欧州におけるデータの質確保に向けた取り組み
～二次利用を促進するFit for purposeのデータ～

医薬産業政策研究所　主任研究員　辻井惇也

要約

欧州において、データ特性に基づくデータの質は、データ利用者の目的への適合性と定義されるとともに、現実を反映するデータであることが求められる。健康医療データの質確保に向け、欧州では、Fit for purposeの概念に基づいたデータクオリティフレームワークがTEHDAS（Towards the European Health Data Space）やEMA（欧州医薬品庁）・HMA（欧州医薬品規制首脳会議）から公開されており、データの質に対するステークホルダー間の合意形成が進められている。また、欧州ではデータ利用者がデータの質を適切に判断できるよう、標準化されたメタデータ・カタログの構築も検討されている。具体的取り組みの一つとして、EMA-HMAは、医薬品規制上の意思決定に用いるリアルワールドデータに対するメタデータ・カタログの推奨事項を提供する世界初のガイドを公開している。加えて、これらの取り組みを実効性あるものとするため、データ保有者や管理者に対し、データ収集並びに二次利用にかかるコストを考慮した金銭的インセンティブの付与が検討されていることを確認した。

1．はじめに

未来の医療ヘルスケアにおいて、人々は出生時から取得されたあらゆるデータを活用し、病気になる前から自身の健康状態を管理・増進するとともに、疾患の早期発見や個人に合った適切な治療介入、予後のケアを受けているだろう。また、様々に取得された健康医療データは、その時々で最適な政策決定や医薬品を含む医療技術の革新、医療資源の最適配分等をもたらし、人々がより健康で過ごすための支援を提供しているだろう。

このような未来の実現の鍵を握るのは、「健康医療データ」であり、近年、わが国において、健康医療データの利活用を促進する制度政策の整備・検討が進んでいる。2023年６月２日に公開された「医療DXの推進に関する工程表」において、国民のさらなる健康増進や切れ目なくより質の高い医療等の効率的な提供等を実現するPersonal Health Record（PHR）の活用、情報共有の仕組みの構築が目指されている[1)]。また、産業界等でのより良い二次利用を支援するため、薬事承認申請への仮名加工医療情報の活用等を盛り込んだ次世代医療基盤法の改正[2)]（2023年５月成立）や明示の同意に必ずしも依存しない医療等データの利活用法制等の検討の必要性を示した規制改革実施計画[3)]（2023年６月閣議決定）のような取り組みも活発化している。一方、“質”の意図するところに違いがある可能性に留意すべきものの、業界紙の中で、

1）厚生労働省、第100回社会保障審議会医療部会　資料２　医療DXの推進に関する工程表について（報告）（令和５年７月７日）（2023年９月７日閲覧）、https://www.mhlw.go.jp/content/12601000/001118552.pdf
2）首相官邸、健康・医療戦略推進本部、第８回　次世代医療基盤法検討ワーキンググループ　資料１　改正次世代医療基盤法とその施行に向けた検討について（令和５年６月28日）（2023年９月７日閲覧）、https://www.kantei.go.jp/jp/singi/kenkouiryou/data_rikatsuyou/jisedai_iryokiban_wg/dai8/siryou1.pdf
3）内閣府、規制改革実施計画について（令和５年６月16日閣議決定）（2023年９月７日閲覧）、https://www8.cao.go.jp/kisei-kaikaku/kisei/publication/program/230616/01_program.pdf

健康医療データのさらなる質・量の確保に関する言及がいくつか見られる[4)、5)]。

このようにデータ利活用、特に二次利用を加速するための制度政策の整備は進みつつあるものの、実際に利活用される健康医療データの質や量の観点での検討は未だ十分とは言えないと考える。以上を踏まえ、本稿では、日本製薬工業協会をはじめとする複数の団体から日本の健康医療データ利活用の参考として言及される欧州の取り組みを俯瞰し、データの質確保の面からわが国でのデータ二次利用を促進するための方策を検討した。

2．データの質確保に向けた欧州の動向

2－1．European Health Data Space（EHDS）とは

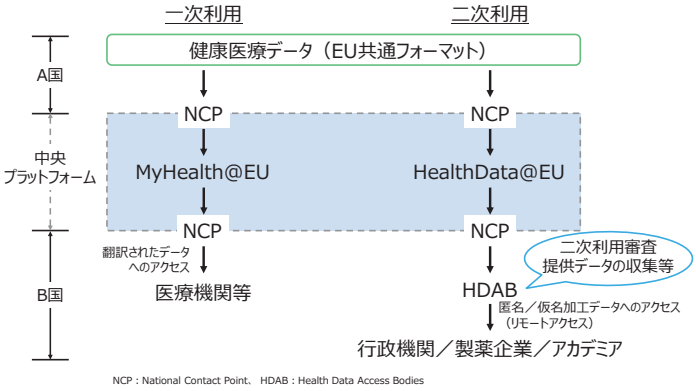
欧州における健康医療データ利活用の基盤となる政策として、European Health Data Space（EHDS）の取り組みが見逃せない。EHDS は、欧州域内の国民に対する情報の越境利用を通じた質の高い医療の提供（一次利用）や医療政策、医学研究、創薬等（二次利用）を推進する欧州共通のデータスペース構想であり、経済的利点として、10年間で約110億ユーロのコスト削減効果が見込まれている[6)]。（内訳として、遠隔医療の普及率向上等による医療分野におけるコスト削減と効率化：54億

補足：国境を越えたデータ利活用の流れ

EHDSに基づく越境でのデータ利活用の流れを図１に示す。

国内の医療機関等が有する健康医療データは、加盟各国に設置されたNational Contact Point（NCP）が窓口となり、中央プラットフォームを介して国境を越えたデータ交換が行われる。各国のNCPと中央プラットフォームを結ぶインフラストラクチャーは利用目的に応じて異なり、一次利用ではMyHealth@EU、二次利用ではHealthData@EUが用いられる。MyHealth@EUは、現状11か国の稼働に留まるが、2025年までにEU及びEuropean Economic Area（EEA）加盟国の大半が参加する予定である[7)]。また、二次利用においては、各国に設置されたHealth Data Access Bodies（HDAB）という機関が、EHDS 第34条に記載の利用目的に合致すると判断した場合に限り、個人が特定されない加工を施したうえで、データの利用が認められる。なお、EHDS 第41条では、データ保有者の義務として、データを利用可能とすること、データ提供要請があった場合、決められた期限内で提供すること等の義務が課されている。加えて、HDABは、アクセスが許可された法的根拠やデータが使用されたプロジェクトの結果等を一般に公開し、かつ容易に検索可能とする義務を負う。

図１　国境を越えたデータ利活用の流れ



出所：医薬産業政策研究所で作成

4）日刊薬業、「医療DX工程表、製薬企業の医療情報二次利用後押し 政府が公表」（2023年６月２日記事）、https://nk.jiho.jp/article/181398

5）日刊薬業、「希少疾病創薬の加速期待、「仮名加工」新設で　製薬協、利用基準や海外申請など論点に」（2023年７月11日記事）、https://nk.jiho.jp/article/182395

6）欧州委員会、Questions and answers - EU Health: European Health Data Space（EHDS）（2022年５月３日）（2023年９月21日閲覧）、https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_2712

7）欧州委員会、European Health Data Space（EHDS）（2023年９月22日閲覧）、https://www.s3vanguardinitiative.eu/system/files/2022-11/European%20Health%20Data%20Space_0.pdf

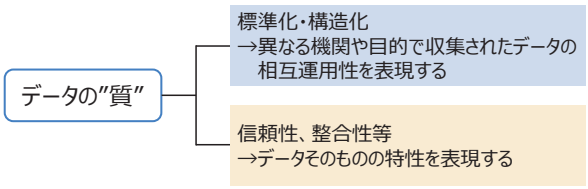
ユーロ、各国のHealth Data Access Bodies（以下HDAB、詳細は補足を参照）を通じたデータアクセスの仕組みの構築等による健康医療データの再利用コストの削減：34億ユーロ[8]等がある。）

EHDSは、データ保護に関しての欧州一般データ保護規則（GDPR）に基づき構築されているが、GDPRでは、第９条において、遺伝子データ、生体データ、健康に関するデータ等の取り扱いに対して、加盟国に独自の裁量を与えたことから、加盟国間で実施や解釈に差異が生じ、実際のデータ利用に障壁があった。さらに、データの質の違いが国境を越えたデータ共有に対する障壁となっていることも指摘されている[9]。EHDSはこれらの課題を解決するための欧州共通のデータエコシステムとして検討されており、2024年の法制化、2025年の施行を目指した取り組みが進んでいる。

2−2. データの“質”とは

データの“質”とは何か。データの質について、筆者は大きく２つに分けることができると考える（図２）。一つは、データの標準化･構造化であり、異なる機関や目的で収集されたデータを相互運用可能な同一フォーマットで保管することを指す。もう一つは、データの信頼性や整合性等、データそのものの特性を表現するものである。わが国において、電子カルテの標準化等のデータ標準化・構造化に関する議論は既に進みつつあることを鑑み[10]、本稿では、“データそのものの特性”の観点からデータの質を考える。そのため、本稿における「データの質」は、特段の断りがない限りデータ特性に基づく質を指す。

図２　筆者の考えるデータの質の分類



出所：医薬産業政策研究所で作成

データ特性にフォーカスしたデータの質の定義については[11]、TEHDAS（Towards the European Health Data Space）の報告を参考にしたい。TEHDASはEHDSの実現に向け、21のEU加盟国と他の欧州４か国の様々な組織・団体（公的機関、アカデミア、医療学会、産業界、患者団体等）が参加し、健康データの二次利用に関する欧州共同原則を開発するプロジェクトである[12]。TEHDASにある８つの作業パッケージのうちWork Package 6（WP6）：Excellence in data qualityでは、データ品質を確保するためのガイダンスを作成している。WP6が2023年９月に公開した「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use[13]」において、データの質は「データ利用者のニーズ（健康研究、政策立案、規制）に適合すること」と定義されている。参考として、

８）欧州委員会、IMPACT ASSESSMENT REPORT, PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space（2022年５月３日）（2023年９月21日閲覧）、https://health.ec.europa.eu/system/files/2022-05/ehealth_ehds_2022ia_2_en.pdf

９）Towards European Health Data Space（TEHDAS）、Recommendations for European countries when planning national legislation on secondary use of health data（2023年３月１日公開）（2023年９月７日閲覧）、https://tehdas.eu/app/uploads/2023/03/tehdas-recommendations-for-european-countries-when-planning-national-legislation.pdf

10）内閣府、規制改革推進会議、医療等データの利活用法制等の整備について（案）（令和５年６月１日）（2023年９月８日閲覧）、データの質として、医療等データの標準化を議論している、https://www8.cao.go.jp/kisei-kaikaku/kisei/meeting/committee/230601/230601general_03.pdf

11）ISO 8000-2 Data quality -Part 2: Vocabulary（2023年９月８日閲覧）、広義のデータの質として、例えばISO 8000では、「データ固有の特性が要件（明示的、一般的に黙示的、または義務的な必要性または期待）を満たす度合い」としている、https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-5:v1:en

12）Towards the European Health Data Space ホームページ（2023年９月８日閲覧）、https://tehdas.eu/project/

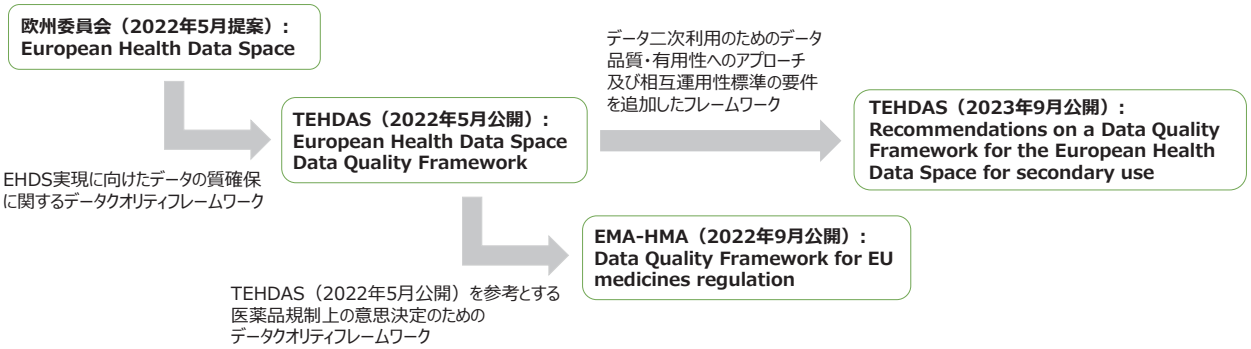
13）Towards the European Health Data Space（TEHDAS）、Recommendations on a Data Quality Framework for the European Health Data Space for secondary use（2023年９月26日）（2023年９月28日閲覧）、本文書では、データクオリティフレームワークの構成要素やステークホルダーに求められる対応事項、実施すべきデータガバナンス（相互運用性、データリンケージ、セキュリティ等）等に関連する13の提言を公表している、https://tehdas.eu/app/uploads/2023/09/tehdas-recommendations-on-a-data-quality-framework.pdf

同じくWP6から2022年５月に公開された「European Health Data Space Data Quality Framework[14]」では、EHDSの中で利活用されるデータの質を「健康研究、政策立案、規制に関する利用者ニーズに対して目的に適合しており、さらにデータが表現しようとする現実を反映していること」とある。以上を踏まえると、データ特性に基づくデータの質は、データ利用者（民間企業、行政機関、アカデミア等）の目的にどの程度適合しているか、つまりデータ利用者の目的適合性により定義されるものと考えるが、前提として現実を反映するデータであることも求められる。データの質は「目的適合性」に加え、「現実の反映性」という２つの観点が重要と言えよう。なお、後述するEMA（欧州医薬品庁）及びHMA（欧州医薬品規制首脳会議）が2022年９月に共同で公開した「Data Quality Framework for EU medicines regulation」においては、2022年５月に公開されたTEHDASの提案を参考にした定義が記載されている[15]（図３）。

2−3. データの質の決定要素

健康医療データの質の決定要素について、TEHDAS及びEMA-HMAより公開されたデータクオリティフレームワークから考える。なお、対象となるデータについて、TEHDASフレームワーク内での直接の言及はないが、EHDSの実現が念頭にあることから、医療データやゲノム、オミックスデータ、臨床試験の電子ヘルスデータ等（EHDS 第33条[16]）が対象となろう。一方、EMA-HMA フレームワークでは、医療データ・リアルワールドデータ（以下、RWD）に加え、特に重要な分野として、生物学的分析オミックスデータや前臨床データ、有害事象自発報告データ、化学物質・製造管理データ等が挙げられている。データの粒度としては可能な限り低いレベル、つまり値レベル（特定のデータ点）に焦点を当てている。

図３　EHDS と TEHDAS／EMA-HMA データクオリティフレームワークの関係性



注：EMA-HMAフレームワークはTEHDASフレームワークを参考に作成されており、文書の上位／下位の関係性を表すものではない。
出所：医薬産業政策研究所で作成

14）Towards the European Health Data Space（TEHDAS）、European Health Data Space Data Quality Framework（2022年５月18日）（2023年９月８日閲覧）、https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf
15）EMA-HMA、Data Quality Framework for EU medicines regulation（2022年９月30日）（2023年９月８日閲覧）、https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf
16）EUR-Lex、Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space（Document 52022PC0197）（2023年９月８日閲覧）、https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197

2−3−1. TEHDAS:Recommendations on a Data Quality Framework for the European Health Data Space for secondary use

データセットレベルでのデータの質担保

データの質の測定にあたっては、質の決定要素を規定する必要がある。質を決定する要素は次元（Dimension）と呼ばれ、現実の１つまたは複数の関連する側面または特徴を表す、測定可能なデータ特性に関する指標である。本フレームワークでは、データ利用者による目的適合的なアプローチを促進するため、データセットレベルでのデータ品質（quality）に加え、データ有用性（utility）にも焦点を当てており、これらに対する次元を規定している。有用性とは、データ利用者を中心とした事前及び事後的な条件に依存するデータの質であり、データ利用前においては、下記の次元のいくつかにより評価が可能である。一方、事後的には、当該データセットの利用（利用実績）、関心（照会実績）、価値（データ利用者による提供データの評価）の尺度により有用性を測定でき、特定の目的に特化した利用者の潜在的な期待への充足に基づき評価されるとしている。

データ品質及び有用性に関する具体的な次元として、下記の６つを挙げている（表１）。

表１　TEHDASフレームワーク：データの質（品質・有用性）に関わる次元

次元	定義
該当性（Relevance）	利用者のニーズをどれだけ満たしているか
正確性（Accuracy） 信頼性（Reliability）	測定のために設計された内容をどれだけ忠実に反映しているか、またそれが長期にわたり一貫性があるか
整合性（Coherence）	データソースやデータ保有者間でデータがどれだけ整合（一貫）しているか
網羅性（Coverage）	データセットが参照する母集団、及びその曝露、事象の代表性の度合い
完全性（Completeness）	変数レベルでの欠測の程度
適時性（Timeliness）	情報がどれだけ最新の状態で収集され、提供されているか

出所：TEHDAS「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」[13] をもとに医薬産業政策研究所で作成

① 該当性（Relevance）
② 正確性（Accuracy）・信頼性（Reliability）
③ 整合性（Coherence）
④ 網羅性（Coverage）
⑤ 完全性（Completeness）
⑥ 適時性（Timeliness）

このうち、該当性（Relevance）、正確性（Accuracy）･信頼性（Reliability）、整合性（Coherence）は品質に関わる次元、網羅性（Coverage）、完全性（Completeness）、適時性（Timeliness）は有用性に関わる次元として、フレームワークに含むべきと提言している。また、データの誤りを修正するために、データ利用者から、データセットの品質や有用性に関するフィードバックを提供することが推奨されている。ただし、データの質の定義のとおり、重要な点は「目的適合性」であり、各次元で到達すべき基準は利用目的ごとに異なることが前提となることに留意が必要である。

データ保有者レベルでのデータの質担保

EHDS 第33条に基づき、データ保有者（医療機関、研究機関、公的機関､EU機関等）は電子データを二次利用可能な状態に整備することが求められる。TEHDAS フレームワーク[13] では、データの質の観点から､EHR（Electronic Health Record）のように純粋な医療目的で収集されたデータの場合（特にデータセットが定期的に更新され、連結される場合）､データ保有者はデータの質の管理及び保証の手順を構築しなければならず、データの質管理はデータライフサイクル全体（図４）で、データの質保証はデータ管理プロセス（モニタリング、インシデントの検出･解決、データ強化等）横断的に適用されるべきと提言している。

データの質管理手順（データガバナンス）は、データ保有者の成熟度に応じて、自動化されることが望ましい。成熟度はデータ保有者（組織）の行動や実践、プロセスがどの程度信頼性高く、持続的に要求される結果を生み出すことができるかを表すものであり、主に能力成熟度モデル（Capability Maturity Model:CMM）により評価される。

図４　データの質管理に関わるデータライフサイクル

✓ 収集 ✓ キュレーション ✓ 保管 ✓ ステージング	✓ 関連ソースやシステムとのデータ統合	✓ データの記述 ✓ メタデータ管理	✓ データ品質評価 ✓ データプロファイリング ✓ 改善	✓ データモデリング ✓ 変換 ✓ 運用 ✓ データサービス

注：「キュレーション」とは、情報を探している人がアクセス可能なよう、データセットを作成、整理、維持するプロセス
「ステージング」とは、情報を取り込み、変換・統合したデータを作成し、永続的／長期的なストレージにロードするためのプロセス
「データプロファイリング」とは、データセットを調査、分析、レビュー、要約するプロセス
「改善」とは、データクレンジング、整理、移行により、データを適切に保護し、本来の目的に最適な状態とするプロセス
「データサービス」とは、データの利用可能性等を高め、より有用なものとするために、データが本来持たない特性を与えるプロセス

出所：TEHDAS「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」[13] をもとに医薬産業政策研究所で作成

CMM では、データ保有者が継続的にデータの質を改善する実際の能力に応じて、成熟度を下記の５つのレベルに分けている。

① Initial：文書化が不十分で、その場限りの（非体系的な）データの質のチェックが行われている
② Repeatable：データの質管理の手順が十分に文書化されており、同じ手順を繰り返すことができる
③ Defined：データの質管理の手順が十分に定義され、標準的なプロセスとして実施されている
④ Managed：データの質管理のプロセスに、品質に対する定量的評価指標が含まれている
⑤ Optimised：データの質管理は、最適化と継続的改善の意図的なプロセスを意味する

上記の成熟度モデルは、データ保有者（組織）を比較するベンチマークとして利用することが想定される。TEHDAS の提言では、成熟度レベルの「Initial」ではデータ保有者自身の自己評価、それ以降のレベルでは外部監査と認証を推奨している。さらに、継続的な改善とレベル向上のためには、データ保有者に対するインセンティブ設計も重要としている。

2−3−2. EMA-HMA：Data Quality Framework for EU medicines regulation[17]

EMA-HMA の共同設立による Big Data Task Force が発表したデータクオリティフレームワークは、医薬品規制上の意思決定に用いられるデータの質を特徴づけ、評価のために関係者が広範なデータソースに適用できる定義や原則、評価手順等を規定している[18]。本フレームワークは、2022年５月公開のTEHDASフレームワーク（European Health Data Space Data Quality Framework）を参考に、EMA、HMA 及び TEHDAS に関係する広範な利害関係者からのフィードバックを考慮し作成されている。EMA-HMAフレームワークでは、以下の５つを次元としており（表２）、TEHDAS の次元と比較すると、正確性（Accuracy）・信頼性（Reliability）を信頼性（Reliability）と表

17）日本製薬工業協会において、当該文書の翻訳版を公開しており、本稿では用語の和訳等の参考としている（2023年９月８日閲覧）、https://www.jpma.or.jp/information/evaluation/results/allotment/g75una0000001lqv-att/DBTF_202305_DQFEU_JP.pdf
18）EMA-HMAフレームワークの策定には、EUの規制評価プロセスのあり方が、文書ベースの提出から文書作成に使用した基礎データの評価（規制上の意思決定の目的に適合したデータかどうかを規制当局が評価）へと移行している背景がある。

表２　EMA-HMA フレームワーク：データの質に関わる次元

次元	定義	下位次元
信頼性（Reliability）	データが測定意図をどの程度正確に反映しているか	・精度（Precision）：どれだけ事実を代表しているかを示す近似度 ・正確性（Accuracy）：データと事実の間の乖離の大きさ ・妥当性（Plausibility）：ある情報が真実である可能性
広範性（Extensiveness）	データがどの程度十分にあるか	・完全性（Completeness）：データ収集プロセスやデータフォーマットを考慮したうえで想定される利用可能な全情報に対する実際の利用可能な情報量 ・網羅性（Coverage）：データ収集プロセスやデータフォーマットは考慮せず、現実世界に存在するものに対する利用可能な情報量
整合性（Coherence）	データセットの各部分の表現や意味が矛盾しないか	・フォーマットの整合性（Format Coherence）：同一形式でデータが表現されるか ・構造的整合性（Structural Coherence）：同じ実態が同一の方法で識別可能か ・意味的整合性（Semantic Coherence）：同一の値が同じことを意味するか ・一意性（Uniqueness）：同一情報が重複せずにデータセットに一度だけ現れるか
適時性（Timeliness）	医薬品規制上の意思決定において適切な時期にデータが利用可能か	・即時性（Currency）：どれだけ新しいデータか
該当性（Relevance）	研究課題にこたえるために有用なデータ要素をどの程度含んでいるか	—

注：整合性と厳密に関連する指標として、適合性（Conformance）と正当性（Validity：より狭義の適合性）が示されており、実際上、適合性が整合性を評価する最良の方法としている。
出所：EMA-HMA「Data Quality Framework for EU medicines regulation」[15] をもとに医薬産業政策研究所で作成

現するとともに、網羅性（Coverage）及び完全性（Completeness）を合わせた次元として、広範性（Extensiveness）を設定している。

① 信頼性（Reliability）
② 広範性（Extensiveness）
③ 整合性（Coherence）
④ 適時性（Timeliness）
⑤ 該当性（Relevance）

これらの次元の詳細を見ると、該当性（Relevance）を除いた４つの次元には、関連する下位次元が存在する。例えば、「データが測定意図をどの程度正確に反映しているか」を評価する次元である「信頼性（Reliability）」では、データがどの程度現実と合致しているかという質問に答えるため、精度（Precision）や正確性（Accuracy）、妥当性（Plausibility）といった下位次元を設定している。精度はどれだけ事実を代表しているかを示す指標であり、測定尺度（metric、例：年齢または月齢）の違いによって変わる。また、正確性はデータと事実の間の乖離の大きさを示す指標であり、例えば、体重の値が服の重さを差し引いているか否かにより変化する。さらに、ある情報が真実である可能性として定義される妥当性は、データセット全体で体重が300kg を超える場合が大半であった場合や妊娠記録に男性が含まれていた場合等、現実世界に起こりそうにない（または起こりえない）誤りの有無に関わる[19]。

ただし、データ生成や収集方法等は一様ではなく、データの質に寄与する要素の影響を考慮した次元評価が求められる。本フレームワークでは、この要素（決定因子）を以下の３つに分類している。

19）EMA、Multi-stakeholder workshop on Real World Data（RWD）quality and Real World Evidence（RWE）use、Data Quality metrics for Real-World Data（2023年９月14日閲覧）、https://www.ema.europa.eu/en/documents/presentation/presentation-data-quality-metrics-real-world-data-kdeli-ema_en.pdf

① 基礎的決定因子

データの生成プロセス（要件定義～収集・生成～管理・処理～公開～入手・集約～検査・受入～提供）及びシステムに関する因子であり、データセットの中身に依存しない。

② 内因性決定因子

特定のデータセットに備わっている固有の特性で、データの生成状況や利用状況に依存しない（例：小数点以下の桁数）。

③ 課題特有の決定因子

特定の課題（目的）に依存する因子。

一般的に基礎的決定因子がデータの質に直接的な影響を与えると整理されている。例えば、信頼性はデータの一次収集及びその処理のプロセスやシステムに依存し、広範性はデータ収集プロセスの仕様が影響する。また、整合性は単一組織のデータの場合、その組織全体のプロセスやシステムの同期に、複数のデータソースを統合する場合は、データ標準の利用に対するデータ生成組織のコミットメントに依存し、適時性はデータ収集並びに利用可能にするために用いられるプロセスやシステムにより決定される。

このように、データの質を適切に測定するためには、データ利用者の目的に応じて、評価すべき次元（下位次元含む）と許容可能な閾値を設定するとともに、次元に影響を与える決定因子を考慮した評価が求められる。

2－3－3. データの相互運用性

本稿のスコープからは外れるが、データの相互運用性について、TEHDASの「European Health Data Space Data Quality Framework」では、相互運用性はデータの高品質な二次利用のための必須条件であるが、品質にとって重要な機能とはみなされていないと言及している。同様に、EMA-HMA フレームワークにおいても、規制当局の意思決定に直接影響を与えない観点（簡潔性、アクセシビリティ等）や標準化等とともに、相互運用性は適用範囲外とされている。ただし、TEHDASの「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」では、データの効果的な二次利用に向けたデータガバナンスとして、データ保有者に対して、相互運用性（意味的相互運用性（交換されたデータや情報の意味が当事者間の交換を通じて保持され、理解されること）や構造的相互運用性（リテラル名、標準的な略語、エンコーディング等の構造的保持））の実装を求めている。

2－4. データ利用者による質の判断

2－4－1. EHDS：メタデータ・カタログの活用

製薬産業等のデータ二次利用者が、Fit for purpose の概念に基づくデータ利用を行うためには、二次利用者がデータの質を事前に判断できる仕組みが求められる。その仕組みの一つに、「メタデータ・カタログ」がある。メタデータとは、他のデータを特徴づける（データ特性を説明するための）記述的データであり、データセットの出所や範囲、主な特徴、健康データの性質及び電子的な健康データを利用可能にするための条件等、データ利用者がデータの質を評価できる情報が含まれる。これらの情報をカタログ化し、公開することは、データセットをプログラム的に検索可能にし、利用者の目的に合ったデータを適切に提供することを意味する。

欧州レベルで相互運用可能なメタデータ・カタログの整備に向け、EHDS では、各国の二次利用窓口となるHDABが、利用可能なデータセットとその特徴について、メタデータ・カタログを通じてデータ利用者に知らせなければならないと規定している（第55条）[16)、20)]。これにより、データ利用者が適切かつ容易に求めるデータを発見できることが期待される。加えて、EHDS第56条（Data quality and utility label）では、データ利用者が利用目的に合わせてデータセットを適切に選択でき

20）EHDSでは、さらに、HDABやその他のHealthData@EUの認定参加者が作成した各国のデータセットカタログを繋ぐEUデータセットカタログを欧州委員会が作成し、一般公開することを規定している（第57条）。

るよう、データセットの特性と潜在的な有用性を明確にするためのラベルを付与しなければならない場合があるとしており、以下の要素に準拠するデータ品質と有用性ラベルが検討されている。

① データの文書化:メタデータ、サポート文書、データモデル、データディクショナリ、使用された標準、出所
② 技術的品質：データの完全性、一意性、正確性、正当性、適時性、一貫性
③ 品質管理プロセス:レビュー、監査プロセス、バイアスの検証を含むデータの質管理プロセスの成熟度
④ 範囲：多領域にわたる電子的健康データの代表性、サンプリングされた集団の代表性、一人の自然人がデータセットに現れる平均的な時間枠
⑤ アクセスと提供に関する情報：電子的健康データの収集からデータセットに追加されるまでの時間、電子的健康データのアクセス申請承認後のデータ提供までの時間
⑥ データの充実に関する情報：他のデータセットとのリンクを含む、既存データセットへのデータの統合と追加

なお、EHDSの理念である国境を越えたデータ共有には言語の違いが課題となっており、メタデータ･カタログも例外ではない。保有されるデータの情報が現地の言語でしか入手できない場合に対し、英語による作成または自動ツール（例えば、欧州委員会の機械翻訳システムであるeTranslation）を活用したメタデータ・カタログの翻訳等により、各国のデータ利用者が情報を理解できるよう対策が検討されている[21)]。

このように、データ利用者が、自身の利用目的に合ったデータを適切に選択するためには、データ保有者がデータの質に関わる情報を標準化された状態で整理・公開することが不可欠である。EHDSへの参加は加盟国の義務（mandatory）であり、メタデータ・カタログ等によるデータ関連情報の公開に対し、データ保有者が必要な情報を提供する必要がある。一方で、データ保有者にとってそれらの対応は少なからず負担となるであろう。そのため、実際の運用に際しては、何らかのインセンティブ設計が必要である。EHDSでは、「料金」についても規定しており、データ保有者並びにHDABは、データ収集や二次利用にかかるコストを考慮した料金をデータ利用者に請求できるとしている（図５）。ただし、現状は概念の規定に留まっており、細かな料金体系（固定・自由料金等）や各国ごとの対応については、偏りが生じないよう欧州レベルでの設計が求められる。その足掛かりとして、TEHDAS（WP5:Sharing data for health）では、価格設定ルールの明確化に向けた議論が進められている。具体的には、フルコスト原理による料金設定や、欧州委員会によるHDABとデータ利用者間のデータ使用契約に対する欧州

図５　EHDSにおけるデータ保有者・HDABに対する料金設計

出所：「European Health Data Space」[16)] をもとに医薬産業政策研究所で作成

21）TEHDAS、TEHDAS requires clarity for cross-border provisions in EHDS（2023年９月22日）（2023年９月25日閲覧）、https://tehdas.eu/results/tehdas-requires-clarity-for-cross-border-provisions-in-ehds/

共通モデルの提供等が検討されている。

2−4−2. TEHDAS：メタデータ・カタログの整備に向けたアプローチ

TEHDASの「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」では、メタデータ・カタログの整備に対する３段階のアプローチを示している。第一段階では、分野やデータ種類にとらわれず、利用可能なデータセットに関するハイレベルな情報を収集することに重点を置く（例：DCAT（国際標準化団体W3C（World Wide Web Consortium）が勧告する国際的なメタデータ・カタログの標準[22]）のようなツールの活用）。第二段階では、様々な二次利用目的を考慮したデータセットの質及び有用性に関するさらなる詳細な情報を提供し、第三段階で、データソースの実際の内容（変数レベル）を踏まえた情報に重点を置く（例：Beacon（Global Alliance for Genomics & Health（GA4GH）により開発された、特定アレルを有するゲノム情報及び関連データの検索システム[23]）のようなツールの活用）というものである。EHDS 第55条で規定されるメタデータ・カタログを通じたデータ特性の公開に向け、上記のような具体的な実施法の開発が進められている。

なお、データ利用者は、データ二次利用の結果を、HDAB を通じて公開することが求められる（EHDS第46条）。さらに、TEHDASからは、データ利用者は、データセットを充実させ、デジタル・オブジェクト（データモデルや注釈、アルゴリズム等）を再利用できるような形で返却する必要性について検討すべきであることも提言されている[13]。

2−4−3. EMA-HMA：医薬品規制上のメタデータ・カタログの活用

2022年９月、EMA と HMA は共同で、「Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources」の初版を公開した[24]、[25]。これは、規制当局や研究者、その他の利害関係者に対し、医薬品規制上の意思決定に用いる RWD のメタデータの使用に関する推奨事項を提供する世界初のガイドであり、「規制目的での適切なエビデンス生成のためのデータソース発見を容易にすること」並びに「利用されるデータソースの適合性に関する情報への迅速なアクセスを提供し、研究プロトコール等の評価を支援すること」を目的としている。つまり、2−3−2項で言及したEMA-HMA データクオリティフレームワークがデータの質の評価の枠組みを規定したものであるのに対し、本ガイドは質の評価にあたっての可視化の方法（標準化されたデータカタログの形式[26]）を示したものと言える。なお、特定の状況を除き、データソースをメタデータ・カタログに登録する法的義務はない。ただし、データソースに関する公開情報がない場合、科学的信頼性や研究結果に対する社会的信頼性に影響を及ぼす可能性があることから、公衆衛生や医薬品規制上の目的での利用が見込まれる場合、データ保有者はデータソースの情報を登録・更新することが望ましい。

本ガイドでは、データソースの適切性評価に際

22）デジタル庁、「メタデータ導入実践ガイドブック」（2022年３月31日公開）（2023年９月25日閲覧）、https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/890e5d96-d63c-4b77-bd3e-cc89487393e3/1ead1a75/data_strategy_gif_469_guidebook_metadata.docx

23）Global Alliance for Genomics & Health、GA4GH Beacon project（2023年10月３日閲覧）、https://beacon-project.io/

24）EMA-HMA、「Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources」（2022年９月）（2023年９月15日閲覧）、https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf

25）Good Practice Guideの公開に先立ち、2022年５月にEMA-HMAから「List of metadata for Real World Data catalogues」が公開されている（2023年９月15日閲覧）、https://www.ema.europa.eu/en/documents/other/list-metadata-real-world-data-catalogues_en.pdf

26）本ガイドで言及されるメタデータ・カタログは、検索可能（Findable）、アクセス可能（Accessible）、相互運用可能（Interoperable）、再利用可能（Reusable）のFAIRの原則を順守し、他のメタデータ・カタログとの連携が可能であることを前提としている。

し、データの質を「信頼性」及び「該当性」の側面から区別する必要があるとしている。

① 一次データの信頼性に関する質
例：エラー、欠損値、非現実的な値の検出と修正、書式等、特定の調査目的とは独立したデータソースの特性
② 疫学的・統計的手法により、特定の研究課題に情報を提供する適切かつ有効なエビデンスをもたらすデータソースの該当性に関わる質
例：研究に必要なデータの有無、対象者数、集団の特徴、データ期間等、研究目的に依存するデータソースの特性

さらにデータの質を評価するために必要となるメタデータの具体的な項目についても本ガイドに記載されている（表３）。加えて、本ガイドではいくつかの利用者視点からメタデータ・カタログのユースケースを例示しており（研究に適したデータソースの特定、研究プロトコールの評価、研究報告書の審査等）、具体的な利用イメージを持つための参考とされたい。メタデータ・カタログはデータソースの信頼性を初期評価するための情報提供や特定の研究課題に対する有効なエビデンス生成のためのデータソースの該当性の初期評価への活用が期待される。

なお、本ガイドに基づくメタデータ・カタログは、European Network of Centres for Pharmacoepidemiology and Pharmacovigilance による既存カタログに置き換わり、2023年後半に公開される予定である。今後、製薬産業においても、医薬品規制上のデータ利活用のため、他のデータ保有者が公開するデータソースの評価や、製薬企業自身が保有するデータソースのメタデータ・カタログ公開への対応を検討する必要があろう。

2−5. 製薬産業の取り組み

欧州の製薬産業においても、データの質担保に向けた取り組みが行われている。2023年６月にEMAが開催した「Multi-stakeholder workshop on Real World Data（RWD）quality and Real World Evidence（RWE）use[27]」において、EFPIA（欧州製薬団体連合会）は、EMA-HMAデータクオリティフレームワークで提唱された次元を独自に評価した結果を発表している[28]。具体的には、EUnetHTA[29] が開発したレジストリ評価及び品質標準ツール（REQueST）[30] を用いて、EMA-HMA提唱の次元（該当性は研究に特化した次元であるため対象外）の評価可能性を検証した結果、「REQueSTを使用してデータクオリティフレームワークの次元を評価することは、患者レベルのデータ要素に間接的にアクセスする分散型レジストリでは複雑である」との見解を示している。この結果を踏まえ、EFPIAは、より効率的なデータの質判断と、全ての利害関係者（規制当局、レジストリ機関、産業界）の業務の持続可能性を調整することを目的に、データクオリティフレームワークを実施するためのデータの質評価と文書化の改善策について、EMA に３つの提言（①対話の主導、②入手可能であるべき最小限の情報に関するガイダンスの検討、③データクオリティフレームワークに沿ったツールの共同開発と試験的実施の主導）を行っている（図６）。

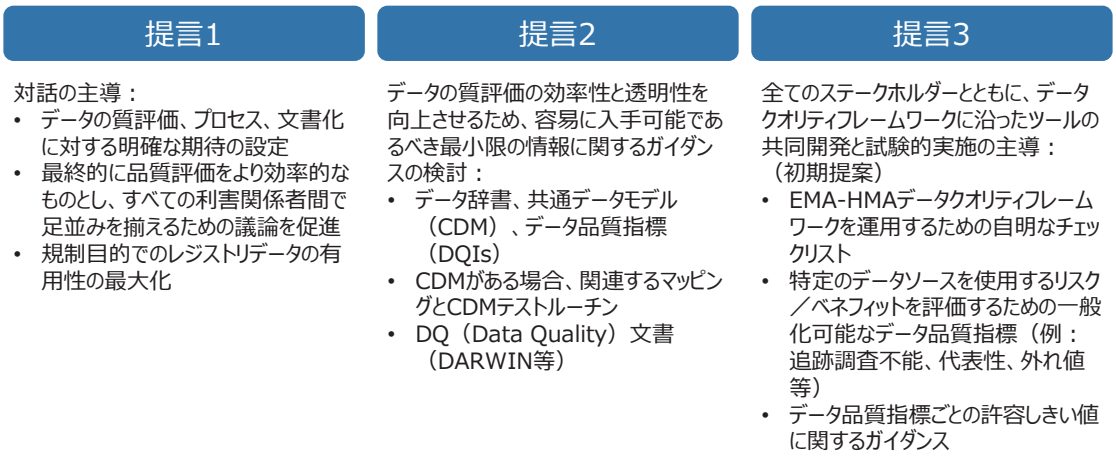
27）EMA、Multi-stakeholder workshop on Real World Data（RWD）quality and Real World Evidence（RWE）use（2023年６月26日～６月27日）（2023年９月15日閲覧）、本ワークショップは様々な利害関係者が集まり、リアルワールドエビデンス生成におけるデータ品質の測定と特徴づけに関する重要な課題について議論する場として開催された、https://www.ema.europa.eu/en/events/multi-stakeholder-workshop-real-world-data-rwd-quality-real-world-evidence-rwe-use
28）EFPIA、「Assessing the EMA data quality framework（DQF）dimensions using REQUeST: a decentralized registry use case」（2023年９月15日閲覧）、https://www.ema.europa.eu/en/documents/presentation/presentation-assessing-ema-data-quality-framework-dqf-dimensions-using-request-decentralized_en.pdf
29）EUnetHTA ホームページ（2023年９月15日閲覧）、EUnetHTA は欧州全土の医療技術評価のための効果的で持続可能なネットワーク構築を目的に設立された欧州共同組織である、https://www.eunethta.eu/
30）EUnetHTA、REQueST Tool and its vision paper（2023年９月15日閲覧）、https://www.eunethta.eu/request-tool-and-its-vision-paper/

表３　EMA-HMA：データソースを特徴づけるメタデータ

分類	情報		説明	メタデータ・カタログから提供される情報	
				信頼性（Reliability）	該当性（Relevance）
管理上の詳細情報	データソース名称		データソースの名称		
	データソース略称		データソースの略称		
	データ保有者		データソースのレコードの収集を維持する組織		
	データソースの連絡先名		データソースの問い合わせ先		
	データソースの連絡先Ｅメール		電子メールによる問い合わせ先		
	データ提供国		データ出所国		○（設定）
	データソース言語		言語		
	データソース地域		地域		○（設定）
	データソースが最初に構築された日付		最初の収集日とは異なる		○（時間要素）
	最初の収集日		データ収集または抽出を開始した日付		○（時間要素）
	最終の収集日		データ収集が終了した日付		○（時間要素）
	データソースのウェブサイト		データソースを説明する専用のウェブページ		
	データソースの出版物		データソース（検証、データ要素、代表性）または薬剤疫学的研究への使用について記述した査読済み論文または文書リスト	○	
	データソースの適格性		データソースが受けている正式な認定プロセス（EMA、ISO 等）	○	
	主な資金援助		自機関や国、欧州、産業界、患者団体等からの資金援助		
	データソースの種類		行政機関、プライマリーケア、セカンダリーケア、レジストリ、バイオバンク、副作用報告等		○（設定）
	データソースのケア設定		プライマリーケア、セカンダリーケア、入院治療、外来治療等		○（設定）
収集されたデータ要素	データソースの特徴	特定疾患	特定疾患の情報		○（アウトカム）
		入退院	入院及び／または退院情報		○（アウトカム）
		ICU 入室	集中治療室への入室情報		○（アウトカム）
		死因	死因情報		○（アウトカム）
		希少疾患	EU における有病率が１万分の５以下の希少疾患		
		処方及び／または調剤	医薬品の処方または調剤に関する情報		○（曝露）
		ATMP	遺伝子治療、体細胞治療、組織加工に関する医薬品の情報		○（曝露）
		避妊	避妊法の使用に関する情報		○（曝露）
		適応症	医薬品の適応症		○（曝露）
		ワクチン投与	摂取したワクチン情報		○（曝露）
		その他の注射剤投与	注射された医薬品情報		○（曝露）
		医療機器	医療機器情報		○（曝露）
		処置	医療処置情報		○（曝露）
		臨床測定値	臨床測定値（BMI、血圧、身長等）情報		○（アウトカム）
		医療従事者	投薬、手術、診断、治療サービスを提供する認可を受けた医療専門家または医療施設情報		
		遺伝子データ	遺伝子タイピング、ゲノムシーケンスの情報		○（アウトカム）
		バイオマーカーデータ	血液学的アッセイ、感染症マーカー、メタボロームバイオマーカー等の情報		○（曝露）
		患者の生成データ	患者（家族、介護者含む）が作成、記録、収集した健康関連データ		○（アウトカム）
		ヘルスケアの利用単位	サービス利用の定量値（年間受診回数、入院日数等）		○（アウトカム）
		人物の一意な識別子	患者の一意な識別子		
		診断コード	診断コード情報		○（アウトカム）
		妊娠・新生児	妊娠中の女性及び新生児（生後28日未満）、乳児、小児の発達データ		○（集団）
	収集される疾患情報		対象疾患情報		○（アウトカム）
	母集団の年齢層		新生児～成人を９段階に分類した年齢層情報		○（集団）
	家族とのつながり		世帯情報		○（集団）
	収集される社会人口統計学的情報		年齢や性別、民族、出身国、配偶者の有無等の情報		○（集団）
	ライフスタイル要因		タバコやアルコール消費量、運動量、食事等の情報		○（集団）
	データソースがカバーする集団		集計対象のうちデータソースがカバーする集団の割合		○（集団）
	データソースがカバーしていない集団		集計対象のうちデータソースがカバーしていない集団の割合		○（集団）
定量的記述事項	母集団の大きさ		データソースに記録されているユニークな人数		○（集団）
	年齢別集団		年齢層別の人数		
	アクティブ集団		記録が作成され、終了していない（アクティブな）患者の数		○（集団）
	年齢別アクティブ集団		年齢層別のアクティブな人数		
	時間の中央値		個人単位の記録の最初から最後までの時間の中央値（年単位）		○（集団・時間要素）
	アクティブな時間の中央値		アクティブな個人の利用可能な最初と最後の記録の期間の中央値（年単位）		○（集団・時間要素）
データフローと管理	ガバナンスの詳細		データ収集、管理、データへのアクセス、品質チェック及び検証結果等の全体的なガバナンス、プロセス、手順を説明した文書またはウェブページへのリンク	○	
	フォローアップ		生物試料の入手、入手条件等の情報	○	
	収集と記録のプロセス		調査ツールやデータソースへの収集・保存に用いるシステムの情報	○	
	記録の作成		データソースへの記録作成のきっかけとなるイベント（退院、専門医との面会等）		
	人物の登録		データソースに登録したきっかけのイベント（出生、疾病診断、治療開始等）		○（集団）
	人物の登録抹消		データソースから登録削除したきっかけのイベント（治療終了、移住等）		○（集団）
	連結		他のデータソースとの連結戦略や変数（患者 ID 等）等の情報	○	
	データ管理仕様		データソースのデータ検証の可能性	○	
	研究のためのデータ利用に関するインフォームド・コンセント		研究に対するインフォームド・コンセントの必要性		
	データソースのリフレッシュ		年単位の固定日での更新月の情報		
	データソースの最終更新		最終更新日		
	CDM（共通データモデル）仕様		変換に用いられる CDM（共通データモデル）情報	○	
	CDM へのデータソースの ETL（データ変換）		共通データモデル（CDM）へのデータ変換（ETL）に関する情報	○	
用語集と標準化辞書	入手可能な医薬品情報		医薬品情報（製品名、バッチナンバー等）	○（利用される場合）	
	使用される医薬品用語		欧州で認可された医薬品情報	○（利用される場合）	○（曝露）
	死因の用語集		死因に関する共通用語	○（利用される場合）	
	QOL（生活の質）の測定		QOL の測定法に関する情報	○（利用される場合）	
	処方箋用語		処方箋に関する共通用語	○（利用される場合）	
	調剤用語		調剤に関する共通用語	○（利用される場合）	
	適応症用語		適応症に関する共通用語	○（利用される場合）	
	処置用語		処置に関する共通用語	○（利用される場合）	
	遺伝子データ用語		遺伝子データに関する共通用語	○（利用される場合）	
	バイオマーカー用語		バイオマーカーに関する共通用語	○（利用される場合）	
	診断／医療イベント用語		診断／医療イベントに関する共通用語	○（利用される場合）	

注：該当性に関係するメタデータは、データソースの設定、集団、曝露、アウトカム、時間要素を評価するためのデータに分類される。
出所：EMA-HMA「Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources」[24] をもとに医薬産業政策研究所で作成

図６　EFPIA による EMA-HMA データクオリティフレームワークへの提言



出所：EFPIA「Assessing the EMA data quality framework（DQF）dimensions using REQUeST: a decentralized registry use case」[28] をもとに医薬産業政策研究所で作成

３．データの質確保に向けた日本の動向

ここからはデータの質確保に関連する日本の動向を見たい。

デジタル庁では、2022年３月に「データ品質管理ガイドブック」を公開している[31]。これは、データ品質指標に関するコンセンサスを日本国内で形成し、官民含めたデータ保有者が高品質のデータを提供可能とする環境の実現を目指すもので、海外とのデータ連携も見据えたデータ品質管理のフレームワークと評価モデルが示されている。本ガイドブックで提唱される「データ品質評価モデル」には、データそのものの評価と管理プロセスを含む。データそのものの評価には、国際規格であるISO/IEC 25012に基づく15の次元（正確性や完全性等）が設定されており、管理プロセスでは、データの品質計画、品質管理、品質保証、品質改善に加え、運用体制（データ関連サポート、リソース規定）も評価対象となっている。また、同じくデジタル庁が、2022年３月に公開した「メタデータ導入実践ガイドブック」では、主に行政機関のデータセットに付与すべきメタデータ・カタログの項目を例示している[22]。ただし、これらのガイドブックは医療ヘルスケア領域に特化した内容ではない。

医療ヘルスケア領域では、バイオバンクの取り組みが先行していると考える。現在、日本のバイオバンク・ネットワーク（９サイト：2023年６月時点）の試料・情報を横断的に検索可能なシステムが構築されており、研究目的への合致を利用者が判断するための情報（バイオバンクの協力者、試料、解析に関わる情報）が格納されている[32]。加えて、CANNDs（AMED が支援した研究開発から得られたデータの利活用を促進するためのプラットフォーム）では、Visiting解析環境において、ゲノムデータに紐づく属性情報（メタデータ：年齢、性別、出生地／居住地、疾患名）に基づき、解析対象となるデータの絞り込みが可能な検索システムの検討が進められている[33]。

制度面から見ると、本年５月に改正案が成立した次世代医療基盤法の検討 WG において、「データカタログの公開など、利活用者が情報を探索・活用しやすくなるような取組の在り方」が示され

31）デジタル庁、「データ品質管理ガイドブック」（2022年３月31日公開）（2023年９月25日閲覧）、https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/890e5d96-d63c-4b77-bd3e-cc89487393e3/0a34b7c6/20220810_policies_data_strategy_460_outline_14.docx
32）バイオバンク横断検索システム（2023年９月26日閲覧）、https://biobank-search.megabank.tohoku.ac.jp/v2/
33）AMED、データ利活用プラットフォームの提供サービス（CANNDs）について（2021年10月20日）（2023年９月26日閲覧）、https://www.kantei.go.jp/jp/singi/kenkouiryou/data_rikatsuyou/dai4/siryou3.pdf

ている[34]。匿名加工等を行う認定事業者が有するデータの情報をカタログとして公表することが提起されているが、具体的な項目については議論の途上にある。一方、内閣府のバイオ戦略においては、バイオデータ連携・利活用促進に向けた関係者間での共通認識の醸成を目指した「バイオデータ連携・利活用に向けたガイドブック」が公表されている[35)、36)]。この中で、データの整理と可視化に資するメタデータフォーマットの項目例（研究プロジェクト情報、データ／データベースに関する情報等）が示されている。ただし、「バイオ分野」には医療ヘルスケア領域のみならず、農業分野等の幅広い領域が含まれることに留意が必要である。

わが国において、上記のような取り組みが進められているが、医療ヘルスケア領域に特化したデータクオリティフレームワークの策定やメタデータ・カタログの整備に向けたさらなる検討が必要と考える。

4．まとめ・考察

データの質は一義的に決まるものではなく、各々のデータ利用目的に適合する質が求められる。TEHDAS並びにEMA-HMAのデータクオリティフレームワークでは、該当性（Relevance）、正確性（Accuracy）／信頼性（Reliability）、整合性（Coherence）、適時性（Timeliness）に加え、TEHDASでは網羅性（Coverage）及び完全性（Completeness）、EMA-HMAでは広範性（Extensiveness）をデータの質の決定に関わる次元として規定していた。一方、データ利用者が自身の利用目的に合ったデータを適切に選択するためには、データの特性を説明する標準化されたメタデータが公開されていることが重要であり、欧州では、加盟国共通のメタデータ・カタログや医薬品規制において利用されるデータに焦点を当てたメタデータ・カタログの構築が検討されている。なお、データの質確保に向けた取り組みを進めるため、TEHDASの「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」では、データ保有者やHDAB、データ利用者等が法的に対応すべきまたは推奨される行動が示されている（表４）。

このような欧州での取り組みを踏まえ、データの質の面からわが国のデータ二次利用を促進するためには、

① ステークホルダー間でのデータの質に対する共通認識の醸成（決定要素等の合意）
② データ利用者が利用目的に合わせたデータを選択するためのメタデータ・カタログの標準化と公開の仕組みの整備

の２点が重要と考える。以下に、この２点を推進するための方策について、筆者の考えを述べたい。

① ステークホルダー間でのデータの質に対する共通認識の醸成

Fit for purposeの概念に基づく健康医療データの二次利用を促進するためには、データの質の定義や決定要素、データ保有者の成熟度等について、データ保有者や管理者、利用者があらかじめ共通認識を醸成する必要があると考える。そのためには、国民やデータを保有する医療機関、アカデミア、データ管理･監視に関わる国、規制当局、データベンダー、利用者となる製薬企業等、様々なステークホルダーが参画する検討プロジェクトを設置し、医療ヘルスケア領域に特化したデータの質に関するガイドラインを策定することが望ましい。例えば、本稿で言及したTEHDASには、EU加盟国21カ国とその他欧州４カ国から、規制当局、

34）首相官邸 健康・医療戦略推進本部、第６回 次世代医療基盤法検討ワーキンググループ（令和４年６月６日開催）参考資料２（中間とりまとめ）（2023年９月25日閲覧）、https://www.kantei.go.jp/jp/singi/kenkouiryou/data_rikatsuyou/dai6/sankou2.pdf
35）内閣府、バイオ戦略関連資料 バイオデータ連携・利活用に向けたガイドブック１（令和５年６月）（2023年９月25日閲覧）、https://www8.cao.go.jp/cstp/bio/data_renkei_1.pdf
36）内閣府、バイオ戦略関連資料 バイオデータ連携・利活用に向けたガイドブック２（令和５年６月）（2023年９月25日閲覧）、https://www8.cao.go.jp/cstp/bio/data_renkei_2.pdf

表４　データの質確保に向けた各ステークホルダーの実践（一例）

アクター	主要な要素	実践	主なガバナンス機構	
			推奨	法的義務
データ保有者	データの質の管理及び保証	能力成熟度モデルの導入	○	
	データセットの説明	国際的なメタデータ標準の実装とメタデータ仕様に従った全データセットの公表		○
	品質と有用性の表示（データ品質と有用性ラベル）	自己評価を実施し、公表		○
Health Data Access Bodies	データセット・カタログ	共通の仕様に従ったメタデータの相互運用可能な公開		○
	ラベリング管理	データの質及び有用性の管理(必要に応じて外部監査と認証を調達)		○
	データ保有者の成熟度の管理	データ保有者の成熟度モデルの実施に関するガイダンスの提供及び監督	○	
データ利用者	充実した／注釈付きデータセットの返却	出所に関する報告	○	
	デジタルオブジェクトの返却	オープンサイエンスを利用した FAIR by design	○	
	利用者の経験	データの質及び有用性ラベルでの経験	○	

注：TEHDAS「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」[13] に記載のうち、データ保有者、Health Data Access Bodies、データ利用者に対して、本稿で言及したデータの質担保に関わる事項のみを抜粋しており、表４の他にも求められる対応があることに留意が必要である。
出所：TEHDAS「Recommendations on a Data Quality Framework for the European Health Data Space for secondary use」[13] をもとに、医薬産業政策研究所で作成

国営のデータ管理組織、アカデミア、学会、民間企業、市民団体等が参加している。わが国でも広範なステークホルダーが自分事として検討する連携体制の構築が重要と考える。

また、資金面での支援も考えなければならない。欧州における様々な分野の研究・イノベーションのための資金助成プログラムである Horizon Europeでは、データ利用者、データ保有者、HDAB及び健康データの二次利用の範囲に関連するその他の関係者の代表で構成されたコンソーシアムを立ち上げている。この中で、幅広いデータタイプやデータ保有者の負担を考慮し、EHDS において提案されているデータ品質と有用性ラベルに関するフレームワークの開発や提案されたフレームワークの試行による最適化等の検討が進んでいる[37]。この活動に対しては、400万ユーロの資金が投入されており[38]、実効性のある検討を継続するにあたっては公的な資金援助も重要な視点となろう。

② メタデータ・カタログの標準化と公開

データ利用者が利用目的に合ったデータセットを適切に選択するためには、データの質の判断項目が整理され、標準化されたメタデータ・カタログとして公開されていることが望ましい。メタデータ・カタログの整備に向けては、二次利用目的で求められる質を考慮したデータ特性項目（メタデータ・カタログに含むべき情報）をステークホルダー間で整理するとともに、データ利用者による情報への容易なアクセスの確保やデータ保有者の負担軽減のため、必要な情報をデータ保有者から一元的に収集･整理し、公開する公的機関（欧

37）European Commission、Funding & tender opportunities「Developing a Data Quality and Utility Label for the European Health Data Space（TOPIC ID: HORIZON-HLTH-2023-TOOL-05-09）（2023年９月15日閲覧）、https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-hlth-2023-tool-05-09
38）Ministère de l'Enseignement supérieur et de la Recherche、Le site français du programme européen pour la recherche et l'innovation「Developing a Data Quality and Utility Label for the European Health Data Space」（2023年９月15日閲覧）、https://www.horizon-europe.gouv.fr/developing-data-quality-and-utility-label-european-health-data-space-33805

州におけるHDABのような窓口）を設置すべきと考える。（なお、日本国内でのデータ利用に限れば、言語の違いは問題とはならないが、国境を越えたデータ利用を想定する場合、メタデータ・カタログの翻訳の仕組みも検討すべきであり、欧州のようにHDABがメタデータ・カタログを一元管理することのメリットがあろう。）

また、質を判断するための具体的なメタデータ・カタログの構成については、製薬産業を含むデータ利用者が各々の利用目的で求める具体的な質をデータ保有者や管理者等と共有することが重要と考える。つまり、データ利用者が、個別のニーズに対し、どういった質を求めているか、どういった視点で質を判断するか等について、分かりやすく提示し、データ保有者、管理者、利用者が共通理解を醸成したうえで、妥当なメタデータ・カタログの項目を設定することが望ましい。EMA-HMAの取り組みも参考に、製薬産業において、具体的なユースケースベースでのデータの質に対する情報発信が求められよう。

ただし、メタデータ・カタログの作成にあたっては、少なからずデータ保有者や管理者の負担が生じる。EHDSでは、データ提供に対する「料金」について規定していると述べたが、わが国においてもメタデータ・カタログの整備も含めたデータ二次利用の推進に対するインセンティブ（金銭的インセンティブ等）を検討する必要があると考える。

5．おわりに

本稿では、健康医療データの二次利用に向けた「データの質」に関して、欧州での取り組みを俯瞰した。わが国では、規制改革推進会議をはじめとする様々な会議体においてEHDSを参考とした日本版 EHDS の構築に向けた議論が進められている。しかしながら、現状、データの二次利用目的や対象となるデータの種類、同意要否を含むデータガバナンス等、データ利活用の概念や仕組み、基盤に関する議論が中心であり、データの質に関する議論は十分とは言えないと考える。日本におけるデータの二次利用をさらに推進するためには、制度政策の整備に加え、実際に利活用される健康医療データの質の確保が欠かせない。本稿で取り上げた欧州でのデータの質に対する考え方や取り組みを参考に、今後、わが国においてもデータの質に対するステークホルダー間での共通理解の醸成や利用者によるデータの質判断のためのメタデータ・カタログの整備等に向け、国全体で連携した対応が求められよう。

また、本稿ではデータの質に着目したが、データの越境利用の観点から見た場合、欧州で検討が進むEHDSは欧州内の自由なデータ流通のためのデータスペースであるのみならず、規格に適合すると認めた域外の機関からのアクセスも認めることから、EHDS 関連の動向は日本にも少なからず影響があると言える。製薬産業視点で言うと、日本のデータ流通基盤や企業等のアクセスが許可された場合、国境を越えたデータ利用が進み、日欧を含めた国際的な医薬品開発の促進が期待されるであろう[39]。一方で、日本のアクセスが許可されない場合、データ流通・利活用の面で他国に取り残されてしまう懸念がある。データの質も含め、わが国における健康医療データ利活用の環境整備にあたっては、国内に閉じたデータ流通システム・制度ではなく、EHDS のような国際動向も考慮したデータ基盤の視点が不可欠と言えよう。

本稿がわが国のデータ利活用を促進する一助となることを祈る。

39）今後の議論により、日本の製薬企業が欧州での治験実施等でデータ保有者と見なされた場合、データ保有者としての義務（EHDS 第３条：自然人がデータに容易にアクセスできるようにする、データを電子化する等、第41条：データを利用可能とする、提供要請に対し、決められた期限内でデータ提供する等）への対応が求められる可能性がある。