

共同企画 | 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） | 共同企画

共同企画1

臨床データベースから見えてくること：ビッグデータを医療に活かす

2020年11月19日(木) 09:20 ~ 11:20 A会場 (中ホール)

[2-A-1-01] データ駆動型臨床研究は、健康・医療をどう変えるのか？

*中島 直樹¹ (1. 九州大学病院 メディカル・インフォメーションセンター)

*Naoki Nakashima¹ (1. 九州大学病院 メディカル・インフォメーションセンター)

キーワード：Real World Data, Data Driven Medical Study, Big Data, Data Quality, Artificial Intelligence

社会に蓄積したデータ（リアルワールドデータ、RWD）を活用した研究をデータ駆動型臨床研究と呼ぶ。前向き臨床研究と比較した場合の特長は「大きいデータ規模」「入力不要」「安いコスト」「早い結果」「悉皆性の確保可能」「探索的研究が可能」である。一方欠点は「低いデータの品質」「必要なデータ項目がない」「バイアスが管理困難」「個別同意取得困難」であり、いずれも医学研究としては致命的障害ともなる。その限界を認識しつつ欠点を可能な限り改善し、従来の臨床研究手法と補完的に用いるべきだ。近年、これら欠点の改善対策が精力的に講じられ一定の成果を得てきた。

日本では公的な事業である NDB や MID-NET の整備が先行した。これらは純粋な後ろ向き研究であり、徹底した匿名化により前向き研究への移行や他のデータベースとの融合は不可である。続いて進んだのが AMED 等の競争的研究であり、その一つが J-CKD-DB（Ext含む）である。戦略的なデザインにより、前向き研究への移行（J-CKD-DB-Next、Next）や他のデータベース（腎臓病総合レジストリー、J-KDR）との融合も可能とした。将来的には、データ規模を効率の良い J-CKD-DB でカバーし、データの深さを Next や J-KDR で補填することができる、日本においては全く新しい手法である。

米国政府は積極的に RWD を活用する方針を打ち出し、FDA は 2019 年には RWD のみに基づき CDK4/6 阻害薬の男性乳癌への適応拡大を承認した。MID-NET も現在の薬剤の副作用検証から、副作用シグナル早期検知へと用途を拡大する。これらは臨床薬理学に則った展開であるが、データ駆動型臨床研究の守備範囲は、薬理学のみならず、臨床医療経済学から臨床生物学に至る大きな分野である。従来の臨床研究手法では把握・解明できなかった Unmet Medical Needs を扱える期待も膨らむ。

データ駆動型臨床研究は、健康・医療をどう変えるのか？

中島 直樹^{*1}

^{*1} 九州大学病院 メディカル・インフォメーションセンター

The Impact of Data-Driven Medical Study on Healthcare and Medical Services

Naoki Nakashima MD PhD ^{*1}

^{*1} Medical Information Center, Kyushu University Hospital, Fukuoka, Japan

In addition to the prospective medical study (PMS), a data-driven medical study (DDMS) has been expected to improve aging society and other social issues in Japan. Comparing with the PMS, the benefits of the DDMS include the following: (1) much larger data size, (2) no need to input data, (3) low cost, (4) rapid outcome, and (5) exhaustive and exploratory survey. On the other hand, the disadvantages include the following: (1) low-quality data, (2) absence of important data items, (3) unmanageable bias, and (4) difficulty to get agreement to use data. Any of these disadvantages can be a setback of the research.

Recent efforts to improve these disadvantages in experience of large-size DDMSs have increased the value of the DDMS. DDMS strategies have also evolved; now, we can use a layered database structure like J-CKD-DB to increase analyzing power and efficiency.

The current DDMS will be connected to the development of artificial intelligence in the future and then expected to improve the safety and quality of medical services. Ultimately, we should clinically integrate scientific specialties coming apart in each field in a human-centered manner while supporting medical staff and considering patients' feelings.

Keywords: Real World Data, Data Driven Medical Study, Big Data, Data Quality, Artificial Intelligence

1. はじめに

社会に蓄積したデータ (Real World Data、以下 RWD) の有効な 2 次利活用の要求が強まっている。医学では、電子カルテやレセプトなどの診療データや健診データにあたるが、これを臨床研究に用いる場合には、データ駆動型臨床研究 (Data Driven Medical Study、以下 DDMS) と呼ばれる。俗に「ビッグデータ研究」とも言われる¹⁾。人類史上初めて超少子高齢社会が到来した日本にとり、健康寿命の延伸、高騰し続ける医療費の適正化や医療安全の確保は喫緊の課題であるが、DDMS はその解決の有効な手法としても期待されている。また、近年増加傾向にある自然災害や 2020 年に爆発的に拡大した COVID19 など Pandemic などの緊急事態への対応にも応用されるであろう。さらにこの手法は、今後実現するであろう人工知能支援医療にも直結するため、倫理や社会状況に対して適正な方針、データの精緻化や解析技術の進化が求められる。

図1に研究に供されるデータ源についての分類を示したが、前向き研究と DDMS は対極にあることが理解される。

本稿では DDMS の長点と欠点を示した後にその課題への取り組みを紹介し、さらにはデータ駆動型臨床研究がどのように発展し、どう健康・医療に影響するのか、概観を述べたい。

2. データ駆動型臨床研究の課題と対策

日本における代表的な DDMS 例として NDB (レセプト情報・特定健診等情報データベース、National Database)²⁾、MID-NET (医療情報データベース、Medical Information Database NETWORK)³⁾などが挙げられる。これらの事業を有効に活用するためにも、さらには後続のデータ駆動型臨床研究を精緻化するためにも、課題への対策が講じられてきた。



図1 研究に用いるデータの分類。横軸が時間軸で、縦軸はデータ源である。

DDMS は RWD を活用するため、当然後ろ向き研究になるが、前向き臨床研究と比較した場合の特長は、

- 1) データ規模が大きい
- 2) 入力が必要
- 3) コストが低い
- 4) 結果が早く出る
- 5) 悉皆性の確保が可能
- 6) 探索的研究が可能

などである。一方でその欠点は、

- 1) データの品質が低い
- 2) 必要なデータ項目がない
- 3) バイアスの管理が困難
- 4) 個別同意の取得が困難

などであり、そのいずれもが医学研究には致命的な障害ともなりうるものである⁴⁾。

従って、データ駆動型臨床研究に際しては、その限界を認識し、欠点を可能な限り改善しながら、従来の臨床研究手法と補完的に用いるべきである。近年、これら欠点の改善対策が精力的に講じられ、一定の成果が得られてきている。

2.1 データの品質が低い

観察的疫学研究報告の質改善(STROBE)のための声明には、データ品質に関する項目が多い⁵⁾。データ駆動型臨床研究の場合にも必ずチェックするべき資料である。

表1にDAMA(Data Management Association)が2009年に策定したデータ管理の機能や活動の枠組み「DAMA-DMBOK(Data Management Body of Knowledge)」を示した。なお、このDMBOKにおける品質の高いデータとは、「ビジネスの目的のためにデータが利用に適した状態」とであると定義している⁶⁾。

表1 データ品質の主要な評価軸例。出典:DAMA-DMBOK Functional Framework Version 3.02 (DAMA International, September 2008)を基にKPMGが作成⁶⁾

評価軸	概要
正確性	データが表そうとしている実体が正しく示されていること
完全性	すべてのデータ要素が揃っていること
一貫性	同じ実体を表す2つ以上のデータに不整合がないこと
最新性	データが期限内の実体を示していること
精度	データの詳細度(有効桁数など)が十分であること
プライバシー	アクセス制御と利用監視がなされていること
妥当性	対象の業務内容においてデータの整合性が取れていること
参照整合性	参照元のデータが存在すること
適時性	必要な時に速やかにデータが利用できること
一意性	同じ実体を表すデータが1つだけ存在すること
有効性	データが定められた属性(型・形式・精度・文字コード等)が有効範囲に収まっていること

表1に示すように、データ品質と言っても多くの評価軸がある。データ品質を論じる際にどの評価軸を指すのかを明らかにしておかなければならない。特にRWDでは十分な品質を得ることが不可能あるいは困難な評価軸が存在することが多い。

例えばRWDでは「完全性=すべてのデータ要素が揃っていること」を満たすことは難しい。欠損値の処理を慎重に行う必要がある。また、気が付かないうちに複数の評価軸が劣化していることが多い。ここでは、MID-NET事業で表出した「一貫性」の低下について説明する。

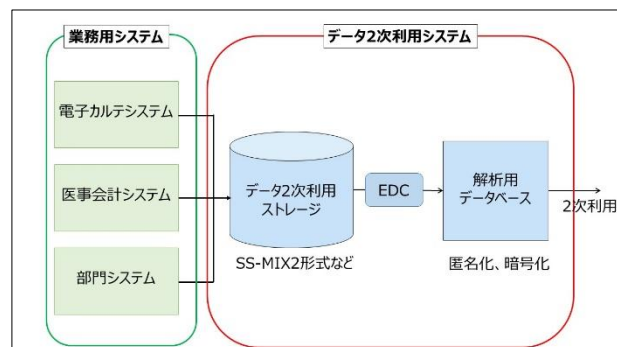


図2 MID-NET等のDDMSに共通の基本構造。データ2次利用のためのストレージにデータを溜めて、Electronic Data Captureによりデータセットを作る。

MID-NET事業は、データ駆動型手法により薬剤副作用の科学的な検知に特化した厚生労働省/PMDA事業で、10の病院グループ23病院が参加する505万人データベース(2019年末)である³⁾。基本的な構造を図2に示すが、この構造は、SS-MIX2をベースにしており、日本腎臓学会のJ-CKD-DBと同様で、2018年度から構築が始まった臨床研究中核病院ネットワーク事業とも類似している。

2013年度からMID-NET事業において、病院情報システム(HIS)からSS-MIX2を経て「統合DS」へとデータが流れる際のデータの一意性の検証を実施したところ、ある施設では図3左のような一意性の低下が認められた。

そこで、MID-NET事業では品質管理作業をPMDAと協力医療グループが共同で2015年まで3年間行った結果、ある施設では図3右のようなデータ一意性の向上を認めた。その原因は多岐にわたり、①標準コードへのマッピングの問題、②HISカスタマイズや独自運用の問題、③SS-MIX2への送信エラー、④診療情報以外のデータ混入、⑤統合DS取込のエラー、など合計300種類以上がみられた。この一連の作業により、DDMSにおける2次利用データの一意性の検証が有効なことが証明された³⁾。なお、これらのエラーはSS-MIX2に特異的な原因によるものは少なく、どのような2次利用ストレージを用いる場合にもでも起こり得るものが多数あり、データ2次利用の際には、このようなデータ一意性の低下には注意するべきであることを認識した。

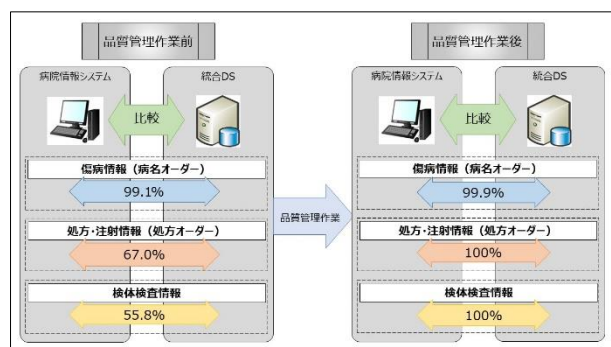


図3 バリデーション事業におけるデータ品質管理作業の前後のデータ一致率。PMDA発表資料より改変。

さらに、一旦データ品質管理が達成された後も、通常の病院業務の運営継続により当然発生する新規薬剤や後発品薬剤の導入、新しい検査方法や検査試薬の導入、HISの

カスタマイズや更新によって、データ一貫性は日々劣化するため、バリデーション作業は継続しなければならない。MID-NET 事業では、AMED 事業（康東天班 20mk0101157h0002）にて九州大学病院にガバナンスセンターを設置し、MID-NET 協力病院グループ同士の標準コードマッピングの整合性を継続的に取ることを試みている。また、これらのバリデーション作業経験を通じて確立した品質管理のための手法に関して AMED 事業（中島直樹班 20mk0101158h0002）では教材を作成し提供している⁷⁾。

このように、いつでも使える形としての 2 次利用データ収集には、経験とコストが必要であることを認識しておく必要がある。今後、病院の診療記録管理担当や検査部、薬剤部、研究者やシステムベンダーが経験を積むことによるエラーの減少、および標準コードの運用の社会全体での整備、などによりコストが減少することが望まれる。

2.2 必要なデータ項目がない

研究に必要なデータ項目の不在は、広くはデータ品質の低下に含まれる。例えば、血圧研究や腎臓病研究などに重要な「血圧項目」の不在などがそれにあたる。

また、表 1 における「正確性＝データが表そうとしている実体が正しく示されていること」とも関連する。例えば、電子カルテには正しい病態を表す病名が構造化データとして存在しない。そこで、それ以外に病名を表す項目がないために保険病名をそのまま用いることが多いが、保険診療上、様々な目的で不正確な保険病名が付与されるために正確性が低下する。

これを補うために、ePhenotyping という手法を実施することがある。保険病名のみならず、処方や検査結果、あるいはレセプト上の加算などの構造化データから、一定のロジックを組んで陽性的中率 (PPV) や感度などを向上する手法である⁸⁾。例えば、「2 型糖尿病」という保険病名だけを持っている症例よりも、処方まで出されている症例を抽出した場合の PPV が向上することで理解される⁹⁾。

厚生労働科研宇山佳明班 (20KC2007) では、MID-NET に用いる薬剤副作用病態を検知するための ePhenotyping 開発研究を続けている。

なお、このような ePhenotyping 手法は DDMS の精度を上げるために当面は用いられるべきであるが、その根本は医事会計システムから発達したレガシーシステムの弊害である。正確な病名・病態をリアルタイムに記録する次世代型の電子カルテの開発が必要である。そもそも病名をつける行為「診断」は医師の有する最も重要かつ崇高な業務の一つである。これを正確に記録する機能が無いことは、現在の電子カルテの最大の欠陥と言わざるを得ない¹⁰⁾。

2.3 バイアスの管理が困難

DDMS では、RWD データを蓄積した目的と、2 次利用による研究解析の目的が違うためにバイアスが幾重にも生じることが多い。例えば大学病院の電子カルテデータでは、医療全体の平均的な症例に比して、重篤疾患、難病疾患、複合疾患などの症例が多い、というバイアスがある。これをそのまま用いて医療全体を代表させる結果を導くことは難しい。そのため、背景などによりマッチングを行う手法が発展してきた。傾向 (プロペンシティ) スコアを用いた手法で、ネステッドケースコントロール研究、ケースコホート研究などがある¹¹⁾。

2.4 個別同意の取得が困難

DDMS は後ろ向きデータであるため、研究開始時には患者の症例数が莫大なこと、および転院や死亡などにより、アプローチが困難なことなどから個別同意を取ることが極めて難しい。公的なデータ駆動型事業では個人情報保護法以外の法制度に基づいて個人情報の利活用が行われることが多い。例えば、レセプト情報・特定健診等情報データベース (NDB) は、「高齢者の医療の確保に関する法律」に、また MID-NET 事業は、PMDA 法に基づき実施されている。

研究の場合には、医学系研究に関する倫理指針では、倫理審査委員会の承認によりオプトアウト (掲示などによりデータの 2 次利用や第三者への提供を本人の求めに応じて停止すること機会を与えること) が認められている。

また、2018 年度には、次世代医療基盤法が実施された。同法では、通知によるオプトアウト (リーフレットなどで個別に通知することより、データの 2 次利用や第三者への提供を本人の求めに応じて停止すること機会を与えること) を行っておけば、データを取得した健康・医療関連機関は実名で個人データを認定匿名加工医療情報作成事業者 (認定事業者) に提供することができ、認定事業者は他のデータ源からのデータと名寄せをしたうえで匿名加工化し、研究者や企業へ有償で提供することができる。これは 2017 年度の個人情報保護法改正により要配慮個人情報の第三者提供のハードルが上がったために研究や企業開発が後退しないことを目的として施行された。現在までに、2 つの認定事業者が承認されている。

以上のように、DDMS の課題を解決する努力により、RWD の欠点が縮小し、その長点を享受する可能性が増大している。

3. DDMS の事例

米国政府は積極的に RWD を活用する方針を打ち出し、FDA は 2019 年には RWD を活用した DDMS のみに基づき CDK4/6 阻害薬の男性乳癌への適応拡大を承認した¹²⁾。

MID-NET 事業も現在の「薬剤の副作用検証」から、「副作用シグナル早期検知」へと用途を拡大する方針である。

日本では公的なデータ駆動型事業である NDB や MID-NET 事業の整備が先行した。これらはともに実績が重ねられているが、徹底した匿名化により前向き研究への移行や他のデータベースとの融合は不可である。

続いて進んだのが厚生労働省臨床効果データベース事業や、AMED 研究事業等の競争的研究である。その一つが日本腎臓学会による J-CKD-DB である¹³⁾。

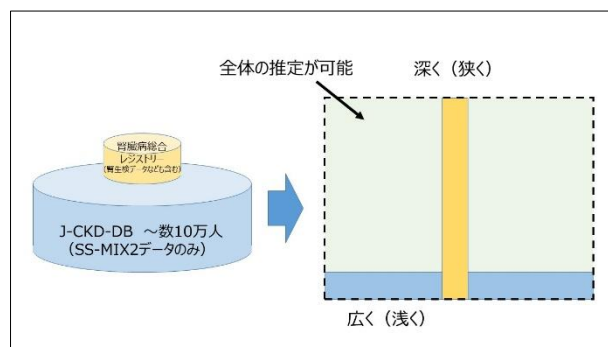


図4 左: J-CKD-DB の包括的・重層的腎臓病 DB 研究計画 (川崎医大・柏原直樹氏スライドより改変) と右: 高効率の RWD 活用戦略

図4に示したように、日本腎臓学会は J-CKD-DB を包括的な DB と捉え、データ項目は限定されていても SS-MIX2 項目のみを収集することにより臨床現場の入力をゼロとし、可能な限り広い症例数を集めることに注力した。また、その J-CKD-DB を1階部分と考えると、2階、3階の症例数は1階ほど多くはないが、腎生検など詳細なデータを含むデータベースあるいは IgA 腎症、糖尿病腎症など疾患別のデータベースを重ねることにより、高効率に CKD 全体を把握することが可能な DDMS を構築している。日本においては全く新しい手法であるが、既に 10 万例を超える症例データを収集し、ベースライン研究も報告され始めた^{14, 15)}。今後も定期的にデータを収集し、縦断的研究および重層データベースを併せた高度解析を実施する予定である。

4. DDMS は何を変えるか？

現在、世界中で情報革命(デジタルトランスフォーメーション: TX)が拡大している。医療もその例に漏れず、Medical TX: MTX が進みつつあるが、その方向性は、

- 1) 空間軸の拡大: 医療施設から日常空間へ
- 2) 時間軸の拡大: 生涯の健康医療データの統合へ
- 3) 領域の拡大: 医療から健康領域・福祉領域・その他へ
- 4) 医療主権の移動: 患者主体医療へ

などが挙げられる。これらを支援するのが健康医療における患者とのコミュニケーションツールである PHR (Personal Health Record) であり、患者の日常データを収集する IoT 技術である。またこれらにより爆発的に増えるデータを活用して医療を支援する人工知能 (AI: Artificial Intelligence) である。DDMS にとって PHR や IoT は大きなデータ源として期待される。

やみくもにデータを集めた RWD では、2 章に示したように決して評価される(つまり社会に有用な)DDMS になり得ない。データ2次利用を行うことを前提として、より戦略的に効率的に品質の高いデータを集積することが重要である。その一例が J-CKD-DB の戦略であり、また次世代の電子カルテに真の病態をリアルタイムに記録する方向性であろう。また、現在日本医療情報学会が日本臨床パス学会と進めている医療プロセス・アウトカムの解析が可能な標準臨床パス (ePath) もそのような戦略と捉えることができる⁹⁾。つまり、これらにより RWD で効率良く、品質の高い2次利用データを収集することが可能となるのである。

ICT 化により情報量が増大し、CT レポートの見落としや処方ミスなどによる医療訴訟も起きるなど社会の混乱がみられる健康医療領域で、DDMS が、いかに現場に負担をかけずにデータを収集し、逆にその混乱を解消するべく臨床現場を支援できるか、が鍵である。

DDMS は究極には AI 開発に繋がる。永井良三氏は、次世代医療 AI の役割は、各科でバラバラになった高度な科学的医学の臨床的統合を人間中心に復活することであり、AI は医療者に対しては支援し、患者にも寄り添うべき、と述べている¹⁶⁾。

5. おわりに

データ駆動型臨床研究の守備範囲は、臨床医療経済学から臨床生物学に至る大きな分野である。従来の臨床研究手法では把握・解明できなかった Unmet Medical Needs を扱え

る期待も膨らむ。

謝辞

本研究は日本腎臓学会、日本糖尿病学会、および日本医療情報学会の合同委員会の企画による。各学会の関係各者に心より感謝申し上げる。

参考文献

- 1) 宮野悟. ビッグデータ駆動型のサイエンスのインパクト. 実験医学 34(5): 676-680, 2016.
- 2) 康永秀生. ナショナルデータベース (NDB) および DPC データを用いた臨床疫学研究. 医学のあゆみ, 265(11): 931-933, 2018.
- 3) Mitsune Yamaguchi, Satomi Inomata, Sayoko Harada, et al. : Establishment of the MID-NET? medical information database network as a reliable and valuable database for drug safety assessments in Japan, Pharmacoepidemiol Drug Saf., 2019. doi: 10.1002/pds.4879. [Epub ahead of print].
- 4) 中島 直樹. PHR を活用した糖尿病療養指導 編集: 門脇 孝, 山内 敏正, 糖尿病 2020 診断と治療社, 120-129, 2020.
- 5) STROBE 声明 [<https://www.strobe-statement.org/fileadmin/Strobe/uploads/translations/STROBE-Exp-JAPANESE.pdf>] cited 2020-Sep-6]
- 6) KMPG ジャパン. [<https://home.kpmg/jp/ja/home/insights/2016/12/data-management.html>] (cited 2020-Sep-6)]
- 7) MID-NET を用いた医薬品等のベネフィット・リスク評価のためのデータ標準化の普及に関する研究. 中島直樹、山下貴範編、Masterpiece 社 2020, ISBN: 9784909288547
- 8) Conway M, Berg RL, Carrell D et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc. 2011: 274-83, 2011.
- 9) 伊豆倉理江子, 山下貴範, 野尻千夏, 高田敦史, 野原康伸, 中島直樹. 病態・疾患特定のためのアウトカム定義 (phenotyping) 調査研究. 医療情報学 37 (Supple.), 1007-1010, 2017.
- 10) 中島直樹. 診療データは溜まるのか、貯めるのか? 医療情報学 39 (Supple.) 1-5, 2019.
- 11) 野間久史. ケースコホート研究の理論と統計手法. 統計数理. 2014; 62(1): 25-44.
- 12) Yanmei Wu, Yu Zhang, Hao Pi, et al. Current Therapeutic Progress of CDK4/6 Inhibitors in Breast Cancer. Cancer Manag Res. 12: 3477-3487, 2020
- 13) 柏原直樹. 腎臓内科 日本の CKD を“可視化”する - ICT を活用した腎臓病データベースの構築: J-CKD-DB. 日本医事新報 4872: 51-51, 2017.
- 14) Naoki Nakagawa, Tadashi Sofue, Eiichi Kanda, et al. J-CKD-DB: a nationwide multicentre electronic health record-based chronic kidney disease database in Japan, Scientific Reports, 10: 7351, 2020.
- 15) Tadashi Sofue, Naoki Nakagawa, Eiichi Kanda, et al. Prevalence of anemia in patients with chronic kidney disease in Japan: A nationwide, crosssectional cohort study using data from the Japan Chronic Kidney Disease Database (JCKD-DB), PLOS ONE, 15(7): e0236132, 2020.
- 16) 永井良三. 特別提言 - AI 時代の“臨床医学のまなざし”. 医学のあゆみ 274(9): 703-711, 2020.