

---

一般口演

## 一般口演24

## DWH・CDISC・その他

2017年11月22日(水) 16:00 ～ 17:30 F会場 (10F 会議室1004-1005)

---

### [3-F-3-OP24-2] 正規化処理を利用した臨床検査値単位と CDISC SDTMのマッピング

西本 尚樹<sup>1</sup>, 田村 菜穂美<sup>2</sup>, 國方 淳<sup>1</sup>, 谷川原 綾子<sup>5</sup>, 赤堀 澄子<sup>3</sup>, 十河 智昭<sup>3</sup>, 谷 佑馬<sup>4</sup>, 横井 英人<sup>4</sup> (1.香川大学医学部 附属病院 臨床研究支援センター, 2.北海道大学 環境健康科学研究教育センター, 3.香川大学医学部附属病院 医療情報部, 4.香川大学 大学院医学研究科 医療情報学分野, 5.北海道科学大学 保健医療学部 診療放射線学科)

[背景及び目的]我々は、これまで、電子カルテデータを臨床研究において二次利用を促進するために、臨床検査値のデータを CDISC SDTM (Study Data Tabulation Model) に対応させる標準的な手法を探索してきた。SDTMの LBドメインに定義された45変数のうち、何らかの変換処理が必要な変数は12変数(27%)であった。そこで、本研究では臨床検査データを SDTMに対応付け (マッピング) する際の一致性及び問題点を明らかにするために、文字列の正規化処理を用いてマッピング手法を探索した。

[方法]香川大学医学部附属病院 (以下、病院) の臨床検査マスターを Unicodeで定義されている互換合成を用いて正規化したのち、統計解析パッケージ JMPを用いて SDTM Terminologyに対応した単位名を dice係数による一致度で評価した。フィージビリティ研究として後ろ向きに330例のデータ収集を計画し、収集可能であった225例 (42,544レコード) を正規化とマッピング処理の調査に用いた。

#### [結果及び考察]

SDTM Terminologyより、Codelist Name=Unitで絞り込みを行ったところ、CDISC Submission Valueの変数から441種類の単位名を取得した。同様に病院臨床検査マスターから166種類の単位名が得られた。SDTM Terminologyと病院臨床検査マスターの単位名をマッチングさせた dice係数は、0.053であった。互換合成を施した後、単位名に関して SDTM Terminologyと病院臨床検査マスターとの一致を評価したところ、dice係数は0.075であった。Dice係数を低下させる要因に、「μ」の代わりに「u」を代替的に用いるなど、SDTM Terminology独自の仕様が見られた。臨床検査値データについて評価指標を算出中である。

# 正規化処理を利用した臨床検査値単位と CDISC SDTM のマッピング

西本 尚樹<sup>\*1</sup>, 田村菜穂美<sup>\*2</sup>, 國方 淳<sup>\*1</sup>, 谷川原綾子<sup>\*3</sup>, 赤堀 澄子<sup>\*4</sup>, 十河 智昭<sup>\*4</sup>,  
谷 佑馬<sup>\*5</sup>, 横井 英人<sup>\*4</sup>,

<sup>\*1</sup> 香川大学 医学部附属病院 臨床研究支援センター, <sup>\*2</sup> 北海道大学 環境健康科学研究教育センター, <sup>\*3</sup> 北海道科学大学 保健医療学部 診療放射線学科, <sup>\*4</sup> 香川大学 医学部附属病院 医療情報部, <sup>\*5</sup> 香川大学 大学院医学研究科 医療情報学分野

## Mapping between laboratory test Unit and CDISC SDTM using text normalization

Naoki Nishimoto<sup>\*1</sup>, Naomi Tamura<sup>\*2</sup>, Jun Kunikata<sup>\*1</sup>, Ayako Yagahara<sup>\*3</sup>, Sumiko Akahori<sup>\*4</sup>, Tomoaki Sogo<sup>\*4</sup>, Yuma Tani<sup>\*5</sup>, Hideto Yokoi<sup>\*4</sup>,

<sup>\*1</sup> Clinical Research Support Center, Kagawa University Hospital

<sup>\*2</sup> Center for Environmental and Health Sciences, Hokkaido University

<sup>\*3</sup> Department of Radiological Technology, Faculty of Health Sciences, Hokkaido University of Science

<sup>\*4</sup> Department of Medical Informatics, Kagawa University Hospital

<sup>\*5</sup> Department of Medical Informatics, Graduate School of Medicine, Kagawa University

[Abstract] The aim of this study was to identify the \*\*\* problem (障害) for mapping the laboratory test name to the CDISC SDTM (Study data tabulation model). We applied the character normalization implemented in Java SE1.8 to the laboratory name used in EMR stored in SS-MIX2 format. The mapping accuracy was estimated by dice index. We obtained 441 types of the laboratory test unit from CDISC controlled terminology, 166 types were obtained from the master file of EMR. The results showed 0.053 in dice index. Symbol was found to be difficult for handling and substituted to the another character, e.g. micro- to u. Further analysis was required for the laboratory test value.

**Keywords:** HL7, Normalization, SS-MIX2, CDISC SDTM, Mapping, LB domain,

### 1. 背景

臨床研究データの標準化によって、シグナルディテクションやメタアナリシスを効率化するために、CDISC が治験等で採用されている<sup>1,2)</sup>。我々は、これまでに臨床研究において電子カルテデータの二次利用を促進するために、臨床検査値のデータを CDISC SDTM (Study Data Tabulation Model) に対応させる標準的な手法を探索してきた<sup>3)</sup>。臨床検査値のデータは、種々の臨床データの中で、最も標準化が進んだデータと考えられるが、SDTM の LB ドメインに定義された 45 変数のうち、何らかの変換処理が必要な変数は 12 変数(27%)であった。電子カルテデータと臨床研究のデータ標準である CDISC を接続する試みはいくつか報告されているものの、標準同士を使用した情報の接続の可能性に関する調査は少ない<sup>4-7)</sup>。

### 2. 目的

そこで、本研究では臨床検査データを SDTM に対応付けする際の一致性及び問題点を明らかにするために、文字列の正規化処理を用いてマッピング手法を探索した。本発表では、対応付けの処理をマッピングと呼ぶ。

### 3. 方法

香川大学医学部附属病院(以下、当院)の臨床検査マスターを Unicode で定義されている互換合成を用いて正規化したのち、統計解析パッケージ JMP を用いて SDTM Terminology に対応した単位名を dice 係数による一致度で評価した。フィービリティ研究として後ろ向きに 330 例のデータ収集を計画し、収集可能であった 225 例(42,544 レコード)を正規化とマッピング処理の調査に用いた。

#### 3.1 マッピングの前処理(正規化)

検査名をマッピングする前処理として、Unicode Standard Annex #15 で定義されている正規化処理を行った。この正規化処理は、文字そのものに対する正規化処理である。例えば、臨床検査値の単位には、 $\mu$  L や  $\mu$  l と言った表記揺れの他、mg/m<sup>2</sup>/day など、m<sup>2</sup> の表記が見られる。これらを一一つルール化して、確実に処理することも考えられるが、個別の対応になってしまい、施設が変わると使えないもの、必要のないものといった余剰のルールが生じる。そこで、Unicode による正規化処理では、そうした表記を一度文字の要素単位に分解し、共通部分を統一した表記にするルールを提供している。

本研究で我々は、Unicode で定義されている Compatibility Decomposition, followed by Canonical Composition を採用した。Java で実装する場合は、図 1 で表されるコードで実装し、図 2 に示す結果を得る。上記の処理では、文字を Unicode で定義される互換等価性によって分解し、正準等価性によって

再度合成される。

### 3. 結果

SDTM Terminology より、Codelist Name=Unit で絞り込みを行ったところ、CDISC Submission Value の変数から 441 種類の単位名を取得した。同様に当院の臨床検査マスターから 166 種類の単位名が得られた。SDTM Terminology と当院の臨床検査マスターの単位名をマッチングさせた dice 係数は、0.053 であった。互換合成を施した後、単位名に関しては、SDTM Terminology と当院の臨床検査マスターとの一致を評価したところ、dice 係数は 0.075 であった。一致した単位には、%, mm, kg, mg, sec cm, U/L, nmol/L などがあった。

一方、不一致だった単位名には、 $\times 10^2/\mu\text{l}$  や mL/min/1.73m<sup>2</sup> などが見られた。

### 4. 考察

Dice 係数の最大値は、1.0 であるから、現時点で標準同士のマッピングはほぼ不可能に近いと言わざるを得ない。ただし、文書の類似度の指標である dice 係数は、最大値が 1.0 であるものの、1.0 はほぼ同じ用語集であり、同じ標準を表すことになるから、標準同士の被覆割合を計算するときには、1.0 はあり得ない。詳細を検討してみると、Dice 係数を低下させる要因に、「μ」の代わりに「u」を代替的に用いるなど、SDTM Terminology 独自の仕様が見られた。また、単位の中で 2 乗や 3 乗を含むものがあるが、そうした指数の扱いが、単に文字の後に数値を並べるのか、「」で指数であることを明示的に示すのかによる違いがあった。本研究で用いた Unicode 正規化は、同じ文字で半角・全角の違いなど、人間が区別しにくいコンピュータによる表記による違いを吸収することを目指していたが、もともとの表記そのものが異なるものも多数あり、単位の互換性を保つだけでも大変な労力が要求されることが明らかになった。臨床検査値データについて評価指標を算出中である。

### 5. 結語

現時点において、臨床研究用に電子カルテから情報を抽出し、自動的に CDISC 標準に対応付けを行うことには、種々のハードルが認められる。近年、文字標準化の技術がプログラミング言語にも実装されているため、信頼性の高い情報抽出の一助になるものと考えられる。

### 参考文献

- (1)CDISC(Clinical Data Interchange Standards Consortium). CDISC | Strength Through Collaboration. 1997. (Accessed Octob 6, 2015, at <http://www.cdisc.org/>.)
- (2)Department of Health and Human Services, Food and Drug Administration. Source Data Capture From Electronic Health Records: Using Standardized Clinical Research Data., 2015. (Accessed June 26, 2015, at <https://www.federalregister.gov/articles/2015/06/26/2015-15644/source-data-capture-from-electronic-health-records-using-standardized-clinical-research-data>.)
- (3) 西本尚樹, 國方淳, 上村幸司, 他. 臨床検査データと CDISC SDTM のマッピング可能性の調査. 第 36 回日本医療情報学会

連合大会(横浜). 2016

- (4)Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. Journal of the American Medical Informatics Association : JAMIA. 2008 Mar-Apr;15(2):130-7. PMID: 18096907. doi: 10.1197/jamia.M2556.
- (5)Dziuballe P, Forster C, Breil B, Thiemann V, Fritz F, Lechtenborger J, et al. The single source architecture x4T to connect medical documentation and clinical research. Studies in health technology and informatics. 2011;169:902-6. PMID: 21893877.
- (6)Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L, et al. Implementing Single Source: the STARBRITE proof-of-concept study. Journal of the American Medical Informatics Association : JAMIA. 2007 Sep-Oct;14(5):662-73. PMID: 17600107. doi: 10.1197/jamia.M2157.
- (6)El Fadly A, Daniel C, Bousquet C, Dart T, Lastic PY, Degoulet P. Electronic Healthcare Record and clinical research in cardiovascular radiology. HL7 CDA and CDISC ODM interoperability. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2007:216-20. PMID: 18693829.

```
import java.text.Normalizer;

public static void main(String[] args) {

    ArrayList<String> altest = new ArrayList<String>();
    altest.add("コピー/μg");
    altest.add("コピー/ml");
    altest.add("コピー/μgRNA");
    altest.add("コピー/アッセイ");
    altest.add("スポット");

    String t1;
    String t2;
    Iterator<String> i = altest.iterator();
    while(i.hasNext()){
        t1= (String)i.next();
        t2= Normalizer.normalize(t1, Normalizer.Form.NFKC);
        System.out.println("正規化前:" + t1+ "正規化後:" + t2);
    }
}
```

図 1. Unicode 正規化の Java コード。

|                                  |
|----------------------------------|
| 正規化前: コピー/μg, 正規化後: コピー/μg       |
| 正規化前: コピー/ml, 正規化後: コピー/ml       |
| 正規化前: コピー/μgRNA, 正規化後: コピー/μgRNA |
| 正規化前: コピー/アッセイ, 正規化後: コピー/アッセイ   |
| 正規化前: スポット, 正規化後: スポット           |

図 2. Unicode 正規化 (Compatibility Decomposition, followed by Canonical Composition) の例。