準ボキャブラリーにおけるコンセプトの例



CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	

(出所)OHDSI Our Journey 2022 edition
https://www.ohdsi.org/wp-content/uploads/2022/10/OHDSI-OurJourney-2022.pdf

(出所)The Book of OHDSI
https://ohdsi.github.io/TheBookOfOhdsi/

図 29. OHDSI 標準ボキャブラリーにおけるコンセプトの例

■ソフトウェア・アーキテクチャ

OHDSI では、標準データモデルに準拠したオープンソースのツールを提供している。これらのツールを解析に活用することで、解析にかかる労力の削減と、再現性や透明性の向上という二つのメリットを得ることができる。幅広いツールがオープンに活用でき、仕様やトレーニング内容も公開されているということが OMOP CDM の一つの大きな特徴となっている。

ATLAS：
OMOP CDM のデータに対して解析を設計・実行するためのツールである。ATLAS は OHDSI WebAPI とともに実装される web アプリケーションであり、個人レベルのデータにアクセスする必要があるため通常は各施設のファイアーウォール内にインストールされる。

HADES：
オープンソースの R パッケージのリストであり、OMOP CDM 形式のデータの取扱いから統計解析、図表の作成まで一通りの観察研究をサポートする。
Data Quality Dashboard（DQD）：
OMOP CDM 形式のデータの質の評価を行うツールである。データをテーブルおよびフィールドごとに確認し仕様に合致しないレコード数を調査し、一定の基準を満たすかどうかを表示する。DQD では 3000 以上の項目を確認する。

ACHILLES：
OMOP CDM 形式のデータベースの特徴の要約と可視化を行うツールである。
ACHILLES は R パッケージであり、結果は ATLAS で表示することができる。

ATHENA：
標準ボキャブラリーのリソース集であり、検索やロードを行うことができる。
WHITERABBIT および RABBIT-IN-A-HAT：
データを OMOP CDM に変換する ETL の作成を支援するツールである。
WHITERABBIT でデータ（テーブル、フィールド、値）の内容を調査し、RABBIT-IN-A-HAT でテーブル間の対応など ETL の設計に必要な情報を提供する。

USAGI：
ソースデータから OMOP 標準コードへのマッピングを支援するツール。記載の類似したコードをマッピング先として提示する。マッピングが正しくない部分はユーザーがマニュアルで修正することが可能である。

■活動内容

OMOP CDM を活用した観察研究や、データモデルやボキャブラリーの実装の研究、解析の方法論やバリデーションの研究など、現在までに 550 以上の文献が発表されている(**)

OHDSI は community forum における web 上での意見交換や、定期的な community call の開催による情報共有や議論などを行う環境を整えている。また、毎年 global symposium を開催している。その他 website 等で仕様やトレーニング内容の公開を積極的に行っており、誰でもアクセスすることが可能である。

OMOP CDM は後述のように米国 NIH, FDA や EU の EMA を始め各国で活用が進められている。

■課題

標準ボキャブラリーにマッピングされていることが最大の利点であるが、その分マッピングを含むデータ変換のための作業量が大きくなる。

[参照]

*OMOP CDM（Observational Medical Outcomes Partnership Common Data Model）
https://ohdsi.org/

**OHDSI Book 2023
https://www.ohdsi.org/wp-content/uploads/2023/11/OHDSI-Book2023.pdf?_gl=1*15d57uy*_ga*ODcwNjM0NjY2LjE2ODg2Mjk0NzA.*_ga_BHVF662WPC*MTcwODczNDMyMC4zNy4xLjE3MDg3NDAxMzcuMC4wLjA.*_ga_PNYYSZVRVX*MTcwODczNDMyMC40MC4xLjE3MDg3NDAxMzcuMC4wLjA.&_ga=2.67675153.1641239374.1708736179-870634666.1688629470

3.3.1 各 RWD CDM の違い

i2b2, Sentinel, PCORnet の CDM はデータの構造を指定しているが、用語（ボキャブラリー）や使用するコードの指定を網羅的には行っていない。そのため、解析ツールによって研究ごとにこれらを定義づけする必要がある。解析にかかる労力が大きい一方で、各データホルダー/施設での実装の負担が少ない・ソースデータに含まれる情報が変換により欠損しにくいというメリットがある。

一方で、OMOP CDM では標準的な用語・コードを指定しており、それらをマッピングしたデータベースができあがる。これにより迅速かつ効率的に解析を行うことができる。ただし、ソースデータの質を補うものではなく、またソースデータに含まれる情報をできるだけ損なわずに表現するためにマッピング手法への熟練も要求される。各施設やデータホルダーにおけるデータの OMOP CDM 変換や、データベースのメンテナンスコストがかかるため、持続可能な形での運営が課題である。また、標準ボキャブラリーに沿ったデータベースを用意していくにあたっては、国内でのコード体系が運用レベルで進むことと、日本のコード体系から OMOP の標準ボキャブラリーとして指定されているコード体系へのマッピングが必要となる。OMOP CDM で標準ボキャブラリーとされているものは SNOMED CT、LOINC、Rx Norm といった欧米の標準規格を参照としていることが多く、それらのコード体系は国内では知的財産権を含め活用が難しいものがあることも課題である。

3.4　まとめ　国内における課題と示唆

RWD の有効な利活用は、健康・医療領域の発展において非常に意義が大きいことはいうまでもない。テクノロジーの進化、爆発的なデータ量の拡大に伴い、RWD として考慮されるデータの種類も広がりつつあり、モバイル・センサー技術の発展に伴う連続生体量（継続的に記録される生体関連データ）は象徴的である。一方で、現在まで利活用の主流となっている RWD は、保険薬局ベース（処方箋、薬剤レセプト）、保険者ベース（医科レセプト）、医療機関ベース（電子カルテ、DPC データ）等の官民により提供されている構造化データである。

本章においては、RWD 活用のためには、広さ（量）、深さ（質）、時間、利用プロセスの 4 つの視点が目的に応じてそろっていることが必要であることを紹介した。また、RWE としてのデータ利活用において、4 つの課題、即ち「正規化」「構造化」「個人情報保護」「データ間の比較と統合」があることを解説した。そして、我が国の状況を鑑みた歳には、この 4 つがいずれも今後解決すべきものとして検討余地が大きい。

特に「正規化」の観点では、電子カルテ情報の標準化を掲げ、医療 DX 令和ビジョン 2030 が推進されており、医療情報交換の国際標準規格である HL7 FHIR の導入が活発に議論されている。これは第 2 章で述べた所の「個人」のデータを「個人」のために活用する（いわゆる、一次利用）のための標準化であり、RWD を「集団」として扱うためには、データの用語および規格の標準化が別途必要になる。我が国において「国際的標準（ガラパゴスにならない）」を見据えた「データ間の比較と統合」を考慮した場合に、この点における議論はまだこれからという状況であり、早急に議論を活発化させることが必要である。また健康・医療領域においては「個人情報保護」と「利活用」をどの様にバランスさせるかという視点を常に持つ必要があることも踏まえて、諸外国において諸外国において産官学による RWD CDM の活用の検討が進んでいる。第 4 章にて詳述する諸外国における先進的な取組ならびに関連技術を参照し、我が国における上記課題の現実的な解決の糸口を見つけていくことが望まれる。

4 分散型データ連携に向けた取組や関連技術

4.1分散型データ連携（連合ネットワーク）事例

EHDEN（European Health Data and Evidence Network）

実施国・地域：欧州

概要

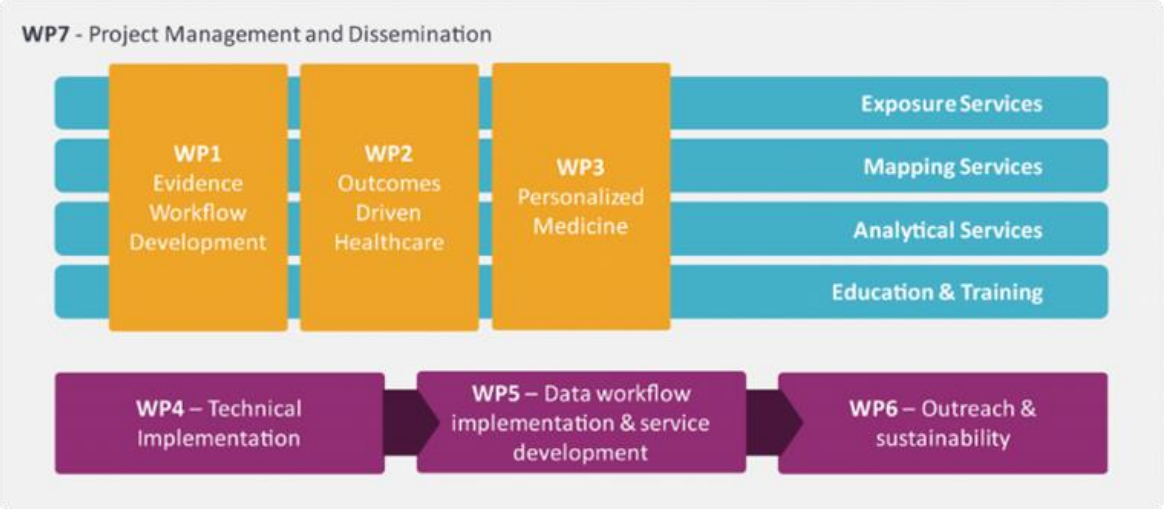
EU と欧州製薬団体連合会（EFPIA）の官民パートナーシップの取り組みである、IMI（Innovative Medicines Initiative）のプロジェクトの 1 つ。データの標準化によって大規模なネットワークをつくり、観察研究のエコシステムを構築してヘルスケアを向上させることを目標とする。「信頼できるオープンサイエンスコミュニティ」をめざす。EHDEN は IMI から約 3000 万ユーロの予算を得て運営されており、EU 各国から 187 の施設・団体が参加している。

設立の経緯および体制、資金

EHDEN は、IMI の Big Data for Better Outcomes (BD4BO)プロジェクトの一環である。2018 年からの 5 年間のプロジェクトで、IMI から約 1400 万ユーロ、EFPIA から約 1600 万ユーロと計約 3000 万ユーロの投資がなされている。

EHDEN は、リアルワールドの臨床データの分析およびエビデンス創出をスケールさせ、患者・医療者・保険者・政府規制当局および産業界が疾患や健康状態、治療、アウトカムをよりよく理解することを助けることを目指し 2018 年に設立された。2018 年 11 月当初、〜〜の 7 つの Work Package からなるプロジェクトとして組織された（図 1）。

EHDENプロジェクトを構成するワークパッケージ



(出所)EHDEN
https://www.ehden.eu/

図 1. EHDEN プロジェクトを構成するワークパッケージ

目的として、データパートナーネットワークの構築、データ/分析基盤の開発、エビデンス創出、および教育を掲げており、OMOP CDM で標準化されたデータにより EU 内で RWD の統合解析ができるエコシステムの構築を掲げる。

EHDEN はデータパートナーの参加をオープンに募り、データの標準化に対して金銭的支援を行っている。一方で、中小企業（small and medium-sized enterprises (SMEs)）に対してはトレーニングを提供し OMOP CDM へのマッピングとエコシステム内でのサービス提供を行えるように支援している。トレーニングを受けた SME に対しては認証も付与している。データの標準化の際に必要となるマッピングに関しては、この SMEs に委託して運営されていくことを期待している。

2024 年で 5 年間の IMI phase2（EHDEN プロジェクト）が終了する。プロジェクト終了後は後継組織として、EHDEN Foundation がオランダに設立されており、プロジェクトの引継ぎとより長期の持続可能なエコシステム発展や研究プロセスの運用に向け

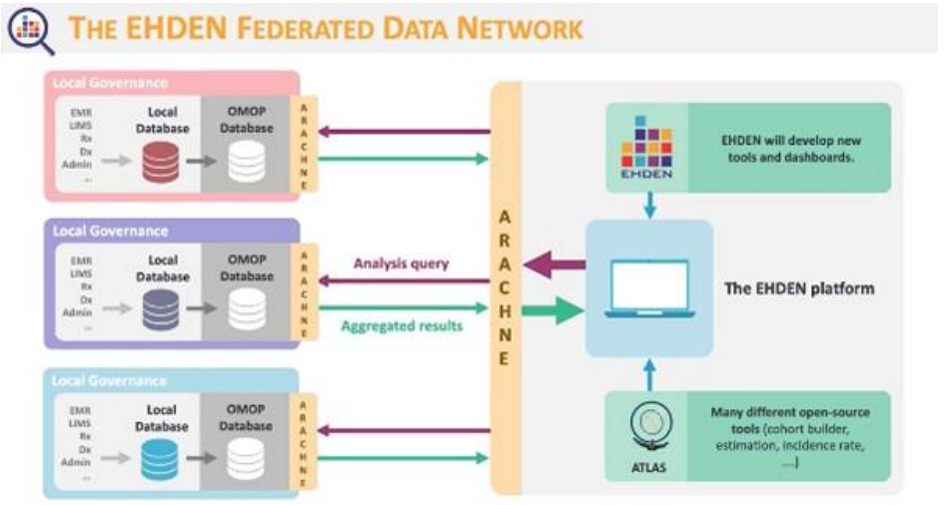
て活動を行う予定である。EHEDN Foundation では引き続き研究支援、トレーニング等を提供するとともに、後述の”study-a-thon”や方法論および技術の開発もすすめていく。

分散型データ連携の手法およびアーキテクチャー

EHDEN では、医療健康に関するデータを OMOP CDM を活用して標準化し、OHDSI の関連ツールの活用を推進しており、ネットワークスタディの仕組みを採用している。データは完全にオーナーのコントロールのもとにおかれるため、倫理面の配慮や各国のプライバシールールに対応することが可能である。

EHDEN Platform ではトレーニングやツールの開発・提供および分析・研究のコーディネーションを行っている。ネットワークスタディの手順は以下のとおりである。解析の際には共通クエリを作成し、各施設に配布する。各施設では施設内でクエリを承認・実行し、解析結果を EHDEN platform に返す。EHDEN platform では各施設から帰ってきた結果を統合解析する。EHDEN では ARACHNE というネットワークスタディを支援するシステムを利用している。（図 2）

ARACHNEのシステム構成



(出所)EHDEN
https://www.ehden.eu/

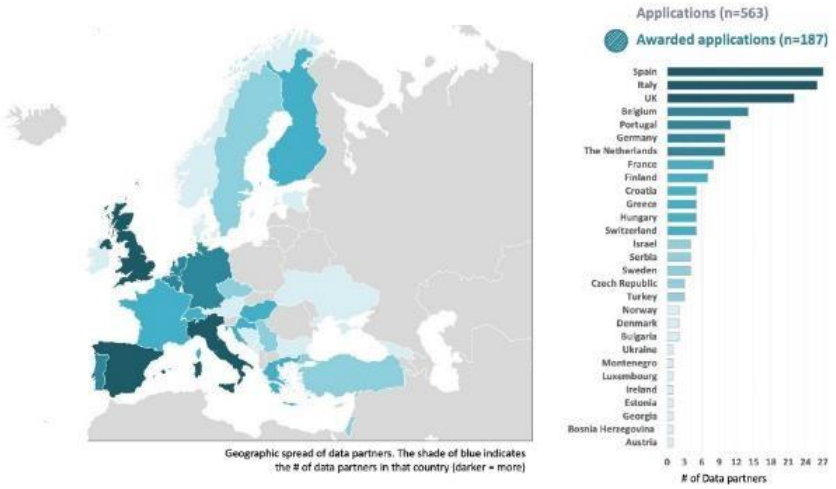
図 2. ARACHNE のシステム構成

活動内容詳細及び成果

・データパートナー

EHDEN プロジェクトでは様々な施設と協力してネットワークの中で OMOP CDM を活用してデータを標準化することを目指している。データパートナーには医療施設や研究施設/プロジェクト、データベースホルダーなどが含まれる。現在 29 か国 187 のデータパートナーを集めており（図 3）、いくつかのデータセットでは実際にネットワーク内で活用できるようにデータを OMOP 変換している。

EHDENのデータパートナー数



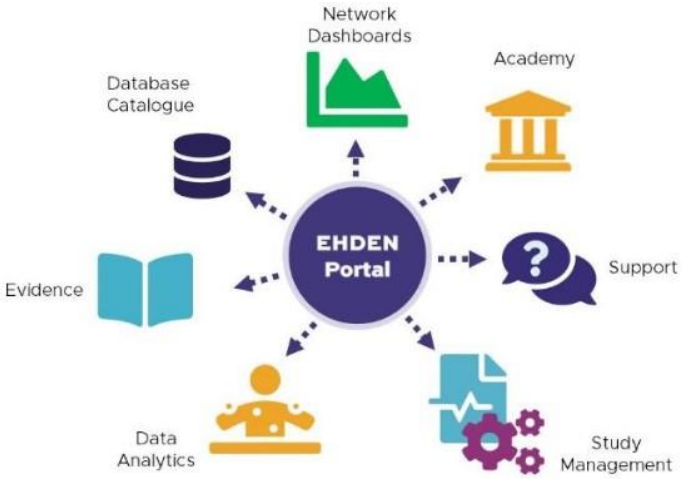
(出所)EHDEN
https://www.ehden.eu/

図 3. EHDEN のデータパートナー数

・EHDEN Portal

EHDEN Portal は、FAIR 原則（Findable, Accessible, Interoperable and Reusable）に基づいてネットワークとデータ分析を促進するアーキテクチャーとしてデザインされている（図 4）。データパートナーと研究者にとって EHDEN における主要ツールとなっており、オープンサイエンスコミュニティの中で調整や承認、契約の前段階でどのステークホルダーが研究に関与すべきかを確認することができると同時に、後続のネットワーク分析や文献作成をスムーズに行うことを可能とする。2022 年 6 月には Data Partner Catalogue をローンチし、データパートナーの一覧がデータの概要とダッシュボードにより容易に参照できるようになった。今後研究アイディアの発案から論文化までのプロセスを標準的な分析パッケージと共に実現できるワークフローを提供予定である。

EHDEN Portalの概念図



(出所)EHDEN
https://www.ehden.eu/

図 4. EHDEN Portal の概念図

・study-a-thon

短期集中で他施設の国際ネットワーク研究を行う取り組み。研究者やデータパートナーが事前に 1 週間ほどかけて集中的に準備をし、完了後の文献化なども含めて行う。トレーニングの機会となるとともに、分析や文献化の過程を見える化することで一種の雛型として類似の研究を展開することも可能になると考えられている。

・EHDEN Academy：EHDEN プロジェクトおよび OHDSI との連携の基礎となる知識を提供するオンライン教育リソース。現在 19 コースを提供しており、これまで 4000 名以上の参加者をサポートしてきた。現状学位や認証の授与は行っておらずサポートは限られているが、RWD/RWE に関わる人向けに広くオープンサイエンスリソースを提供している。

今後の課題と展望

29 か国 187 のデータパートナーが参画しており、また、22 か国 64 の中小企業（SMEs）がトレーニングおよび認証をうけてデータパートナーと継続的なソースデータの標準化に取り組んでおり、これまで OMOP 形式に統一された 8 億 5000 万以上の匿名化レコードが存在する。

2024 年で IMI2 の 5 年間の期限を迎えるため、今後も持続可能な取組とするための財政支援やビジネスモデルの構築が必要であると考えられる。また、より広く産業・医療・政策分野で活用されるためには、継続的に普及活動を行っていくことが必要となるであろう。

[参照]

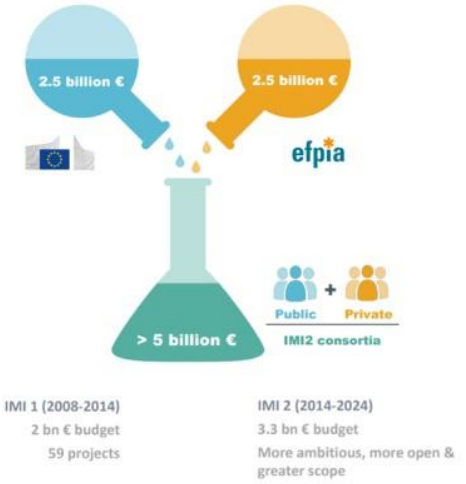
*EHDEN website: https://www.ehden.eu/

**IMI factsheets, EHDEN: https://www.imi.europa.eu/projects-results/project-factsheets/ehden

【参考】IMI（Innovative Medicines Initiative）

EU（European Commission）と製薬業界（EFPIA, the European Federation of Pharmaceutical Industries and Associations）の官民パートナーシップであり、ライフサイエンス領域では世界最大級である。第 1 紀期の IMI1 program (2008-2013）では計約 20 億ユーロ、第 2 期となる IMI2 プログラム（2014 年~2020 年）では、計約 33 億ユーロの予算を確保した[図 5] 。このうち半分は EU の研究に対する投資のフレームワークから、半分は EFPIA および一部他の産業からも拠出されている。

IMIの出資形態



(出所)IMI
https://www.imi.europa.eu/

図 5. IMI の出資形態

大学や研究所等のヘルスケア領域における研究に関わるステークホルダー、製薬その他の産業、中小企業、患者団体、医薬品規制当局のキープレーヤーの間での連携を促進することで、革新的医薬品の開発や患者アクセスの向上を加速させ、健康を向上させることをミッションとする。IMI の目的は、次世代のワクチン、医薬品等の治療を創出し、EU の人々に提供することである。その意味で、産業界の連携はより信頼性が高く迅速な臨床試験やよりよいレギュレーションにつながるし、新規の製品やサービスを生

み出すことにつながると考えられている。以下のような指標を目指し、180 以上のプロジェクトを推進してきている。

・医薬品開発プロセスの改善、POC までの期間の短縮
・診断および治療バイオマーカーの開発
・治験の成功率の向上（特に WHO による優先度の高い医薬品）
・AD や耐性菌の領域など、アンメットニーズの高い医薬品の開発
・新規バイオマーカーでの有効性と安全性の早期チェックを通した、ワクチン候補物質の PhaseⅢでの脱落率の減少

[参照]

*IMI

https://www.imi.europa.eu/

Data Analysis and Real World Interrogation Network (DARWIN EU)

実施国・地域：欧州

概要

DARWIN は、リアルワールドデータから医薬品の使用実態および有効性と安全性に関するタイムリーで信頼性のおけるエビデンスを提供するために European Medicines Agency (欧州医薬品庁、EMA)により設立されたコーディネーションセンターである。DARWIN は、i)データソースカタログの整備、ii)医薬品の使用・有効性・安全性に関する高品質でバリデートされたリアルワールドデータの提供、iii)特定のリサーチクエスチョンに対する質の高い非介入研究の実施、によって規制当局の意思決定をサポートしている。

設立の経緯および体制、資金

もともと HMA（Heads of Medicines Agencies）/EMA Big Data Task Force により提唱されたものであり(*)、Erasmus University Medical Center Rotterdam をサービスプロバイダーとして 2022 に設立された。EMA-HMA Big Data Steering Group workplan(*)および European medicines agencies network strategy to 2025(**)において重要な取組と位置付けられている。

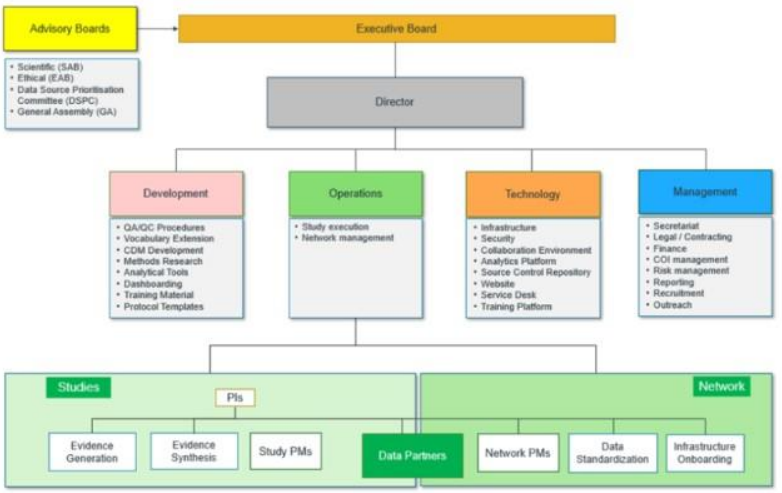
2022~2023 年にかけては DARWIN の体制及びプロセスの構築が行われ、DARWIN のデータを活用した研究などにより EMA および各国規制当局の意思決定の支援およびヘルスケアデータの活用基盤整備の取り組みである EHDS(****)への支援・連携も行われた。2024 年からは通常運用に移行し EMA 及び各国規制当局の評価業務を定常的に支援する予定である。　EMA は DARWIN EU Advisory Board 設置しており、本取り組みの指揮監督および調達契約やコーディネーションセンターの監督も行っている。

分散型データ連携の手法およびアーキテクチャー

DARWIN では The Erasmus University Medical Center Rotterdam が中心となっているコーディネーションセンターが重要な役割を果たしており、分散型データパートナ

ーネットワークの拡大と管理、トレーニング、研究実施に適しているかの評価、データ品質のモニタリングなどをいながら、研究実施にあたって実施可能性の調査から EMA への報告まですべてのステップに責任を持つ[図、https://www.darwin-eu.org/index.php/about/coordination-centre]。また、分析等方法論の開発や EU の関連する他の取り組みとの連携にも携わる。

DARWINの組織体制

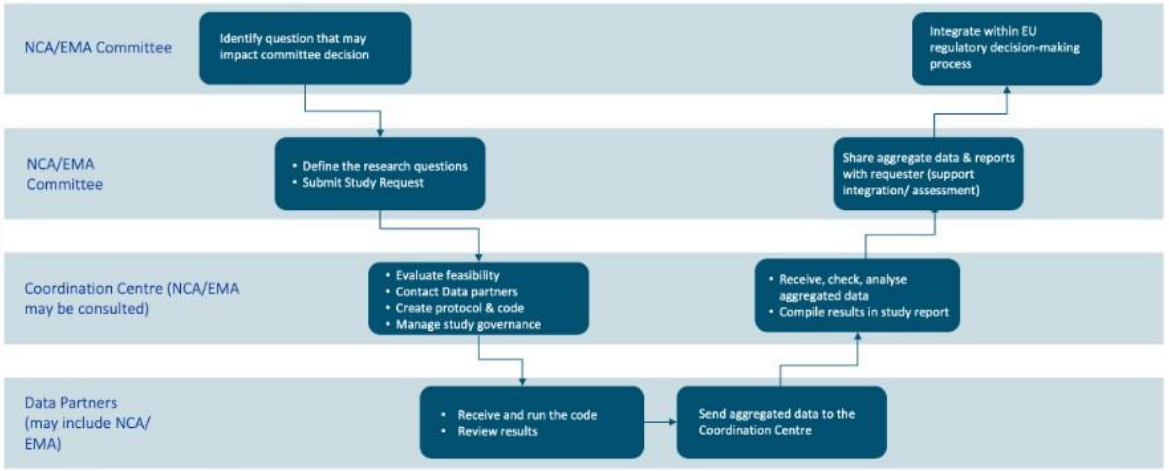


(出所)DARWIN
https://www.darwin -eu.org/index.php/about/coordination -centre

図 6. DARWIN の組織体制

EU Medicines Regulatory Network からエビデンスによるサポートが必要なクエスチョンが提示されると、EMA はリサーチクエスチョンとスタディデザインを設計して DARWIN のコーディネーションセンターに連携する。スタディは、分析コードが各データパートナーに共有され解析結果のみがコーディネーションセンターに共有されるという分散型の手法で行われる[図 7]。また、各国のプライバシー規制に従って追加の規制を研究デザインに組み込むことも可能である。データパートナーは科学的独立性が担保されており、各施設の承認プロセスに則って研究への参加を決定する。

DARWINにおけるスタディの実施プロセス



(出所)DARWIN
https://www.darwin -eu.org/index.php/about/coordination -centre

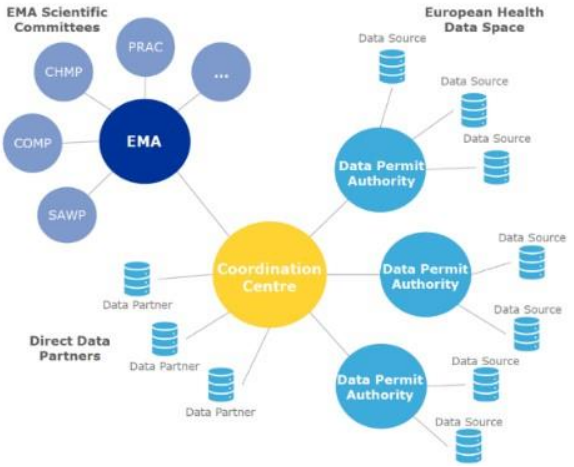
図 7. DARWIN におけるスタディの実施プロセス

データパートナーは研究に必要なデータの解析を DARWIN が活用できるようにプライバシーが保護された形で提供する。データパートナーは直接あるいは複数まとめられてデータネットワークの一部となっている[図 8]。DARWIN のデータパートナー候補開拓にあたってはデータの内容や種類・地理的あるいは属性からみた集団の代表性、同一患者データが一意の ID に紐づいていること、薬剤の用量と投与日が同定できること、臨床イベントが一定の形式で記録されていること、データが CDM に変換されている（あるいは変換されることが予定されている）ことなどが考慮される。2022 年 11 月にはその中から DARWIN EU Advisory Board と協議の下記の基準をもとに病院、プライマリケア施設、保険組合、レジストリー、バイオバンクといったデータをもつ最初のパートナーが選定された。

・継続的にデータ収集を行っている（最低年に 1 回の更新）
・分析や調査に際して 6 か月以内にデータを活用可能
・データが OMOP CDM に変換されている

データパートナーは単に解析結果を返すデータ提供者であるだけでなく、科学的・疫学的な知見の提供や文献化など共著者として研究そのものに貢献することも期待されている。データパートナー側にも DARWIN のサイエンスコミュニティに所属してインパクトの大きな研究に参加するだけでなく、トレーニングを受けることができ、研究の質や効率を上げるといったメリットもある。参加を希望する施設は Expression of Interest フォーマットを提出し、コーディネーションセンターおよび EMA のレビューを経て参加が決定される。

DARWINにおけるデータパートナーネットワーク



(出所)DARWIN
https://www.darwin-eu.org/index.php/about/coordination_-centre

図 8. DARWIN におけるデータパートナーネットワーク

活動内容

EMA により提示されたリサーチクエスチョンをもとにコーディネーションセンターがネットワークスタディを実施する。データパートナーはコードを受け取り自施設のデータベースで実行し、実行結果はコーディネーションセンターに返される。コーディネーションセンターでは統合分析および結果のまとめを行い、EMA に報告する。研究をその複雑さによって”Off-The-Shelf Studies”、” Complex Studies”、” Very Complex Studies”に大きく分類しているが、そのうち定期的に同じコードで実施可能なものは”

Routine Repeated Analyses”とよび定期的に実行されるものもある。既に多くの研究がコーディネーションセンターを中心に実施されており[図 9]、研究の手法や内容は参照可能なように保存されている。

DARWINにおいて実施された分析・研究

Status	Type	EU PAS Register number	Official Title	Last Updated
Finalised	oft-the-shelf	EUPAS50789	DARWIN EU® - Drug utilisation of valproate-containing medicinal products in women of childbearing potential	28/03/2023
Finalised	oft-the-shelf	EUPAS103381	DARWIN EU® - DUS of Antibiotics in the 'Watch' category of the WHO AWaRe classification of antibiotics for evaluation and monitoring of use	28/03/2023
Ongoing	complex	EUPAS103936	DARWIN EU® - Background rates of serious adverse events to contextualise safety assessments in clinical trials and non-interventional studies in adolescent and adult patients with severe asthma	14/03/2023
Finalised	oft-the-shelf	EUPAS50800	DARWIN EU® – Prevalence of rare blood cancers in Europe	28/03/2023
Ongoing	oft-the-shelf	EUPAS105033	DARWIN EU® CC - Multiple myeloma: patient characterisation, treatments and survival in the period 2012-2022	28/08/2023
Ongoing	oft-the-shelf	EUPAS105841	DARWIN EU® CC - Drug utilisation study of prescription opioids	28/08/2023
Ongoing	complex	EUPAS106479	DARWIN EU® CC - EHDS Use Case: Natural history of coagulopathy in COVID-19 patients and persons vaccinated against SARS-CoV-2 in the context of the OMICRON variant	23/10/2023
Ongoing	oft-the-shelf	EUPAS106052	DARWIN EU® CC - Co-prescribing of endothelin receptor antagonists (ERAs) and phosphodiesterase-5 inhibitors (PDE-5is) in pulmonary arterial hypertension (PAH)	28/08/2023
Ongoing	oft-the-shelf	EUPAS105844	DARWIN EU® CC - Use of take-home naloxone for opioid overdose treatment	28/08/2023
Ongoing	oft-the-shelf	EUPAS106759	DARWIN EU® CC - Drug utilisation study of medicines with prokinetic properties in children and adults diagnosed with gastroparesis	23/10/2023
Ongoing	oft-the-shelf	EUPAS106436	DARWIN EU® CC Treatment patterns of drugs used in adult and paediatric population with systemic lupus erythematosus	06/09/2023
Ongoing	oft-the-shelf	EUPAS106885	DARWIN EU® CC – DUS in patients with major depressive disorder	24/10/2023
Ongoing	oft-the-shelf	EUPAS107706	DARWIN EU® Age specific incidence rates of RSV related disease in Europe	5/12/2023
Ongoing	oft-the-shelf	EUPAS107630	DARWIN EU® Characterization of patients with chronic hepatitis B and C	30/11/2023
Ongoing	Complex	EUPAS107615	DARWIN EU® Effectiveness of COVID 19 vaccines on severe COVID 19 and post acute outcomes of SARS Cov 2 infection	28/11/2023
Ongoing	oft-the-shelf	EUPAS107454	DARWIN EU® CC – Natural history of dermatomyositis (DM) and polymyositis (PM) in adults and paediatric populations	13/11/2023

(出所)DARWIN
https://www.darwin -eu.org/index.php/about/coordination -centre

図 9. DARWIN において実施された分析・研究

今後の展望と課題

規制当局が医薬品に関する判断に資するエビデンスを供給するという明確な目的をもって進められている点や、当局から出されるリサーチクエスチョンに対してコーディネーションセンターが分析の実行を行うという明確な役割分担とリソースの確保などが功を奏していると考えられる。

[参照]

*HMA/EMA JOINT BIG DATA STEERING GROUP (HMA)

https://www.hma.eu/about-hma/working-groups/hma/ema-joint-big-data-steering-group.html

** Big Data (EMA)

https://www.ema.europa.eu/en/about-us/how-we-work/big-data

*** Network strategy to 2025

https://www.ema.europa.eu/en/about-us/how-we-work/european-medicines-regulatory-network/european-medicines-agencies-network-strategy

****EHDS

https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

The Digital Oncology Network for Europe（DigiOne）

実施国・地域：欧州

概要

DigiOne は The Digital Institute for Cancer Outcome Research (DIGICORE)の傘下で実施されている複数のプロジェクトのうちのひとつであり、がん領域で連合型データ連携を志向している。

DIGICORE は現在 34 の主要ながんセンター、フランスおよびイタリアの 2 つのがんネットワーク（Unicancer、Alliance Against Cancer）、および 2 つの民間企業（IQVIA、Illumina）からなる官民パートナーシップであり。OECI (Organisation of European Cancer Institutes)(*)との協議を経て、European Economic Interest Grouping (EEIG) (**)として 2021 年ベルギーに設立された。参加施設組織での相互運用性の上にデジタル研究基盤を構築することを目的としており、がん診療施設が臨床試験の自動化、アウトカムリサーチ、診療の質マネジメントのデジタル化などに向けて EHR（electronic health record）や分子遺伝学的診断情報の活用を行えるように実装していく取り組みである(***)

設立の経緯および体制、資金

DIGIONE は DIGICORE により資金拠出されたプロジェクトであり、欧州の 6 つの医療機関が参加して 2023 年 1 月に開始された。

分散型データ連携の手法およびアーキテクチャー

DigiOne では欧州 6 か国 6 つの施設から、日常的に収集できる診療及び分子診断情報を活用するための研究ネットワークである。Minimal Essential Description of Cancer (MEDOC)というコンセンサスによって定められた必要最小限の診断・バイオマーカー・治療・アウトカムのデータセットに従って全患者の情報を活用可能にする[図、https://doi.org/10.1038/s41591-023-02715-8]。これらのデータセットで診断からアウトカムまでの記述に必要な最低限の情報及び主要な研究に必要な適格基準の情報をカ

バーしており、質の高いリアルワールドエビデンス及び診療の質マネジメントのための分析データとなる。MEDOC はオープンイノベーションコンペティションの形式で、16 のがん診療施設によりデータ活用における重要性と各施設での実装可能性の観点から設計された。各施設におけるデータ抽出を含めた連合型ネットワークのためのソリューションに関してはオープンイノベーションチャレンジの形式で各施設によりプロトタイプ構築が行われた。その結果選定された 6 施設により DigiOne のパイロットが進められている。

MEDOCで定められたデータセット

Table 1 | Data concepts in MEDOC

Area	MEDOC concept
1. Demographics	1.1 Date of birth (month) 1.2 Sex 1.3 Weight (with timestamp) 1.4 Height 1.5 Healthcare ID (or other unique identifier) 1.6 Legal basis for data processing
2. Clinical phenotype	2.1 Primary cancer diagnosis and comorbidities, typically in International Classification of Disease standards such as ICD10, ICD9 or ICD-O-3 2.2 Charlson comorbidity index (derived from 17 comorbidities in 2.1) 2.3 Date of primary cancer diagnosis 2.4 Method of primary cancer diagnosis 2.5 Performance status (for example, coded by ECOG or Karnofsky standards) 2.6 Disease stage in a recognized standard such as TNM 2.7 Histological cell type, typically in ICD-O-3 standards 2.8 Menopausal status (for example, for patients with breast cancer)
3. Biomarkers	3.1 Biomarker name 3.2 Biomarker measure 3.3 Biological sample ID
4. Treatment	4.1 Line of therapy (derived algorithmically within each cancer type) 4.2 Anti-cancer treatment name, including systemic treatment and supportive therapy 4.3 Molecule generic name 4.4 Start date for drug treatment 4.5 Treatment dose 4.6 End date for drug treatment 4.7 Radiotherapy type 4.8 Radiotherapy start date 4.9 Radiotherapy dose 4.10 Radiotherapy end date 4.11 Surgery type 4.12 Surgery date 4.13 Participation in clinical trial 4.14 Date of trial consent
5. Outcomes	5.1 Date of death, at any location 5.2 Time to next treatment (derived from treatment start dates) 5.3 Metastasis presence/absence 5.4 Metastasis location 5.5 Date of clinical visits (with cancer related visits separated from other visits) 5.6 Vital status (derived from visits or death linkage) 5.7 Extent of debulking surgery (for example, for patients with gynecological cancer)

Note: implementation of 1.5 and 5.1 is influenced by national regulations; 2.8 and 5.7 are essential only in some cancers for which risk-normalized audit or research cannot take place without their capture.

(引用) Mahon, P., Chatzitheofilou, I., Dekker, A. et al. A federated learning system for precision oncology in Europe: DigiONE . Nat Med 30, 334 –337 (2024).

図 1. MEDOC で定められたデータセット

前述の OMOP CDM の活用に加えて自然言語処理（NLP, natural language processing）技術によるより広範なデータ取込を行っている。そのため、がん領域の日常診療において得られる質の高いリアルワールドデータを、プライバシーを保護した形で分析に活用することが可能となっている。

活動内容

教育、啓発活動としては DigiOne meeting を開催したほか、DGICORE Masterclass として連合型分析に関するレクチャー/トレーニングを提供しており、YouTube channel でも公開されている。

今後の展望および課題

現在パイロット中であるが、データセットを実装可能性も考慮して必要最小限のものに絞っているところが特徴的である。実装コストを抑えて連合学習のモデルを実現する一つの方策となる可能性がある。長期的には DIGICORE からの資金拠出がなくても多くの施設にシステムを導入しネットワークを拡大していくために、いかにエコシステムを構築できるかが課題となる。

[参照]

*OECI
https://oeci.eu/
**EEIG
https://eur-lex.europa.eu/EN/legal-content/summary/european-economic-interest-grouping.html
***DIGICORE
https://digicore-cancer.eu/

Feeder-Net（Federated E-health Big Data for Evidence Renovation Network）

実施国・地域：韓国

概要：

韓国での連合データ活用の取り組みとして、Feeder-Net（Federated E-health Big Data for Evidence Renovation Network）の事例を紹介する。Feeder-Net は、韓国国内の医療機関が保有する EMR データを OMOP-CDM を用いて標準化する事をゴールとした公共医療情報交換（HIE）プラットフォームであり、規模としてはアジア最大の RWE プラットフォームと言われている。

Feeder-Net プロジェクトは韓国の貿易・産業・エネルギー省（MOTIE; the Ministry of Trade, Industry &Energy）の主導により、2018 年から直近 3 年間予算を約 1,000 万 US ドルとして開始された。（*）

2019 年以降は Feeder-Net+に名称を変え、継続中の Feeder-Net プロジェクトでは 41 の医療機関における共通データモデル（OMOP-CDM）への変換、データネットワークの構築、それに伴うサービスの PoC の実行などを目標として予算が割り当てられた。すべての CDM 変換プロジェクトに政府から支出された予算は合計約 4,300 万 US ドルで、内訳は以下の通り。

プロジェクト名／時期	目的	担当省庁	予算額（US ドル換算）
Feeder－Net（2018 年から継続）／2018-2020	• CDM 変換（41 医療機関） • Feeder－Net プラットフォームの開発 • PoC プロジェクト	産業省	1,000 万 US ドル
Feeder-Net 拡大／2019-2022	• CDM データネットワーク拡大 • Feeder－Net プラットフォーム改良	産業省	1,000 万 US ドル
企業向けサービス開発／2019-2022	• 2020 年以降合計 6 案件への投資	産業省	700 万 US ドル

臨床試験プロジェクト／2019-2022	• 2020 年以降合計 20 件の共同試験	保険省	1,400 万 US ドル
標準化プロジェクト／2019-2022	• CDM 拡大モデルの開発 • メディカルコードのマッピング • 匿名化加工のガイドライン作成	産業省	150 万 US ドル
プライバシー―保護／2019-2021	• CDM に関連するプライバシーガイドラインの開発、合計 10 プロジェクト	保険省	50 万 US ドル

活動内容詳細及び成果：

Feeder-Net の取り組みによる OMOP-CDM 推進プロジェクトの成果として、2023 年時点で、61 の病院と 1 軒のクリニックがネットワークに参画している。これは医療機関全体の 70%となり、患者レコード数は 9,700 件を超えている。
また、12 の医療機関では EMR の CDM 変換を完了している。

今後の課題：

韓国の Feeder-Net プロジェクトの成果としてネットワークに参加する医療機関は増えたものの、いくつかの課題が残っている。一つ目は運用資金の確保の問題である。Feeder-Net プロジェクト運用期間に医療施設側は OMOP 変換に係る費用を負担していない。Feeder-Net イニシアティブでは政府からの資金提供の大半が、病院と企業に OMOP 変換に係る費用の支払いに充てられた。今後の運用には、ネットワーク維持のためにより多くの政府助成金、または参加医療施設・企業からの投資が必要となる事が考えられる。すでに一部の医療施設ではリソース不足を理由に OMOP データベースの更新を停止している。二つ目の課題は、商用利用の活性化である。商用目的でのネットワークの使用は学術利用より低い傾向である。この取り組みには当初ネットワークの商用利用に関するマイルストーンがいくつか含まれたものの、ネットワーク活用を希望する企業の高い期待と病院の保守的な性質との間に大きなギャップが存在している。現在進行中の唯一の商業的使用は研究用であるが、病院は商業的研究の申請を評価する際により厳格であり、これも学術研究ほど活発ではない。

[参照]

* https://www.oecd-ilibrary.org/sites/c4e6c88d-en/index.html?itemId=/content/publication/c4e6c88d-en&_csp_=2e36e9568b34042ffecfde8809e7fd98&itemIGO=oecd&itemContentType=book#section-d1e121

4.2 分散型データ連携関連技術

バイオデータ連携・活用のデータ連携関連技術の現在および今後の大きな流れは以下の 3 つのカテゴリが中心になっている。

（A） 分析の技術（人工知能、AI、自然言語処理など）

（B） 分散型のデータ連携の技術 （データ分散保持、データ標準化など）

（C） データプライバシーの技術（匿名仮名加工、連合学習など）

この章ではこれらの技術について説明する。

4.2.1 分析の技術（人工知能、AI、自然言語処理など）

バイオデータ連携・利活用で使われる分析の技術は多岐にわたり、分析の手法は数えきれないほどの数が存在している。その中で重要な概念である人工知能・機械学習・深層学習、自然言語処理、その他の重要な分析手法の概略の説明をする。

人工知能（AI）・機械学習・深層学習

人工知能（Artificial Intelligence, AI）とは人間の思考結果を模して機械が思考結果を出力するプログラムである。簡単な例をあげる。ある人の身長のみを入力データとして体重を人間が推定・回答するというプロセスを数式などでプログラムすることで、同じ入力データを使って機械が人間の回答に準じた結果を推定・回答することができる。
人工知能は包括的な概念であり、機械的に人間の知能のようなプロセスを実行するアイディア全般を含んでいる。人工知能のうち、人間の学習に相当する部分をコンピューターープログラムが自動で特徴の学習をするものを機械学習と呼ぶ。例として、不良のネジと正常なネジを見分ける作業において不良ネジの特徴を人間が定義せずに機械に学ばせるなどがある。

機械学習の手法の一つに深層学習（ディープラーニング、DL）がある。ニューラルネットワークと呼ばれる特定の機械学習の手法のうち、中間層と呼ばれる層が特定の層数以上の条件で学習した状態が原理的にはディープラーニングと呼ばれる。入力データと出力データの表面的な関係だけで予測するのではなく、複数の入力データが示唆する表面的には見えないような特徴を機械が自動的に特定して結果を予測する。通常の機械学習よりも深層学習の方が一般的に高い予測精度になるポテンシャルが大きいが、深層学習は分析コスト（データ収集コスト、分析作業者コスト、分析処理時間コストなど）が大きくなることが多い。

分析の手法・分析処理の例

分析には様々な手法が存在する。この項ではバイオデータ関係の分析手法をいくつか紹介する（自然言語処理、画像処理、異常検知、時系列解析）。下記は全て機械学習で

実装されることが多い。なお、分析手法はどれか一つだけ使用して実装することは稀であり、たいていは複数の手法や概念を統合して使う。

自然言語処理

自然言語処理とは人間が日常で使用している言葉や文章のテキストデータや会話音声をコンピュータで処理する技術である。自然言語処理系の処理・分析は大きく 2 つのカテゴリに分かれる。非構造化データ処理と構造化データ処理である。構造化とはデータをテーブル形式などの機械が処理しやすい形に整形する処理を意味する。非構造化データとは人間が日常で使用している言葉を文章のまま保存しているデータを指し、非構造化データ処理とは文章データをそのまま分析処理する技術である。

前者の非構造化データ処理系の自然言語処理ではテキストデータを単語などの小さい単位ではなく文章・段落・章のまま分析処理する。非構造化データ処理の多くは語順考慮ありで処理・分析する。非構造化データ処理では少量のテキスト（通常は数千文字程度まで）を深く処理・分析するのに用いられ、2023 年から話題に上がっている ChatGPT を含む生成 AI もこちらの非構造化データ処理の自然言語処理に分類される。非構造化の自然言語処理の機能として文章要約、文章抽出（ユーザー指定のトピックの記載がある部分の抽出）、推論、文章生成、感情推定などがある。非構造化の自然言語処理はコンピュータによる処理時間が長くかかり、数百文字の処理に数十秒から数時間かかることも多く、大量のテキストデータ処理・集計との相性は悪い。その反面、文章要約、文章抽出、推論、文章生成などの深い処理・分析ができる。

後者の構造化データ処理系の自然言語処理ではテキストデータを処理する単位が単語などの小さい単位または単語や数値をテーブル型にした単位で実施される。構造化データ処理の多くは語順無関係で処理・分析する。文章の意味や単語をベクトル（数値の集まり）に置き換えて分析処理する。構造化の自然言語処理の機能として、文章同士の類似度計算、文章のトピック抽出、トピックのクラスタリング、感情推定などがある。構造化の自然言語処理はコンピュータによる処理時間が短く、億単位以上のレコード（エクセルのセルに相当して文章を格納しているレコード）などの大量のテキストデータ処

理・集計との相性が良い。その反面、文章要約、文章抽出、推論、文章生成などの深い処理・分析には不向きである。

自然言語処理の用途

バイオデータ連携・利活用において自然言語処理の活用範囲は広い。例えば数億件の消費者のコメントの感情をコメント単位で機械に推定させたり、コメントで頻出するトピックを抽出したり、コメント全体の要約をさせるなどが可能である。また企業側から消費者へのメッセージを自然言語処理で生成や、メッセージを消費者に出す前に複数観点からメッセージを評価させることも可能である。

自然言語処理とデータ連携

データ連携・利活用を実施するには自然言語処理の標準化が必要となる。各施設（またはデータを保持する各団体）間でデータ保持する決まりが標準化されていればそれらの施設のデータを連携や、同じ観点で比較するなどの利活用が可能となる。今後数年でChatGPT のような生成 AI が急速に発展することが予想され、処理時間が長くかかる生成 AI などのデータ処理は各施設で実施してその要約情報のみが施設外に共有されていくと思われる。前述のように生成 AI の処理は重く、中央集権的にすべてのデータを一か所で処理することは考えにくい。連合学習のように施設毎に分散して分析の処理（生成 AI 処理、自然言語処理）を実施して分析結果要約を中央の分析機関などが収集する形になると思われる。

画像処理

画像処理とは静止画や動画をコンピュータが処理して画像を編集・生成したり情報を取り出したりする技術である。バイオデータ連携・利活用において画像処理技術の活用範囲は広い。農産物の個体ごとのスコア化・分類、農作地域のスコア化・分類、医療画像のスコア化・判定、製品の欠陥検査、製品の文字情報の読取り、人間の動きの読取りを使った調査やアラートシステム、などが考えられる。

画像処理とデータ連携

画像処理のデータ連携・利活用の一例として医療系の画像診断が進んでいる。レントゲン写真や皮膚の写真などと人間（医師）による疾患診断のマッチングパターンを機械学習で学習させ、機械に特徴をつかませる。この際に各施設内では疾患ごと患者数が少なく機械学習の精度が低迷するが、複数施設で画像データを連携することで機械学習の精度が高まる。画像処理のデータを施設間で標準化し、連合学習などでデータ連携が進められる。

異常検知

異常検知は分析のカテゴリの一つである。バイオデータで正常ではない個体などを見分ける際に用いる。トマトなどで一定の品質基準を保ちたい場合などに用いられ、その場合には正常なトマトと異常なトマトで分類する課題として機械学習を利用することができる。また、異常として分類される教師データが少ないまたは皆無の場合には正常状態を定義しておき、正常状態から極端に離れた場合に異常検知が反応する仕組みを機械学習で構築することも可能である。

時系列解析

時系列解析は分析のカテゴリの一つである。時間的な変化を解析・予測する際に用いられる。イチゴの熟し具合の変化などを解析し、消費者に最高の状態で届けるために機械学習を利用することができる。医療分野であれば、一人一人の疾患の時系列予測などにも時系列解析の機械学習を利用できる。

データ標準化

上記いずれの分析手法の場合にもデータ連携・利活用という観点から見るとデータ標準化や連合学習という枠組みが有効である。現時点で考えられる主なデータ利活用方法（異常検知、時系列解析など）を想定してデータ連携の方式（OMOP の利用、連合学習の利用など）を定めていくのが順当と思われる。

4.2.2 分散型のデータ連携の技術 （データ分散保持、データ標準化 など）

この項ではデータを一か所ではなく複数箇所に分散して保持し、必要に応じて分散したデータを連携して活用する技術について説明する。

分散型データ分析基盤

分散型データ分析基盤とは複数のチーム（企業・施設・部署など）がデータをつないでデータを利活用する仕組みである。分散型データ分析基盤と対極にある考え方が統合型データ分析基盤で複数のチーム（企業・施設・部署など）のデータを一括して一か所に保存する考え方である。データ連携する複数チームが同一社内であれば統合型データ分析基盤は現実的であり、データの取り回しもしやすく、社内の別部署のデータにアクセスして新たな利活用を生み出しやすい。また一社内での複数部署からのデータアクセスであればデータセキュリティもコントロールしやすい。
一方でデータ連携する複数チームが複数の企業・病院などにまたがっている場合は分散型データ分析基盤と相性がよく、各チームのデータはそれぞれのチームで管理・保持し、データ連携の協定を組んでいる他チームのデータの詳細を見ることはできないものの統計処理した代表値などを見ることができる。
技術の面でいうと、分散型データ分析基盤とは複数の異なる種類のデータベースをつなぎデータをやり取りする仕組みである。データベースを持つ施設のことをデータセンターと呼ぶ。データセンター間のデータのやり取りはセキュリティで保護され、セキュリティ設定の許す範囲で受け渡しが可能なデータの粒度や種類などを設定する。このデータ連携の仕組みによりデータセンター内の秘匿したい個人情報やセンシティブなデータは連携グループに共有することなく代表値に丸めた統計値などのみを共有することが可能となる。
分散型データ分析基盤の利用にはコーディネーションセンターが必要になる。上述のデータセンターが施設単位のデータの集合であり、コーディネーションセンターは複数のデータセンターのデータを管理する役割がある。コーディネーションセンターの役割として、データの標準化の管理やセキュリティ確保のためのアクセス制御を実施する。

分散型データ分析基盤の課題

連携グループ内の各チームはそれぞれが既存のデータベースを使っておりデータベースの製品種類も異なっている。異なるデータベース製品間でのデータのやり取り・コミュニケーションは難易度が上がるが、データ連携ツールの利用などでデータ連携させることは可能ではある。ただし連携グループ内のチーム数が多くなるほど対処する製品間対応が多くなり難易度があがる。

またデータ連携ツールの使用で異なる製品間でデータのやり取りができたとしても、チーム間ではテーブルの項目名も項目定義も異なっているため、複数チームのデータを一括して利活用は一般的には不可であり、可能であっても多くの工数がかかり分析精度も低下する。

データ標準化

そこでデータ連携・利活用するにはテーブルの項目名や項目定義の標準化が必要となる。連携グループ内において異なったデータベース製品を利用するのは今までも今後も避けられないと考えられるため、各チームが既存のデータの形式は継続使用し、既存データを標準フォーマット形式のデータに変換して保持し、その標準フォーマットデータをグループ間で連携・利活用することが現実的と思われる。

更に言えば国際的な標準化の枠組みを見据える必要がある。連携グループは今後統廃合される可能性が大いにあり、連携グループ統廃合の度に標準フォーマットを設定しなおすのは効率がわるい。データ標準化をする際に当初から国際的なデータ標準化フォーマットを利用することが望ましい。医療分野では OMOP CDM などが国際的なデータ標準化フォーマットの有力な一候補になっている。

4.2.3 データプライバシーの技術（匿名仮名加工、連合学習 など）

データの利活用とプライバシー保護を両立させるための重要な概念には匿名化、リスクベースアセスメント、連合学習、K 匿名化などがあり、特に連合学習が有望視されているため連合学習に関連するアイディアを中心に説明を進める。

リスクベースアセスメント

データ利活用とプライバシー保護を完全に両立することは困難である。データ利活用の利便性を向上させるほどプライバシーデータ漏洩のリスクが高まる。二者の完全な形での両立が困難なためリスクベースアセスメントという考え方が導入されている。リスクベースアセスメントアプローチとはリスクを完全にゼロにはせずにある程度のリスクを受け入れるアプローチで、潜在的なリスクを評価して適切な対策を講じるためのプロセスである。データ利活用においてのリスクベースアプローチとは個人情報が再識別されて個人特定されてしまうリスクがゼロではないことを認識しつつ実際にはほぼ発生しない状況を作り、かつデータ利活用を実施する方法である。

リスクベースアセスメントに基づきデータプライバシーを保護する基本技術である匿名加工、仮名加工、およびそれらの応用概念である K 匿名化、連合学習の説明をする。

匿名加工・仮名加工

匿名加工および仮名加工とは個人を特定できる情報を除去することで匿名性を実現する方法である。例えば個人名の情報はデータ抽出の際にカラムそのものを抽出対象から除外する。また別の例では、生年月日と受診日など組み合わせることで個人を特定できてしまうデータの粒度を年単位などに粗くする（粗化）などで個人特定できない状況にする。匿名化の基本技術には仮名加工や匿名加工がある。
匿名加工は年齢や住所などを非特定化する処理である。年齢を通常の 1 歳刻みではなく 5 歳刻みなどに丸めることで同値になる人数を増やすことで個人特定のリスクを下げる。

一方で、仮名加工は個人名やメールアドレスなどを非特定化する処理である。仮名加工は別のデータセットと名寄せできる情報を残す場合が多い。入力データに対して一意に定まる文字列に置換することで同じ名前や同じメールアドレスを仮名状態でも名寄せ可能にできる（*）。

K 匿名化

K 匿名化とは個人が特定される確率が K 分の 1 以下になるように匿名化する処理である（*）。具体的には上述の仮名加工処理や匿名加工処理を使用する。ある集団の中に 97 歳の特定の疾患の患者が 1 名しかいない場合には個人特定につながりやすいが、その集団の 90 歳以上のくくりにすると 20 名の患者が該当して個人特定につながりにくい場合などがある。その場合は 90 歳以上の全患者の年齢を「90 歳以上」などに粗化・匿名加工することで個人特定のリスクを減らすことができる（**）。

K 匿名化データの利用、利点、欠点

K 匿名化データは施設外に共有されることがある。K 匿名化された匿名加工データは年齢などの情報を 10 歳単位などに粗化(粒度粗く)した個人ごとのデータの状態で施設外の研究機関や政府などに共有され、他の施設のデータと一緒にまとめて集計される。K 匿名化の利点として施設外に個人ごとのデータを出せる強みがあげられる。K 匿名化による個人特定リスク減少により個人ごとのデータを施設外に共有したり、複数施設の個人ごとのデータを研究機関や政府でまとめて集計したりすることが可能になっている。一施設内の数百人程度の集団では見えにくい関係性も複数施設をまたいだ数万人以上の集団では因果関係を発見できることもある。万一データが流出しても K 匿名化されており直接的な個人特定にならないためリスクベースアセスメントの意図に沿っている。

K 匿名化には欠点もある。個人単位のデータが施設外に出ることが強いリスクとなりうる。K 匿名化された個人単位のデータが施設外に出ると背景知識攻撃などのリスクにさらされる。背景知識攻撃とは特定のターゲット人物の背景情報を使いターゲット人物を個人特定しようとする攻撃である（**）。

K 匿名化の別の欠点として施設側の負担の重さがあげられる。K 匿名化は多次元になるほど弱く、個人特定につながりやすい（**）。そのため K 匿名化を利用して複数施設のデータを合わせて分析する場合には都度 K 匿名化のリスクを考えて個人データが特定されない粒度への粗化処理が必要となり施設側の負担が大きくなる。

K 匿名化の時代から連合学習の時代へ

上述のように K 匿名化は直接的には個人特定ができないためデータ活用に重宝されているが、個人ごとのデータを施設外に提供することに強い懸念を示す業界もある。例えば医療業界などは個人データの秘匿要望性が高いため、個人ごとデータの施設外への共有は敬遠され、集団の代表値に要約したデータのやり取りが好ましい。それを実現した方法が連合学習である。

分析におけるデータ連携の 3 手法

分析を実施する際に施設間のデータ連携の取り扱いは主に 3 つの手法があり、個別施設学習、集合型学習、連合学習に分けられる。表 1 は 3 手法の比較である。総合的な有望度合いが高い連合学習が今後大きなトレンドになると考えられている。まずは既存技術の個別施設学習、集合型学習を説明し、その後に注目の連合学習を説明する。

表 1. データ連携手法の比較表

	分析に考慮されるデータの量	分析の精度	データセキュリティ	データ入手の容易度合	総合的な有望度合い	例 体重 ＝0.73 x 身長センチ - 66 1 施設 30 人 x 10 施設
個別施設で分析	× 1 施設内のデータ	× 低い	○ 施設外にデータ非共有	○ 容易	△ 低い分析精度	施設 A は 30 人のデータで分析

集合型で分析	○ 全施設のデータ	○ 高い	× 施設外にデータ共有するためセキュリティが困難または複雑	× セキュリティ観点で複数施設データを収集承認困難	△ セキュリティ難色	集合型では300人のデータで分析
連合学習で分析	○ 全施設のデータ	○ 高い	○ データの中身が見えない状態で複数施設のデータを連合させて分析可能	○ 個人データを収集しないため収集容易	○ 分析精度高い。セキュリティ考慮済み	連合学習では施設毎のパラメータ（w=a*h-b)を収集

個別施設で分析

個別施設での分析とは施設内で得られたデータで集計・分析を実施する手法である。病院の例で考えた際、特定の疾患 A の患者が 1 施設に累計 30 人いた場合。30 人のデータで見える範囲の因果関係を分析によって明らかにすることができる(表 1)。この手法の利点として、高い機密性と簡易性が挙げられる。データを他の施設やデータベースと接続しないためデータ漏洩する機会が減る。またデータ連携・データ共有をしないためシステム構築が比較的簡易である。一方、この手法の欠点は分析精度の低さである。1 施設内で得られたデータのみを使用するため知見がかなり限定的になり、因果関係が見えにくく、機械学習や深層学習を実施しても分析結果はかなり限定的になる。

集合型で分析（生データ）

集合型(中央集権型)で分析する場合、複数の施設でコンソーシアム(共同体)を結成しコンソーシアム内の全施設の全個人のデータを一か所に集めて分析を進める。この方法の利点は個人データの人数が多いため分析の精度が高くなる。欠点はプライバシー保護リスクである。各施設から見ると施設外に個人ごとのデータを提供することになりデータの漏洩リスクが高まる。

K 匿名化と集合型で分析

K 匿名化と集合型を組み合わせてコンソーシアムで中央集権型の分析が可能である。この方法の利点は上記と同様に個人データの人数が多いため分析の精度が高くなる傾向がある。ただし K 匿名化の処理で個人ごとのデータが粗化されているため生データによる集合型の分析ほどの精度は見込めない。反対に欠点も薄まる。施設外に個人ごとのデータを提供するが、仮に漏洩しても K 匿名化されており直ちに個人特定には至らないためリスクが強くない。

K 匿名化と集合型を合わせたこの方法は現在頻繁に用いられる方法で、データ利活用と個人特定リスクのバランスが取れており、広く使われる手法になっていた。しかし近年、K 匿名化では個人特定リスク対処が十分ではないと考えられることも多く、次に述べる連合学習が注目されている。

連合学習で分析

連合学習とは各施設に分散している個人ごとのデータを施設外に共有しない状態で複数施設データを連合して精度の高い分析を進める手法であり、プライバシー保護に優れている。具体的には統計集約の数式や機械学習の学習モデルを分析機関から各施設に投げ、各施設で集約結果や機械学習モデルパラメーターを分析機関に返信し、分析機関が各施設の結果をまとめて分析する。この手法により、各施設は個人ごとのデータを施設外に提供せずに、分析機関は複数施設の個人ごとのデータを反映した精度の高い分析結果を取得できるという両者(分析機関、データ提供施設)ともに嬉しい手法となっている。

連合学習に必要な整備

連合学習には事前の準備が必要である。統計の集約や機械学習をどの施設でも同じ方式で進められるように各施設のデータの保存方法を一定のルールに従って進める必要がある。

データの保存方法の一定のルールを Common Data Model (CDM)と呼ぶ時がある。CDM には多くの種類が存在し、例えば前述の OMOP CDM などがある。

今後、施設間や多国間で共通のデータ保存方法で蓄積したデータの利活用が容易となるOMOP CDM は注目されている。

プライバシー保護の主要技術

この項ではプライバシー保護した上での分析の技術の比較として、K 匿名化を利用した方法、連合学習、秘密計算技術を説明する。表 2 の比較表を利用して説明する。

表 2. プライバシー保護をしたうえでの分析の技術

	目的	用途、利用例	概略	分析者(政府など)にとっての集計容易度合い	データの標準化の度合い	プライバシー保護の信頼度	バイオデータ連携・利活用での総合的な有望度合い
秘密計算技術	プライバシー保護	個人が疾患 A の罹患確率の AI 結果を取得	個人データを暗号化	× 用途外	○ データ標準化	○ 個人データ暗号化	○ 個人のデータ活用が可能
K 匿名化	プライバシー保護	データ連携、 地域ごとの疾患 A の罹患率の集計	個人特定できない人数単位で集計	△ 収集要求の度に施設毎の対応必要	× 標準化未対応	△ 個人単位のレコードを施設外に共有するため名寄せなどで再特定リスクあり	△ データ連携工数多い。集計の度に施設毎に対応必要。

連合学習	プライバシー保護	データ連携、地域ごとの疾患 A の罹患率の集計	施設単位での代表値・パラメタで集計	○ 即日集計も可能	○ OMOP 利用などで標準化	○ 施設外に共有されるデータは集団の代表値であり名寄せ再特定リスクはほぼない。	◎ データ連携が容易に分析者命令を即日で実施できるようになる

秘密計算技術

秘密計算技術とはデータを暗号化したまま計算する技術である。例えば住宅ローンの審査申し込みをする際に購入希望者が年収情報などを入力しても即刻暗号化されセールスマンは年収情報を見ることなく、分析機関でローン審査の計算が行われて審査結果のみが返ってくる。今までの技術では人工知能などへの入力データはデータが分析モデルに入力されるまでの各中間地点で見られてしまう状態にあったが、秘密計算技術を使用すると入力データが暗号化された状態で分析モデルに入力されるため中間地点で第 3 者に見られなくなる。秘密計算技術の他の例として、自身の医療情報を中間地点の第 3 者に見られることなく入力して疾患予測の結果だけを受け取ることも可能となる。

プライバシー保護の主要技術の比較

秘密計算技術、K 匿名化、連合学習を比較すると秘密計算技術は個人のデータ分析用途であり、K 匿名化と連合学習は集団のデータ分析用途と言える(表 2)。秘密計算技術はブロックチェーンなどの暗号化技術により個人データを暗号化して分析モデルに送るためプライバシー保護に強い反面、暗号化技術のために処理が重く大量のデータ処理には向かない傾向がある。

バイオデータ連携利活用ではビッグデータで大人数のデータ処理を進めることが予想されるため K 匿名化や連合学習が注目技術となる。その中でも上述の通り個人特定リスクが低く施設側の負担が低い連合学習が特にトレンドとなると予想される。

[参照]

*荒牧英治. 医療言語処理. コロナ社, 2017.

** Khaled El Emam, Luk Arbuckle 著 ; 笹井崇司訳. データ匿名化手法 : ヘルスデータ事例に学ぶ個人情報保護, オーム社, 2015.

4.3　国内におけるデータ連携事例

SIP 第 3 期　統合型ヘルスケアシステムの構築

SIP 第 3 期「統合型ヘルスケアシステムの構築」は、戦略的イノベーション創造プログラム（SIP）第 3 期の課題のひとつであり、医療デジタルツインの実装を通じて、医療・ヘルスケアにおける「知識発見」と「医療提供」の循環を自律的に促進し、医療の質向上、健康寿命延伸、医療産業振興、持続可能な医療制度に活用することを目指すものである。本プログラムにおいて掲げられている「現状と問題点」は大きく 3 点に要約される。

- **医療データの断片化**：
  医療データが様々なシステムやプロバイダーに分散し、統合されていないこと
- **AI 技術の活用不足**：
  医療現場における AI の適切な導入と活用が進んでいないこと
- **プライバシーとセキュリティ**：
  プライバシーを保護しながら、データを有効活用する方法が十分検討されていないこと

上記課題を解決し、SIP 終了時点（2028 年）には、50 病院程度で医療デジタルツインのシステムが全面的あるいは部分的に導入され、医療データが標準化・統合された医療情報統合と医学知識発見が始まることを目標としている。さらに SIP 内外の取組とこれを活用したソリューションの普及により、SIP 終了から 5 年経った段階（2033 年）では、約 1,000 病院で、部分的であれ医療データが標準化・統合され、医学知識の発見が継続的に行われる。その基盤として、本デジタルツインと互換性のある医療データプラットフォームが、全国の中核病院 100 か所で導入されることを目指している。

また、期間中に医療 DX やデータヘルス改革等における電子カルテ情報の標準化に向けた取組や、次世代医療基盤法等の法整備による環境整備・社会的受容性の向上が進むこと、電子カルテ情報の標準化及び交換方式の標準化等、医療デジタルツイン構築の観点から重要度が高い取組に対しては、関連するステークホルダーとの連携と意見交換を通じて相互理解を深めることが期待されており、本調査報告書における海外先進事例を参考とした、データ連携・利活用を検討する余地も大きいものと考えられる。

Real World Evidence 創出のための取り組み（通称臨中ネット）

AMED が実施する医療技術実用化総合促進事業の一環として、2018 年から臨床研究中核病院(*)を対象に開始された取り組みである。電子カルテを中心とした病院情報システムより得られる患者データを リアルワールドデータ（RWD） として集積し、リアルワールドエビデンス（RWE）の創出を目指す。病院情報システムより RWD を抽出するにあたっての課題の整理とそれを解決するための方策の検討と実装、ならびにユースケースを中心とした個別研究の試みとそれを通じた課題解決のために、サブワーキンググループ(SWG)を中心とした議論と臨床研究中核病院間で協調した RWE 創出のための体制構築、実装を推進している(**)

本取り組みにおいて整備が進んでいる臨中ネット標準 DB では、標準データテーブルが用意され、研究者はクエリを作成することで各医療機関から検索結果を集めることが可能となる。検査・薬剤等のコード（用語集）に関しても課題として検討、対応がすすんでいる(***)。

今後このようなネットワークをいかに多数の施設に拡大していくか、ネットワークの持続可能性を如何に担保するか、データの標準化やインターフェースを含め如何に利便性を高くし利活用を推進していくかが課題となってくると考えられる。

*臨床研究中核病院について（厚生労働省）
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/tyukaku.html
**臨床研究中核病院で試みる Real World Evidence 創出のための取り組み
松木絵里他、日本医療情報学会 第 40 回医療情報学連合大会（第 21 回日本医療情報学会学術大会）
***武田理宏他、臨床研究中核病院による「Real World Evidence 創出のための取組み（臨中ネット）」での標準化の検討、MID-NET シンポジウム 2023（独立行政法人医薬品医療機器総合機構）
https://www.pmda.go.jp/safety/symposia/0025.html

4.4　まとめ　国内における課題と示唆

本章では、海外の連合型ネットワーク先進事例に焦点を当てた紹介と、データ活用に関する関連技術について紹介した。まず、海外における連合型ネットワーク事例をみると、立ち上げにあたってはすべて政府あるいは官民共同で資金拠出がされていることがわかる。そして、各取り組みの成果指標としては以下のように段階に分けて整理すると理解しやすい。

①データモデルの浸透・研究期間及び研究者のネットワーク構築
②データの RWD CDM への変換・データベースの準備
③目的に応じた分析や研究の実施
④これらを持続可能にするためのエコシステムの構築

EHDEN を例にとると、OMOP CDM を用いて①のネットワーク構築に国を跨いで成功しており、②（データ準備）・③（分析・研究）も実現している(*)。また、医薬品の安全性調査を目的とした FDA BEST や DARWIN においても RWD CDM を活用してその目的（③）を実現しつつある。

これら連合型データ連携事例の多くは取り組みが始まってから 10 年以内であり、今後の拡大・発展を目指しているフェーズであるため、④（エコシステム）に関してはまだ正解といえるモデルが出てきていないのが現状であるが、ファンディングスキームとしては、EHDEN における官民連携モデルは一つの見本といえよう。「民」の部分も散発的な企業の取り組みに頼るのではなく、業界団体（欧州製薬団体連合会：EFPIA）として出資が行われており、より公共の利益につながりやすいモデルと考えられる。今後の課題としては、運営を純粋なコストとして捉えられるのではなく、連合型ネットワークからいかに研究開発等の事業成果が生み出され、その収益の一部ネットワーク維持費として還元されていくエコシステムの構築であろう。

RWD CDM の観点では、本章で紹介した事例では OMOP CDM が採用されていることが多い。海外有識者へのヒアリングからはその理由として、以下の観点があげられていた。

・大規模な観察研究を連合型ネットワーク構築の目標として掲げていること

・用語の意味内容（ボキャブラリー、用語集）まで踏み込んだデータの標準化をしていること
・米国だけではなく世界中で活用されており、患者数で 8 億人以上(**)と OMOP 変換されているデータも多いこと
・開かれたオープンコミュニティで啓発活動も積極的に行っていること

中央集権型ではなく連合型のネットワークを選択する背景は、主に各取り組みの目的・参加組織の意向・各国の規制の 3 つの要因に依存する。例えば、リアルワールドデータを活用した取り組みである N3C や All of US では統合型データベースのモデルを採用している。参加組織がその目的と照らし合わせてデータを外部に提供して統合することに同意すれば、必ずしも連携型である必要はない。一方で、FDA BEST、DARWIN、EHDEN のとりくみにおいては規制の問題からデータを各施設から外に出さないスキームが求められるため、連合型のモデルが望ましい。特に EU においては国を跨いだネットワークを構築する必要性が高いこと、GDPR に準拠することが求められ、結果として連合型の h 取り組みが進んでいると考えられる。

我が国においても、次世代医療基盤法に基づく認定匿名加工医療情報作成事業者により、匿名加工データの活用が成果をあげつつある。一方で、第 3 章でも述べた国際標準を目指した RWD CDM の活用議論、さらには本章で事例にあげたような連合型ネットワーク構築に関しては、ほとんど事例がないのが現状である。そうした中、1.2 章でも触れた産学連携による次世代創薬 AI 開発（DAIIA）(***)は、創薬基盤構築として期待される事業であり、今後もバイオデータの連携・利活用を促進していく目的に鑑みて、上述の海外先行事例のエッセンスを十分に取り入れた、国内における連合型ネットワーク構築の検討が進むことが期待される。

[参照]

*Publications (EHDEN)

https://www.ehden.eu/publications/

**OHDSI

https://ohdsi.org/who-we-are/

***産学連携による次世代創薬 AI 開発（DAIIA）：

https://www.amed.go.jp/program/list/11/02/001_02-04.html

5. 日本が今後取り組むべき方向性

現在に至るまで、バイオデータ利活用の中心は、集合型のビッグデータベースの構築と利用目的に応じた匿名・仮名化加工によるデータの利用許諾形式であった。我が国において、個々のデータベースならびにその統合事業においては一定の成果が産まれているものの、データベースの構築ならびに維持管理にかかるコスト、利用許諾プロセスなどの課題も大きい。また諸外国と比較した場合には、必ずしもビッグデータの囲い込みで競争優位を築けているとは言えない状況でもある。今後我が国がバイオ関連産業の育成において国際競争力を獲得していく目的においても、国内外に分散しているデータを効率的、効果的に連携していくことにより、これまで以上の“ビッグデータ”の活用を推進していくことが不可欠である。

本調査報告書では、分散型データ連携の手法として一貫して連合型のデータ連携について概念、手法、実例を通して紹介をしてきた。連合学習の概念を応用した連合型データ連携は、データそのものは各々の発生元に保持しながら解析結果のみを統合する方式であり、秘匿性や個人情報保護に配慮した新たな方法として、実際に米国、欧州、アジア各国が先行する形で産官学が連携した戦略的な取組を進めており、実際に成果が出ている。この流れに我が国も遅れを取ってはならない。

連合型データ連携を進めるにあたって、先進事例からの学びと国内の状況を踏まえた際の課題は大きく 2 つに分けられる。

１つ目は連合型連携の基礎となる、バイオデータの規格、方式の統一である。特に電子カルテや病院情報システムなどの診療情報に関しては構造ならびに用語体系の非統一（非正規化）が利活用に際して大きなハードルとなっている。一方、この課題は諸外国においても同様であり、先進事例からの学びを含めると、一斉にデータ方式の統一規格を構築し統一規格の普及を進めることのみが解ではないことがわかる。海外の事例を見ても、健康・医療関連情報の一切を統一するのではなく、COVID-19 に必要な単位、特定の疾患（オンコロジー領域など）、利活用の目的に応じた項目単位など「単位ごとの標準規格」を産官学がスピード感を持って構築、普及させていることが成功要因となっ

ており、我が国においても現実的な解決策を検討するうえで重要な視点と考えられる。また、その際には利活用の余地を狭めてしまうガラパゴスの規格を採用するのではなく、医療情報を交換規格である HL7 FHIR と同様に、「国際標準規格」を前提とすることが、ひいては国際競争力の獲得につながることを意識すべきである。

2 つ目は DX の概念を強く意識して、連合型データ連携により実際に利活用されていくためのプロセス、法制度、システム構成、人財育成などを入念に企画しておくことである。諸外国における CDM の取組ならびに連合ネットワーク先進事例を見ても、構築後にその運営段階で躓くケースが散見される。データ利活用は構築、運用、活用（マネタイズ）のエコシステムが循環していくことが存続要件である。いかにいい仕組みを構築しても、活用されず経済が循環していかなければ持続的な存続は不可能である。そのためには 1 つ目の課題と同様に、産官学が密に連携することにより「持続可能かつ現実的な仕組み」を構築していくことが不可欠である。