登録情報の流用性の現状
−項目標準化の課題−

横井英人[*1]、西本尚樹[*2]、谷川原綾子[*3]
*1 香川大学医学部附属病院　医療情報部、*2 香川大学医学部附属病院　臨床研究支援センター、
*3 北海道科学大学保健医療学部

Current status of Data Item Standardization
- What shall we do for Data Item Standardization?-

Hideto Yokoi[*1*2], Naoki Nishimoto[*2], Ayako Yagahara[*3]
*1 Department of Medical Informatics, Kagawa University Hospital,
*2 Clinical Research Support Center, Kagawa University Hospital,
*3 Faculty of Health Sciences, Hokkaido University of Science

Abstract: Many research and development projects supported by AMED established some data capture schemes for clinical research. A data extraction from Electronic Medical Record system (EMR) and secondary use for clinical research are emphasized because of workload decreasing and needs of source document keeping. We have a lot of terminologies, master data, and code lists, and it is not easy to unify the master data in our hospital. The realistic solution is to make a map between local master and standard master. Mapping workload is too heavy for hospital staffs because standard master will be revised annually. Recently we developed a supporting system for mapping. The system would release us from the workload. We also give attention to master contents distribution. CDISC release Controlled Terminology every 3 months, and one useful solution, SHARE API, which enables us to retrieve newest resources via web access is available. We will discuss reasonable maintenance of master data for EMR.

Keywords: terminology, mapping, master data maintenance

1. 背景

多施設の電子カルテからのデータ抽出を行うためには、標準となるデータの格納方法と、標準となる項目セット、更にはそれぞれの項目に於ける標準コードという3つの要素が必要となる。これらは密接に関連し合い、互いに他者の要素を持つこともあり得る。標準となるデータの格納方法については、SS-MIX 標準ストレージが 2017 年 3 月現在、1114 件導入されており、この格納方法に対応したデータ抽出であれば、原理的には単一のシステムで抽出が可能となる。後者については、格納されるときの項目名が標準化されていることにより、目的とする項目の特定が容易になる。更にそれぞれの項目に格納されるデータが標準的な用語集やそれに付随したコードに準拠していると、抽出したデータの集計の労力が大幅に削減される。CDISC で提供される様々な規格若しくはコンテンツを用いると、上記の3つの要素を、例えば標準となるデータの格納方法に Operational Data Model (ODM)、標準となる項目セットに Clinical Data Acquisition Standards Harmonization (CDASH)、標準用語集（コード集）として Controlled Terminology を用いることで、統一的な運用を行うことが可能となる。ただし、CDISC は臨床研究のみに限定した運用のみを目的としているので、一般的な臨床上の連携に於いて必要十分であるわけではない。

医療情報データベース基盤整備事業により構築された MID-NET は、SS-MIX により格納方法を統一し、ICD-10・HOT コード・JLAC10 などを用いることにより項目データのコード統一を果たし、平成 30 年から、薬剤の市販後調査の一部をデータベース研究として行う際、データ抽出等に用いられる予定である。また山口大学が主幹となって行っている AMED 研究「既存の診療情報と一体的に運用可能な症例登録システムの構築とアウトカム指標等の分析・利活用に関する研究」では、National Clinical Database (NCD)の症例登録のために電子カルテからデータを抽出し、できるだけ標準的な形式で送信することを目的に CDISC などの標準を積極的に使用している。

2. 当院で経験したマッピング作業の現状

異なる項目セット・用語集・コード（以後、まとめて用語集と呼ぶ）を用いたデータを合わせて集計を行う時や、一方のデータを他方のデータの用語集に合わせて変換を行う場合に、双方の用語集同士の対応表となるマッピング表を作成する必要がある。マッピング表には方向があるとされている [1]。

二つの用語集 A と B があるとすれば、A を用いたデータを B に変換する場合と、その逆を行う場合があり、それぞれにマッピング表が存在することになる。マッピングを行う際には、片側にはあってもう一方にはない概念や項目があれば、当該項目がある用語集から当該項目がない用語集にマッピングする際には、マッピング表に対応が記載されるが、その逆の方向では、対応が記載されることはない。このことは粒度が違う場合にも同様で、細かい粒度の用語を、それらを包含する用語にマッピングすることはできても、その逆はできない。

このようにデータの関連性を精査し、対応を標記するマッピング作業は、多大な負荷を施設に強いることになる。更にどちらかの用語集が更新されれば、必然的に更新部分についてのマッピングが必要になり、これを用手的に行う負荷は想像に難くない。

2.1.1 臨床研究に於ける薬剤のグルーピングの例

香川大学が受託している臨床研究で、被験者の服薬情報として、薬剤名称ではなく薬剤の種類（グループ）の入力を要求している例がある。例えば、「脂質治療薬」として「スタチン」「スタチン以外」という情報のみを収集している。国内で上市されているスタチン製剤は 6 種類であるが、ジェネリック薬は数え切れないほど存在し、日に日に増えている。スタチン以外の薬剤も同様である。

香川大学で採用している薬剤マスタは月に一回程度更新されており、持参薬処方などで全件マスタとして使用されている。同マスタには YJ コードが収載されており、当該研究に際しては、この YJコードを元にグルーピングを生成しようとした。YJ コードは、上位4桁（それより粒度が粗い分類の場合は、3桁・2桁）に注目すれば「日本標準商品分類（平成 2 年 6 月改定）」で規定されている分類情報を得ることができる。ただし、YJ コードで表される分類は、薬剤によって複数該当する可能性のあることが指摘されている。そのような状況下では、YJ コード上位4桁の情報のみでなく、該当する薬剤を個別に指定して処理することとした。また上記分類では対応できない分類がある（例えば糖尿病用剤の「選択的 DPP-4 阻害剤」・「選択的 SGLT2 阻害剤」などは日本標準商品分類の項目に存在しない）。このように日本標準商品分類で対応できない分類については、YJ コードの上位7桁により、続々と上市される後発品を含め、適切に一般名レベルで特定し、そのリストをもって、分類することとした。

2.1.2 薬剤グルーピング支援システムの開発

前項にて作成したマッピングアルゴリズムを実現するためのプログラム（図1）を作成した。今回、要求されているのは服薬情報であり、外用薬は除く、とされている。例えば「痔疾用剤」（255 で始まる YJコード）のほとんどは外用薬であるが、薬剤名を見たところ「～錠」や「カプセル」という名称が見られ、その多くが内服薬であることを確認したので、これを除外することとした。（検索：「255 で始まる YJ コードを持つレコード」から「薬品名が「錠」若しくは「カプセル」を含むレコード」を除外）このように複数の検索条件を組み合わせて特定の薬剤グルーピングを形成することがあった。今回のアルゴリズムでは検索条件として最大2つの組み合わせしか案出されていなかったので、システムとしての実装は、2つの検索条件の組み合わせを想定して行った。

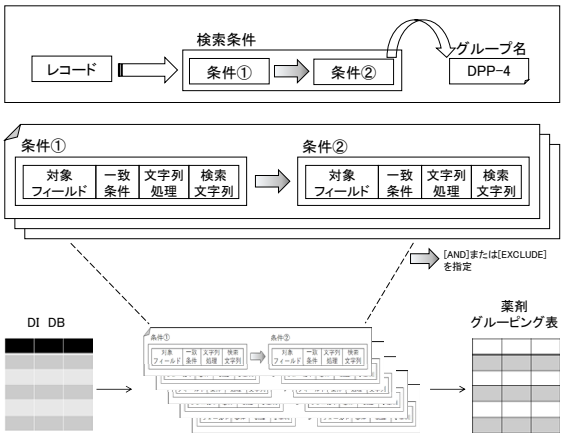

図1　薬剤グルーピング支援システム

データのグルーピングについて、2つの条件を適用して作成した検索条件をいくつも組み合わせて、行う。

システムは Excel VBA を用いて開発し、Excel ファイル中の特定のシートを対象に、2つの検索条件を適用して、グルーピングする機能を実装した。

検索条件としては、対象として当院で採用している薬剤情報(DI)用データベースの一部（一つのテーブルとして Excel ファイルの1シートに収納され、原則的に1薬剤1レコードとしたもの）「検索対象フィールド」「検索文字列」「対象フィールドへの文字列処理（左側若しくは右側から n 文字を抽出、m 文字目から n 文字を抽出）」「一致の種類（完全一致か部分一致）」を指定して一致の有無を調べることが可能となっている。また2つの検索条件を用いる場合には、一つめと二つめの関係性（一つめと二つめとも一致（AND））か（一つめに一致するが、二つめには一致しない（EXCLUDE））の二つから選ぶことができる。なお「検索文字列」は「^」で区切ることにより、一度に複数の文字列との一致を調べることができる。これらの機能で、桁毎に表現する情報が決まっている YJ コードの解釈を自動的に行うことが可能となった。

本システムを用いて薬剤のグルーピングを行った。グルーピングのアルゴリズムの確立とデータの検証には延べ数日間以上かかったが、そのアルゴリズムの実装は約1日、その後1月毎のマスタ更新に際するマッピング表の更新は1時間程度で終了した。別に作成した EDC 入力支援システムで、このマッピング表などを元に、院内の SS-MIX の処方結果や検査結果を臨床研究用の EDC に提供する情報に変換する（図2）。現在、このシステムの有用性を検証中である。

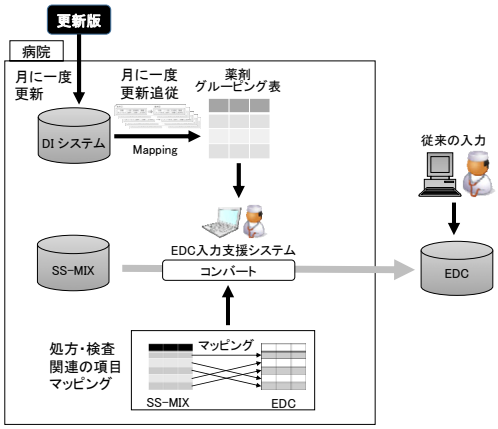

図2　SS-MIX から EDC へのデータ変換・送出

データ項目のマッピング及びデータ自体のマッピングが行われている。薬剤データ自体のグルーピングは DI 用 DB の定期的更新に合わせて、マッピングが必要になる。

2.2.1 自然言語に関するマッピングの例

筆者らは医療機器不具合用語集に関するメンテナンスの支援の依頼を受けて実施している [2]。同用語集は、医療機器の不具合事象を表現する「不具合用語」、人体への健康被害を表現する「健康被害用語」、医療機器を構成する「部品・構成品用語」から成り、用語集としては 89、合計 6000 を超える用語が収載されている。

13 医療機器団体がそれぞれ専門分野に於いて 89 の医療機器グループを想定した用語集を構築した。それぞれの用語集は独立して運用することを前提にしていたが、互いの用語集に於いて、同様の事象を扱った用語が散見され、用語集間の整合性を担保することが求められている。医療機器不具合用語集の精緻化に向け、定義文の表記から同義・異義語の自動判別手法の検討を行ってきた。用語自体、若しくは

定義文の違いについて、軽微な違いの検出には編集距離が有効であった反面、言い回しが違うものの実際には述べている内容に変わりがないケースを検出することは困難であった。現在、定義文を形態素解析し、その結果の共起性を検証した結果、同じ内容を似たような言い回しで何通りにも表現されているケースがあることが見いだされた。そのような表現を集約化したテンプレートを用いることにより、表現方法が統一化できるきっかけとなった。用語自体の表現の他、定義文の内容などについては、自然言語処理の諸手法により、マッピングの人的な作業を一定量減らせる可能性がある。

2.2.2 検査項目のマッピング

CDISC SDTM をはじめとしたコンテンツは基本的に英語で提供される。SDTM の LB ドメインに定義された変数に各病院の検査項目名をマッピングするには、英語と日本語の翻訳をした上で、それぞれが表現することを意図した概念が一致しているかを検証しなくてはならない。そこには様々な知識が必要で、そのナレッジベースの作成が必要となる。JLAC10 のマッピングも同コードに精通した臨床検査技師が必要で、各病院の負担は相当に大きい。AMED 研究「医薬品等の安全対策のための医療情報データベースの利用拡大に向けた基盤整備に関する研究」を実施中の康班では、各病院のローカルコードとJLAC10のマッピングを行うガバナンスセンターの運用について検討しており、そのような業務の集約化は一つの現実的ソリューションであると考える。

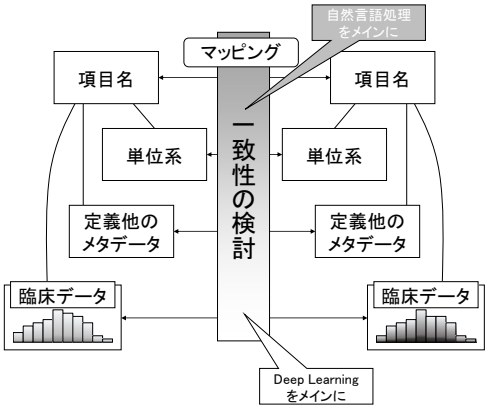

図3 マッピングのための知識表現

項目名同士のマッピングなど文字レベルの作業は、自然言語処理を基本とした技術を使用。また大量データに対しては Deep Learning をはじめとした機械学習により特徴量算出を想定。

その他、我々は、マッピングにオントロジー（意味関連に基づいた構造を持った用語集）表現技術を用いて、Web 上の Machine Readable な情報を利用する手法や、Deep Learning を用いて、その類似性を機械学習させることも今後の研究対象として検討している。前者は定型的な情報表現に基づき、既にある「知識」を利用し、自動的にマッピングできる項目の数を増やすことを目的とする。CDISC は現在 Resource Description Framework (RDF)[3]を用いたコンテンツ表現を提供している。RDF は Web 上の資産を意味論的に使用するための表現手法であり、より機械によるコンテンツ処理範囲の向上が期待できる。CDISC では平行して、SHARE API[4]と呼ばれる Web 上の API を介した情報アクセスサービスも提供しており、こちらも機械的な処理への親和性に富んでいる。

後者の Deep Learning は、言語化しにくい大量データの特徴を機械学習によって特徴量として表し、それを項目同士の異同判別に用いることが目的となる。いずれも一朝一夕に成るものではないが、様々な努力により、機械による知識表現範囲が拡大し、それにより人的作業負担（人間が判断する量）を減らすことを考える必要がある。マッピングの作業はいずれ我々の日常業務に於いて、負担として重くのしかかってくることになるからである。

参考文献

1) ISO/TR12300 Health informatics – Principles of mapping between terminological systems
2) 谷川原綾子，西本尚樹，横井英人，他：編集距離を用いた同義語同定手法の検討．第 21 回日本医療情報学会春季学術大会抄録集.2017.
3) CDISC Standards in RDF Reference Guide Version 1.0 Final
4) https://www.cdisc.org/standards/share