様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３１　日現在

機関番号：　１２６０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１５２１８

研究課題名（和文）標準データモデルの導入による臨床研究データマネジメント基盤技術の開発

研究課題名（英文）Fundamental software library for clinical data management with standard-based data models.

研究代表者

岡田　昌史（Okada, Masafumi）

東京大学・医学部附属病院・特任講師

研究者番号：７０３７５４９２

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：臨床研究で得られるデータの品質向上を目的として実施されるデータマネジメント活動の標準化および効率化を可能にするためのツール、Define2Validateを開発した。医学研究における電子データの国際標準規格であるCDISC標準にのっとったデータの品質向上活動を行う際に、本研究成果を用いることで、標準規格形式で記述されたデータの検査ルールをR言語を用いた再利用可能なプログラムに変換し、自動的にデータ検査を行うことが可能となる。

研究成果の概要（英文）：We developed a software tool to help data management process for clinical research: Define2Validate. When the research will be conducted using CDISC standards, one can apply validation rule for collected data with standard-based rule description, by converting the rules into re-usable script of R statistical language.

研究分野：　データマネジメント，疫学，医療情報学

キーワード：　データマネジメント　臨床研究　CDISC　R言語

１．研究開始当初の背景

実際の臨床現場で日々発生するデータを用いた医学研究、すなわち臨床研究は、近年の電子カルテの急速な普及にともない、診療に関連する情報が何らかの形で電子データ化されるようになったことを通じて、効率的に実施できる状況が整いつつある。

その一方で、電子カルテを用いた臨床研究の品質は従来に比べて上がったか？という問題には、明確な回答は示されていない。電子カルテから研究向けに出力されるデータは、多くの場合ほぼ無構造なカンマ区切りテキストファイルの形式で交換されていることが、研究代表者が 2007 年に実施した調査結果でも示されている [1]。一方、電子カルテ上のデータは診療の実務のために収集されたものであり、データは単位の混在、測定日などの重要な情報の欠落などを多く含む、いわば「汚い」データである。それを科学的な研究に活かすためには、データを測定方法、妥当な値範囲等の情報を含む標準的データモデルに適合させて品質を検査・向上させるデータマネジメントの過程が必須である。

データマネジメントの品質は臨床研究の品質に直結する要素だが、その具体的手法や使用されるツールは研究実施機関によってばらばらであり、最適な手法を議論するための基盤となる規格やツールも存在しない。

そこで、本研究では医療情報分野で標準化が進められているデータモデルと、広く普及しているオープンソースの統計解析ソフトウェア「R」[2]とのインターフェースを開発することにより、データマネジメントの作業工程を再利用可能なプログラムパーツとして、共通のプラットホーム上で評価するための基盤技術を提供する。開発成果をオープンソースライセンスにより公開し、製薬企業や医薬品開発受託機関におけるデータマネジメント業務の標準化に寄与することを目的とした。

[1] 岡田昌史. 診療情報からの地域共通臨床研究データベース構築のための基礎調査. 平成 18-19 年度科学研究費補助金研究成果報告書. 2008
[2] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL https://www.R-project.org/.

２．研究の目的

1) 医療情報分野におけるデータモデルのうち、臨床研究に適したモデルを調査する。
2) 統計解析ソフトウェア R と上記のデータモデルとのインターフェースを開発する。
3) 開発したインターフェースの性能を評価する。

３．研究の方法

まず、医療情報データモデルの現状を整理し、本研究での統計解析ソフトウェア R とのインターフェース開発対象となるデータモデルを選定した。候補としては、openEHR プロジェクトが採用しており、ISO13606 として標準化されている Archetypes データモデル [3]、および、国際的な臨床研究データ交換基準を提供する CDISC 標準 [4] を検討対象とした。なお、電子カルテシステムにおいては、HL7 標準規格がすでに広く実用されているが、HL7 は原則として通信のための標準規格であり、データ項目の意味について標準化するものではないので、検討対象とはしなかった。

R 内でのデータ表現形式との整合性、データマネジメントの観点からの再利用性、実際の応用事例の多寡などの点から両者を比較し、適したモデル 1 種を選定した。

その後、R から該当データモデルにアクセスして、データマネジメントを支援する機能を備えた R プログラム（以下「支援プログラム」と呼ぶ）を開発した。

支援プログラムは、図 1 のように、臨床研究における電子症例報告書自体が内部的にデータモデルを使用することで、個々のデータの単純な記載ミス等を未然に防ぐために利用されることを想定して設計した。したがって、以下のような機能を備えるものとした。

1) 研究プロトコル上のデータ収集時期、欠測値の許容範囲等の制限を、電子症例報告書およびデータベースの実装と独立して R コードで記述することができる
2) データの分布の出力、時系列的に異常な変化の検出等のデータマネジメント実務のアルゴリズムを、臨床研究の内容とは独立して R コードで記述することができる
3) 実行時の性能が実用的である

開発完了した支援プログラムはソースコードをオープンソースライセンスで一般公開した。

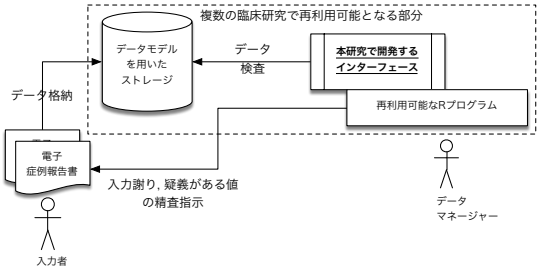


図 1: 研究成果の利用イメージ

[3] openEHR - An open domain-driven platform for developing flexible e-health systems. https://www.openehr.org/
[4] CDISC - Clinical Data Interchange Standards Consortium. https://www.cdisc.org/

４．研究成果

4.1 対象とした医療情報データモデル

本研究では、openEHR Archetypes データモデルおよびCDISC CDASH のデータモデルを対象として、R 内でのデータ表現形式との整合性、データマネジメントの観点からの再利用性、実際の応用事例の多寡等の観点から調査した。

openEHR Archetypes はオブジェクト指向データモデルであり、データ項目それぞれに応じてそれぞれ別の属性情報が設定されている。R から取り扱う場合、年齢、性別、血圧などの個々のデータ項目それぞれに対して R のリスト、もしくはデータフレームを対応させることが妥当と考えられたが、それぞれの臨床研究で用いられる項目全てを表現するには多数のリストやデータフレームが必要となってしまうので、R からの操作の効率は必ずしもよくはないと考えられた。再利用性については、きめ細かい標準化がなされており、精密なデータ検査が可能となることが考えられたが、必須とされる属性値が少ないため、データマネジメントに用いられるコンピュータープログラムの再利用性は低いと考えられた。ただし、このことはopenEHR が本来電子カルテのためのデータモデルであり、できるだけ多様な臨床データに対応するように開発されているためであるから、openEHR 自体の問題点ではないことを注記しておく。応用事例については、openEHR のArchetypeを取り扱うライブラリはJava 言語で実装されたものが複数存在するものの、R 言語ではまだ公開されたものは存在しないと思われた。

CDISC のデータモデルは古典的な２次元の表であり、データ項目それぞれは１つの列で表される。R の基本データ型であるデータフレームとこの表を対応させることで、自然に取り扱うことが可能であった。再利用性については、CDISC では必須のデータ項目も比較的多く、また個々の臨床研究プロトコルごとに利用されるデータ項目が変更されていても、プロトコルの情報を反映したメタデータ定義ファイル(Define-XML)を読み取ることで、データマネジメントプログラムの動作を変更することができる。そのため、データマネジメントに用いるプログラムの再利用性は高くなることが期待できた。また、応用事例については、研究開始時点でもR の拡張パッケージとしてCDISC データを取り扱うライブラリが存在した。

以上のことから、臨床研究のデータマネジメントの標準化とソフトウェア基盤の強化のためには、R と CDISC 標準を組み合わせた場合に利用できるソフトウェアを開発することが必要と考えた。

4.2 データマネジメント支援プログラムの開発

本研究で開発するデータマネジメント支援プログラムは、R を利用して CDISC 標準に則ったデータセットの内容の検査、記述統計、異常検知等のデータマネジメント業務を実施することを支援するためのコンピュータープログラムとして定義した。CDISC 標準は、臨床研究のプロトコルの電子的記述、症例報告書の内容の記述、データセットの書式、解析結果の記述、データのコーディングに利用される統制用語辞書などを包括的に定めている標準だが、支援プログラムの対象業務はデータマネジメントなので、CDISC 標準のうちデータセットを表現する Dataset-XML、およびそのデータセットのメタデータを表現するDefine-XML を対象とした。

Define-XML はデータセットが準拠するべきルール、すなわちデータ項目ごとに、そのデータ型、必須か否か、カテゴリー値であれば利用する統制用語番号といったメタデータを XML 形式で表現するための規格であり、CDISC 標準の一つである。CDISC 標準は医薬品の承認申請時の電子的データ提出形式として採用されていることでも知られているが、その際にも Define-XML の添付が必要となっている。Define-XML はデータが従うべきルールの記述であるから、臨床研究のデータ収集後のデータマネジメント過程で行われるデータ検査のルールについても、Define-XML で記述することが可能である。

しかし、実際に収集されたデータセットが、Define-XML で定められたルールと矛盾していないかどうかを実際に調査、検出するためのソフトウェアはCDISC からは提供されておらず、また研究グループで調査した限りでは、利用できるものは存在していなかった。そこで、本研究では Define-XML で記述されたルールと Dataset-XML で記述された収集データの間の整合性を検査するためのソフトウェアをR 言語で実装した。

4.3 支援プログラムの詳細

支援プログラムはR 言語のスクリプトとして開発し、Define2Validate という名称をつけた。CDISC Dataset-XML、およびDefine-XML を読み取って、含まれる情報をR のデータフレームとして格納する機能には、R の拡張パッケージとしてオープンソースライセンスで公開されているR4DSXML[5]を利用した。また、R 言語で記述されたデータ制約ルールに対してデータの整合性を検査する機能には、やはりR の拡張パッケージとしてオープンソースライセンスで公開されている validate[6]を利用した。今回開発したスクリプトは、R4DSXMLにより Define-XML に記載されているデータ制約ルールを読み取った上で、それを解釈して適切なR 言語の式の形に変換する。この式

を validate に渡し、整合性の検査を実施するものである。Define-XML の一部の例を図 2，変換された R 言語式の例を図 3 に示す。

```
<!-- ORIGINAL Define-XML -->
<ItemDef OID="IT.LB.LBSEQ" Name="LBSEQ" DataType="integer" Length="2"
    SASFieldName="LBSEQ">
    <Description>
        <TranslatedText xml:lang="en">Sequence Number</TranslatedText>
    </Description>
    <def:Origin Type="Derived"/>
</ItemDef>
```

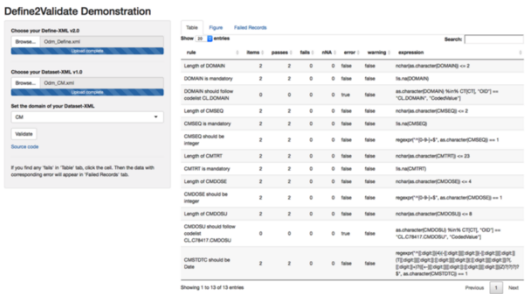
図 2: Define-XML の例

```
# Generated YAML Rule
-
  expr: 'nchar(as.character(LBSEQ)) <= 2'
  name: Length of LBSEQ
-
  expr: '!is.na(LBSEQ)'
  name: LBSEQ is mandatory
-
  expr: 'regexpr("^[0-9-]+$",as.character(LBSEQ)) == 1'
  name: LBSEQ should be integer
```

図 3: 変換された R 式の例

Define2Validate で特徴的な機能は、Define-XML で定義される Value-level Metadata と呼ばれる制約のサポートである。CDISC は全てのデータが 2 次元の表で表現されるが、臨床検査値など、1 つの列に別の意味を持つ値を混在して格納せざるを得ない場合がある。Define-XML では、このような場合に Value-level Metadata を導入し、ある列の値に応じて、別の列に課される制約の内容が変化するというルールを記述することができる。これにより柔軟かつきめ細かいデータ制約ルールを定義することができるが、コンピュータープログラムでこの制約を正しく取り扱おうとすると複雑なプログラミングが必要となる。Define2Validate を使うことで、Value-level Metadata の定義が自動的に複数の条件式の組み合わせに展開され、複雑なルールでも機械的なデータ検査が可能となっている。

Define2Validate は基本的には R のコマンドプロンプトから、関数呼び出しの形で利用することを想定している。しかし、機能をわかりやすく表現するため、Web アプリケーションも作成した。画面写真を図 4 に示す。こちらのソースコードも同時に公開している。



[5] R4DSXML: R package for handling CDISC Dataset-XML and Define-XML. https://github.com/DataDrivenInc/R4DSXML

[6] validate: Data Validation Infrastructure. https://cran.r-project.org/web/packages/validate/index.html

4.4 支援プログラムの機能評価

本研究で開発した支援プログラムの機能を、以下の点から評価した。

1) 研究プロトコル上のデータ収集時期、欠測値の許容範囲等の制限を、電子症例報告書およびデータベースの実装と独立して R コードで記述することができるか

Define2Validate は、入力を CDISC で定められている Dataset-XML として受け取り、データの許容範囲を Define-XML として受け取る。電子症例報告書の実装とは関係なく利用することができ、Define-XML で表現できないルールについては直接 R 言語の式として検査ルールを追加することができる。

2) データの分布の出力、時系列的に異常な変化の検出等のデータマネジメント実務のアルゴリズムを、臨床研究の内容とは独立して R コードで記述することができる

既存の R4DSXML パッケージの機能として、R のデータフレームにデータを読み取ることができ、CDISC に準拠したデータセットであれば表の名称、列の名称、型などは定まっているので、研究プロトコルの内容に関わらずアルゴリズムを R コードとして記述し、再利用することが可能である。

3) 実行時の性能が実用的である

実用性の目安として、Apple Macbook Pro (CPU 2.9GHz Intel Core i7, 16GB メモリ, macOS 10.13.4) 上で、7661 行の Define-XML(CDISC による Define-XML 2.0 規格の配布パッケージに含まれている SDTM 向けの Define-XML の例)から、LB ドメインのルールを生成させた場合の実行時間を測定した。3 回実行させ、平均 1.137 秒であった。

また、同配布パッケージに含まれる SDTM LB ドメインのサンプルデータ(83 行)を、生成されたルールを用いて検査した場合の所要時間は、平均 0.073 秒であった。

以上のことから、Define2Validate を用いたデータ検査は、通常の臨床研究の規模のデータに対しては実用的であり、データが入力されると直ちにデータ検査を行う、という利用法にも耐えうるものであると考えられる。

なお、当初の研究計画では、実際に実施されている臨床研究のデータマネジメント業務に対して本研究で開発された支援プログラムを導入するフィージビリティスタディを実施する計画であったが、日本ではアカデミアにおける CDISC 標準の導入への取り組みが未だ十分でなく、CDISC の導入事例自体がまだ少ないこと、また、CDISC が導入されていても、Define-XML を理解し記述できる人材

がアカデミアのデータセンターにはほぼいないことから、研究期間中にはスタディを実施するまでには至らなかった。しかしながら、臨床研究に対するCDISC標準の導入はメリットが大きく、すでに製薬工業における CDISC 標準の導入がほぼ完了しているため、今後本研究を実際に臨床研究の現場に導入する試みを実施することは可能と考えている。

５．主な発表論文等

〔雑誌論文〕（計　0件）

〔学会発表〕（計　2件）
Masafumi Okada. Define2Validate - Validate CDISC Dataset-XML with corresponding Define-XML metadata. CDISC Europe Interchange 2017(London). 2017.
Masafumi Okada. Define2Validate - Validate Your CDISC Dataset with Metadata Defined in Define-XML. CDISC Japan Interchange 2017(Tokyo). 2017.

〔図書〕（計　0件）

〔産業財産権〕0件
〔その他〕
ソースコード・関連資料公開URL:
https://github.com/mokjpn/Define2Validate

６．研究組織
(1)研究代表者
岡田　昌史（OKADA, Masafumi）
東京大学・医学部付属病院・特任講師
研究者番号：70375492

(2)研究分担者
土井　麻理子（DOI, Mariko）
和歌山県立医科大学・医学部・講師
研究者番号：70636860

(3) 研究分担者
上野　悟（UENO, Satoshi）
国立研究開発法人国立精神・神経医療研究センター・臨床研究支援部・科研費研究員
研究者番号：20595706

(4) 研究分担者
木内　貴弘（KIUCHI, Takahiro）
東京大学・医学部附属病院・教授
研究者番号：10260481