

公募シンポジウム

シンポジウム3

臨床研究中核病院ネットワークにおける基盤構築と人材育成の進捗と発展

2021年11月19日(金) 14:10 ～ 15:40 F会場 (2号館2階224)

[2-F-2-01] 臨中ネットにおけるリアルワールドデータ利活用のための基盤構築の取組み

*山下 貴範¹、山下 暁士²、竹村 亮³、永島 里美⁴、諸橋 朱美²、中村 直毅⁵、青江 伯規⁶、郷原 英夫⁶、大江 和彦⁴
(1. 九州大学、2. 名古屋大学、3. 慶應義塾大学、4. 東京大学、5. 東北大学、6. 岡山大学)

*Takanori Yamashita¹, Satoshi Yamashita², Ryo Takemura³, Satomi Nagashima⁴, Akemi Morohashi², Naoki Nakamura⁵, Michinori Aoe⁶, Hideo Gobara⁶, Kazuhiko Ohe⁴ (1. Kyushu University, 2. Nagoya University, 3. Keio University, 4. The University of Tokyo, 5. Tohoku University, 6. Okayama University)

キーワード：Real-World Data (RWD), SS-MIX2, Data quality management, Data validation

日々の診療で電子カルテに蓄積されるデータはリアルワールドデータ（以下、RWD）と呼ばれ、世界的に利活用が進められており、国家レベルでその技術獲得が推進されている。RWDは診療の重要情報を保持しているものの、生データの状態でデータ解析は困難であり、RWD活用の成否はデータの品質が大きく影響する。さらに複数施設による統合解析のためには、データの品質課題を検証し、解決する必要がある。

臨中ネットでは、様々な臨床研究課題に対応できるデータ品質の高いRWD基盤整備のためにデータ蓄積からデータセット生成、その利活用を範囲として「データ項目の標準コード対応、データベースが保持する値の正確性、意味的な情報の保存性の検証、共通データベースの構築」の項目について各SWGで活動を進めている。データの正確性と保存性の検証の取り組みとしては、電子カルテ～SS-MIX2（またはDWH）間の品質管理のためにチェックリストとデータバリデーションツールを開発し、手順を整理した。チェックリストは、電子カルテデータとSS-MIX2データの整合性確認のために、各データ項目の一致件数や一致率を記録できる。ツールは、チェックリスト対象のデータを突合検証するためであり、その結果をチェックリストに反映できる。そして、品質保証確保のために品質管理作業の実施手順を定めたバリデーション手順書を作成した。令和2年度には、これらを用いて3施設でのデータ品質管理の作業を実施した。抽出された課題の対応とその評価のために2クール実施し、各施設とも第2クールで改善が見られた。一方、一定のデータ評価後の品質のレベル設定は現時点では定まっていない状況であり、品質管理指標や品質管理モデルについて議論している。シンポジウムでは、臨中ネットRWD基盤整備の状況を俯瞰的に報告し、臨中ネットデータ品質管理の内容について共有する。

臨中ネットにおけるリアルワールドデータ利活用のための基盤構築の取組み

山下貴範^{*1}、山下暁士^{*2}、竹村 亮^{*3}、永島里美^{*4}、諸橋朱美^{*2}、中村直毅^{*5}、青江伯規^{*6}、郷原英夫^{*6}、大江和彦^{*4}

^{*1} 九州大学、^{*2} 名古屋大学、^{*3} 慶應義塾大学、^{*4} 東京大学、^{*5} 東北大学、^{*6} 岡山大学

Construction of a Platform for utilizing Real-World Data on Core Hospitals for Clinical Research

Takanori Yamashita^{*1}, Satoshi Yamashita^{*2}, Ryo Takemura^{*3}, Satomi Nagashima^{*4}, Akemi Morohashi^{*2},

Naoki Nakamura^{*5}, Michinori Aoe^{*6}, Hideo Gobara^{*6}, Kazuhiko Ohe^{*4}

^{*1} Kyushu University, ^{*2} Nagoya University, ^{*3} Keio University, ^{*4} The University of Tokyo,

^{*5} Tohoku University, ^{*6} Okayama University

The data accumulated in electronic medical records in daily medical care is called Real-World Data (RWD), and its utilization is being promoted worldwide. Although RWD holds important information on medical care, it is difficult to analyze data as the raw data. RWD utilization is greatly affected by the quality of the data. Furthermore, for integrated analysis by multiple facilities, it is necessary to verify and solve data quality. Each SWG in on Core Hospitals for Clinical Research is working from data accumulation to data set generation for development of a RWD platform with high data quality that can respond to various clinical studies. A checklist and data validation tool were developed and the procedure was organized with the aim of improving the data quality of the SS-MIX2 data at each institution to a level that can be used for observational studies to verify the accuracy and completeness of the data. Data validation for quality management was carried out at three institutions. They worked twice to solve extracted problems and to evaluate them, and improvement was seen in each institution the second time. On the other hand, the quality level setting after data quality management has not been decided at this time. We are discussing quality management indicators and quality management models. We will comprehensively introduce the status of the development of the RWD platform and share the SWG activities that are proceeding with the establishment of the method theory of data quality management and the clarification of problems.

Keywords: Real-World Data (RWD), SS-MIX2, Data quality management, Data validation

1. はじめに

日々の診療で電子カルテに蓄積されるデータはリアルワールドデータ(以下、RWD)と呼ばれ、世界的に利活用が進められており^{1), 2)}、中央行政もその技術獲得を推進・支援している。RWDは診療の重要情報を保持しているものの、生データの状態でデータの解析は困難であり、RWD活用の成否はデータの品質が大きく影響する。さらに複数施設による統合解析のためには、データの品質課題を検証し、解決する必要がある。

臨中ネットでは、様々な臨床研究課題に対応できるデータ品質の高い RWD 基盤整備のためにデータ蓄積からデータセット生成、その利活用を範囲として「データ項目の標準コード対応、データベースが保持する値の正確性、意味的な情報の保存性の検証、共通データベースの構築」などの項目について各 SWG で活動を進めている。

特にデータの正確性と保存性の検証の取り組みとしては、各施設に設置している SS-MIX2 標準化ストレージのデータ品質を観察研究に用いることが可能なレベルにまで向上することを目的として、電子カルテ～SS-MIX2(または DWH)間の品質管理のためにチェックリストとデータバリデーションツールを開発し、手順を整理した。チェックリストは、電子カルテデータと SS-MIX2 データの整合性確認のために、各データ項目の一致件数や一致率を記録できる。ツールは、チェックリスト対象のデータを突合検証するためであり、その結果をチェックリストに反映できる。さらに、品質保証確保のために品質管理作業の実施手順を定めたバリデーション手順書を作成した。令和2年度には、これらを用いて3施設でのデータ品質管理

の作業を実施した。抽出された課題の対応とその評価のために2クール実施し、各施設とも第2クールで改善が見られた。一方、一定のデータ評価後の品質のレベル設定は現時点では定まっていない状況であり、品質管理指標や品質管理モデルについて議論している。

本発表では、臨中ネット RWD 基盤整備の状況を俯瞰的に触れながら、データ品質管理の方法論の確立と課題の明確化を進めている SWG 活動の内容について共有する。

2. データバリデーションチェックリスト

SS-MIX2 標準化ストレージのデータ品質管理をするにあたり、各施設のデータ品質状況を一定程度等しく評価する必要がある。各施設自身で評価を同じ対象・同じ評価項目で行うためにバリデーションの対象範囲・項目を記載した施設共通のチェックリストが必要である。人的資源が限られた臨中ネットの体制で行うためにはスモールスタートで始める必要があることから、その対象となるデータ項目の選定から開始した。

SS-MIX2 標準化ストレージ仕様書(ver.1.2f)から臨床研究のために必要となる11データ種別(病名オーダ、処方オーダ、注射実施、検体検査結果、患者基本の更新情報、入院実施、退院実施、外来受付、アレルギー、転棟転科、給食オーダ)を選定した。その中から臨床研究に必要と考えられるデータ項目を抽出し、「HISの情報とそこから出力されたSS-MIX2のデータが同等であること」を示すために必要なチェック項目を検討した。その際、以下の2点を考慮した。

1. 複数施設での観察研究を想定し、各施設に設置され

た SS-MIX2 標準化ストレージに格納されているデータの品質管理活動を一定のレベル以上かつバリデーション定義に沿って行えるように整備すること

2. 観察研究を行う研究者(利活用者)が統合解析用データセットのデータ信頼性を確認する際、抽出元となる SS-MIX2 標準化ストレージ内の RWD データの信頼性を評価し各研究課題に応じてデータ利活用の可否を判断できる評価材料となり得ること

チェックリストの確認項目は、各データ種別の検証条件における HIS 及び SS-MIX2 標準化ストレージのデータ件数(以下、データ件数)、及び各データの意味合い(以下、データ内容)とした。また、SS-MIX2 標準化ストレージに蓄積された RWD(以下、SS-MIX2 データ)を利活用する者がそのデータの信頼性を客観的に評価できるよう、HIS データと SS-MIX2 データの一致率を記載することとした。なお、HIS データ・SS-MIX2 データの突合比較検証において不一致データが存在する場合、原因究明とともに、その経過についても追記する仕様とした。最終的に、確認事項や記載方法等に関する検討の結果をまとめてデータバリデーションチェックリスト(SS-MIX2 使用時)第1版が令和2年度に完成した(図1)。

3. データバリデーションツール

SS-MIX2 標準化ストレージは患者 ID・診療日ごとのフォルダとデータファイルの組み合わせで診療データを保存する仕組みであり、検索性に乏しく、HIS データと突合して内容の確認をするには適さない。そのため、SS-MIX2 データと HIS データの一貫性の確認を容易にすることを目的にツールを開発した(図2)。

平成30年度に MID-NET®でのデータ品質管理を参考にして、SS-MIX2 標準化ストレージサーバ内のトランザクションストレージに蓄積されたデータを Relational database(以下、RDB)に保存し、データ検索を可能とするツールを開発した。当初のデータバリデーションツールは4データ種別(処方オーダー、注射実施、検体検査結果、病名オーダー)の SS-MIX2 ヘッダー情報のみをデータベース化するツールであったが、その後、SWG での議論の進行に伴い、単にメッセージ(データ)件数だけの検証ではなく、そのメッセージ内のデータ内容の確認も必要であるとの方針となったため、4種別のデータ内容をカバーするための機能強化を令和元年度に実施した。

さらに、データバリデーションチェックリスト第1版には11データ種別のデータ内容について評価項目を対象となったため、その項目をすべて網羅するために、令和2年度に以下の内容の機能強化を立案した。

1. データ種別を11データ種別に拡張すること
2. データ種別ごとに、出力が必要なセグメントとそのメッセージ内容を個別に格納できるようにすること
3. データバリデーションチェックリストに記載されている評価項目をすべて実行できるようにすること

1については、前述の通りである。2については、各データ種別にあるセグメントの中には、データバリデーションチェックリストの評価項目1つも含まないものがあること、1つの SS-MIX2 メッセージの中に複数のセグメントが存在する(例、処方オーダー(OMP-01)の中に、患者識別情報を格納するPIDは1回しか出てこないが、薬剤自体の情報を格納するRXEは複数回出てくるなど)ため、1つの SS-MIX2 メッセージ全体を1つのテーブルに格納することは困難で、セグメントごとに別々のテーブルに格納する必要がある。3については、デー

タバリデーションチェックリストが完成するまでは、対象となるデータ項目に対して評価内容が定まっていなかったことから、データバリデーションツールは SS-MIX2 データをデータベースに格納する、までの機能しか有さず、実際の比較は評価者が SQL などを作成して実施していた。データバリデーションチェックリストが完成し、個々のデータ項目について評価内容が明確となったため、SS-MIX2 データと HIS データをチェックリストに従って比較する機能を付加した。

さらに今回の機能強化では、データ項目の拡張だけでなく、HIS データ抽出や SS-MIX2 データとの突合比較機能などを搭載した。SS-MIX2 標準化ストレージサーバへ格納される RWD は、SS-MIX2 標準化ストレージ仕様書(Ver.1.2f)に定められた形式で作成されるため、HIS データと SS-MIX2 データを比較実施する際は、HIS データを SS-MIX2 標準化ストレージサーバ仕様書と同様に抽出し変換させる必要がある。このプロセスはとても煩雑なものであり、SS-MIX2 データを再現することは非常に困難を極めた。各データ種別の評価のためのシステムが完成するごとに納入と評価を繰り返し、その結果をフィードバックすることで、開発を加速する方法を採用し、令和3年2月末に名古屋大学に納品された。

バリデーションツールは Microsoft Access を用いた。初期のバリデーションツールのデータベースは MySQL であったが、病院側の担当者が上手に利用できず、バリデーションツールを利用するためだけに相当量の時間を要していた。今後、臨中ネットに参加する施設に展開する際、各施設にデータベース技術者がいない可能性も考え、比較的初心者でも理解しやすい Access を採用した。しかし、データベースのサイズが2GBを超えると動作しなくなる問題が発生している。これに対しては、トランザクションログから取得するデータを分割して格納することで対処した。

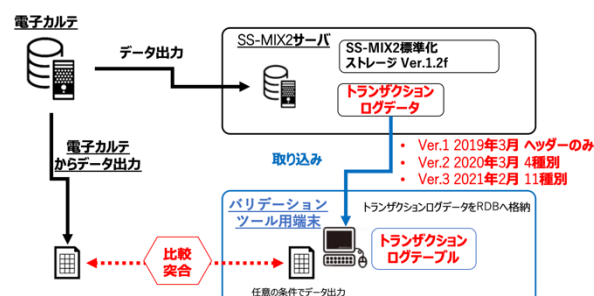


図2 データバリデーションツールの概念図

4. データバリデーション手順書

データ品質管理は定められた手順に従って実施し、必要なエビデンスを残し、後で適切に実施されたことを確認・再現できることでその品質が保証される。そのため、データ品質管理の手順を定めそれを文書化すること、どのように HIS データ・SS-MIX2 データの突合比較検証(バリデーション)実施の記録を残すかを定めることが必要となる。バリデーションツールを利用してどのように作業を行うかについて必要な作業手順を文書化し、「臨中ネットデータバリデーション手順書(SS-MIX2 使用時)第1版」(以下、バリデーション手順書)を作成した。手順書の構成は以下の通りである。

- ・ 第1章 総則
- ・ 第2章 運用管理体制
- ・ 第3章 品質管理

- ・ 第4章 資料の保存
- ・ 第5章 実施状況の確認

5. バリデーション作業の実施

令和2年度のバリデーション実施の目的として、「データバリデーションチェックリスト第1版を利用してバリデーション作業を実施し、各施設のSS-MIX2データの品質を評価すること」とした。名古屋大学、慶應大学、岡山大学でバリデーション作業を実施し、その過程で発見された課題を記録し、ノウハウを共有することを第2の目的とした。

5.1 バリデーション実施方法

3施設とも、事前にバリデーション手順書の第1章・第8条・1項に従い「2020年度データバリデーション実施計画書」を作成し、作業を実施した。対象となるSS-MIX2標準化ストレージサーバは臨中ネット用に各施設で調達したVer. 1.2fである。対象となるデータは、名古屋大学では過去データ送信(過去の指定した期間のHISデータをまとめてSS-MIX2データに変換する方式)と、リアルタイムデータ送信(HISのデータが生成されるたびにSS-MIX2データに変換する方式)の両方のデータとした。慶應大学と岡山大学はリアルタイムデータ送信のデータのバリデーション作業を行った。

比較対象となるHISデータ・SS-MIX2データは、その比較内容によって異なり、バリデーションツールで抽出・変換された後、メッセージ作成件数、データ(セグメント件数)、メッセージの内容、セグメントデータの内容ごとに定めた条件下での一致率を算出した。その結果をデータバリデーションチェックリスト第1版に抽出条件や比較条件と共に記載した。バリデーション作業でデータの不一致が見つかった場合や、ツールの不具合や問題点が発見された場合は、ベンダーとの間で課題を共有し、解決を図った。

5.2 バリデーション実施結果

3施設とも、バリデーションツールを用いた初回の評価(第1クール)を実施し、課題を抽出した。課題を可能な限り対応し、その状態の評価を再度実施した(第2クール)。結果を図3に示す。図3に記載されている値はバリデーションチェックリストの項目のうち、一致率が指定の範囲内に入っている項目の数を表している。一致件数Aは対象となるHISオーダとSS-MIX2メッセージヘッダーのうち、突合キー(患者ID、オーダNo.等)が一致した件数を指す。一致件数Bは対象となるHISデータとSS-MIX2セグメントのうち、突合キー(患者ID、オーダNo.等)が一致した件数を指す。全11データ種別総計で、一致率が100%であった割合は、名古屋大学の過去データ送信第2クールで76%、岡山大学のリアルタイム送信第2クール目で60%、慶應大学リアルタイム送信第2クールで47%であった(図3)。これは第1クールよりもいずれも改善した。

3施設とも同じ状態のSS-MIX2サーバを利用していることもあり、データの一貫性に関する課題はほぼ同一であった。共通の課題例として、食事オーダ(OMD)出力設定マスタの不備や、名古屋大学での事前確認で発見した外来受付(ADT-12)の日付取得元情報のSS-MIX2プログラム送信プログラムの不備があった。前者については、来年度に2施設にも修正適応、後者はプログラム修正後、3施設に適応し、SS-MIX2メッセージのリカバリを実施する予定である。逆に、施設毎の差異としては、各施設でのマスタの違いや実装の違いが原因のものが大部分であった。例えば、慶應大学、岡山大学では統一診療科マスタの登録を行っていなかったことから、診療科に関して不一致が多く見られた。慶應大学の結果については、患者基本情報(ADT-00)にてデータの形式などが原因と思われるツールエラーが発生したこと、バリデーション対象期間の移動系動系のデータ種別が出力されていない事象が確認されたことが起因となり一致率が2大学と比べ低水準となった。ただし、後者の事象は第2クール後に対応済みであり、次にバリデーション作業を行った場合、一致率は改善していると想定される。

また、データ不一致における課題の分類分けの結果と比較を行なっているが、これについては一般口演「諸橋ほか、SS-MIX2標準化ストレージ内に蓄積されたRWDの品質管理方法の検討と信頼性検証」での報告を参照されたい。

5.3 バリデーション実施評価

第2クールで残存している課題のほとんどは解消可能であり、継続したバリデーション作業により臨床研究に用いるデータとなり得ると考えている。今回のバリデーション作業により、SS-MIX2標準化ストレージ送信プログラムの不備やシステムマスタが起因となる不一致、及び施設での運用がデータに与える影響を明らかにすることができた。

今後は、不一致の原因に対する対応策を講じ、再度評価を実施、残された不一致の原因を調査し、その対応策を講じる。いわゆるPDCAサイクルに従った形でSS-MIX2データの品質を向上させていく必要があると考える。その結果、品質が高レベルに達したとしても、データ品質はマスタの更新、システムの改修や微調整などで容易にデータの品質は低下していることが知られている。そのため、SS-MIX2標準化ストレージサーバ内のRWDを臨床研究に用いるためには、継続した品質管理が必要である。

6. 品質管理モデルと品質指標

データ管理の教科書ともいえるDAMA-DMBOK第2版³⁾によると、データ品質の度合いとはデータ利用者の期待と要求を示す度合い、つまりデータが果たすべき目的に合致しているかを示す度合いとされている。しかし、臨中ネットは広く観察研究に用いることのできる基盤整備を目的としており、データを蓄積し、品質管理を行う時点ではデータ利用者の要件は不明である。そのため実際にどの評価軸で評価すべきなのか検討が必要である。その中で、臨中ネットの品質管理モデルについて、「データベースの品質」と「データベース研究の品質」に切り分けて整理した。

「データベースの品質」は、データベースが原資料を反映していること、現資料を扱うシステムの正確性、データの即時性、正確性の担保することであり、原資料を適切に変換・格納していることを示すことが重要である。「データベース研究の品質」は、データベースを用いた「研究」の品質であり、研究デザインとの整合性、正確な抽出アルゴリズム、データ項目の可用性と匿名性の適切なバランスを要素として、結果の信頼性に注目し、信頼性を示す情報の提示が重要である。

また、品質管理モデルを3フェーズ/6層に分割した品質管理マトリクスモデルを提案した(図4、図5)。3フェーズは、工程管理、品質検証、品質改善を定義し、6層は下位より診療録層、構造化データ層、データベース層、匿名化層、セッション層、プレゼンテーション層を定義した。

さらに、それぞれのセルにおける定義について議論を重ね、DAMA-DMBOK第2版の評価軸を品質指標として整理し、議論を始めたところである(図6)。

