様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ２４ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2013～2015
課題番号：２５８７０１５６

研究課題名（和文）オントロジーを利用した診療情報検索システムの開発に関する研究

研究課題名（英文）Development of Knowledge Integrated Database for Clinical Information Retrieval System

研究代表者
河添　悦昌（Kawazoe, Yoshimasa）

東京大学・医学部附属病院・助教

研究者番号：１０６２１４７７

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：生命科学分野のオントロジーを利用したクエリ拡張によって診療情報の検索を可能とするシステムの開発を目指した。本邦の標準的診療データのストレージであるSS-MIX2に含まれるHL7 v2.5メッセージをRDFデータへと変換する一般的な手法を開発した。また、一般に公開され利用可能な医薬品データベースを相互に関連付けたLinked Drug Databaseを開発するとともに、これを利用したクエリ拡張によってRDF化したHL7 v2.5メッセージを検索する方法を開発した。有害事象の検出をユースケースとして実診療データを対象とした手法の評価を行い、現実的な速度によって検索が可能であることを確認した。

研究成果の概要（英文）：This study aimed to develop a clinical information retrieval system that utilize life-science knowledge resources for query expansion. We developed a generic method to convert HL7 version 2.5 messages into the RDF, and also converted five kinds of existing drug databases into RDF and provided explicit links between the corresponding items among them. With those linked drug data, we then developed a method for query expansion to search the RDF-ized HL7 v2.5 messages using semantic information on drug classes. An evaluation of the proposed method in use case for detecting adverse drug events showed that the execution time of the developed queries increased with the amount of clinical data without diverging.

研究分野：医療情報学

キーワード：電子的診療データ　オントロジー　セマンティックウェブ　RDF　SPARQL　Health Level Seven

１．研究開始当初の背景

電子化された診療情報が増加するに伴いこれを積極的に活用することが期待されている。診療情報を二次活用するデータベースを構築する際に考慮すべき事項として、大規模データが取り扱えること、検索速度が十分であること、個人情報を安全に取り扱えること、データ抽出要件を検索式が十分に表現できること等が挙げられる。本研究ではこのうち検索の表現力に着した。臨床研究や疫学研究を目的とした活用のためには、医療分野の専門家によるデータの利便性を更に向上させる必要がある。専門知識を持つユーザーにとっては、彼らの背景知識に基づいてデータの検索が行えることが重要であり、これを実現するためには、外部の知識を用いて診療データを検索するクエリ拡張が容易に行えるフレームワークが求められる。例えば、「併用禁忌の関係にある医薬品」のうち、「CYP3A4 により代謝される医薬品」と「CYP3A4 活性を阻害する医薬品」の組みが投与されている症例をスクリーニングする際には、CYP3A4 に関連する何百とある医薬品コードを調べクエリに列挙するのではなく、詳細化された医薬品の概念と概念間の関係、つまりオントロジとして記述される知識を利用した問い合わせが可能となれば、多くの研究者のデータベース利用が促進されることが期待される。

２．研究の目的

本研究は、生命科学分野のオントロジをクエリ拡張の知識として用い、これと診療データとを連合して利用可能な RDF データベースの構築を行う。検索のユースケース検討し、実診療データを対象としたクエリ表現と実行性能の評価を行う。

３．研究の方法

診療データを検索する際のクエリ拡張の知識として利用可能なオントロジを調査する。また、これらオントロジと診療データとの対応状況を調査し、検索のユースケースを検討する。次に、本邦の標準的診療データのストレージである SS-MIX2 に含まれる HL7 v2.5 メッセージを RDF データへと変換する手法を開発する。この RDF データに対して、オントロジを使用した SPARQL クエリ拡張の手法を検討する。この際に、診療データの時間関係や時系列順序をクエリに表現できることを考慮する。開発した手法を検証するために、所属病院の実診療データを利用して、クエリの事項結果と実行速度を評価する。

４．研究成果

(1) 生活習慣病の適格基準に出現する概念の調査、ならびに利用可能な診療データとの対応状況の調査

研究当初は検索のユースケースとして生活習慣病に関する臨床試験の適格基準を満たす症例の抽出を考えていた。生活習慣病として糖尿病に焦点を絞り、診療ガイドラインと大学病院医療情報ネットワーク研究センターで公開される約 800 件の糖尿病に関する臨床試験の適格基準中に出現する医療概念と利用可能な診療データとの対応状況を調査した。その結果、既往歴、併存疾患、検体検査結果、処方歴などに関する概念の多くは、登録病名や検査・処方などのオーダ情報など構造化された診療データに対応付くが、残りは嗜好歴、家族歴、妊娠の有無、薬剤の副作用歴、血圧値、眼底所見、画像所見、心エコー所見など、診療録や各種レポートの自然言語で記載された情報源を必要とすることがわかった。また、多くの適格基準においては、後者の情報を必須とすることがわかった。各種レポート類など自然言語で記載された診療情報の種類は多岐に渡り、これを利用可能とするために時間を要すること、またそこに含まれる概念を有するオントロジの構築に時間を要することから、検索のユースケースを投薬による有害事象の検出に変更し、研究の概念実証を行うことを優先した。

(2) 複数の医薬品データベースを関連付けた Linked Drug Database の開発

薬剤の有害事象を検出するクエリを記述するためには、医薬品を様々な角度から表現できることが有用である。医薬品リソースを Linked data として活用し、複数の医薬品リソースの特徴を活かした表現で薬剤を同定し、個々の医薬品コードに解決することで、処方オーダを検索することができると考えた。例えば、非定型抗精神病薬のうちセロトニン 2C 受容体とヒスタミン 1 受容体の阻害作用を有する医薬品の処方オーダ検索は次のステップから行うことができる。１）USP 分類を利用して非定型抗精神病薬に分類される薬剤を同定する。２）USP 分類と KEGG とのリンクを利用して対応する KEGG の薬剤を同定する。このうちセロトニン 2C 受容体もしくはヒスタミン 1 受容体の阻害作用を有する薬剤に絞り込む。３）KEGG と MEDIS 医薬品マスタとのリンクを利用して MEDIS の医薬品を同定し、その医薬品コードを得る。４）得られた医薬品コードを元に処方オーダを検索する。このような考えのもと、公開される医薬品データベースを RDF とし、各リソース間で識別子のレベルで対応がとられている項目に明示的なリンクを設けることで Linked data とした。それぞれのデータベースに含まれる各項目を *rdfs:Class* とし、階層構造を有する項目間の上位下位関係には *rdfs:subclass0f* を用いた。RDF フォーマットで公開されるリソースはなかったため、提供されるデータを個別に RDF に変換した。本研究では、表 1 に示す 5 種類の医薬品データベースを相互に関連付け RDF データし、これを研究成果の一部として公開した。

| 名称 | リソースの概要 | 他リソースへのリンク | 薬剤・薬品クラス数（トリプル数） |
|---|---|---|---|
| ATC | WHOによって開発された5階層からなる薬剤の分類体系。治療的、薬理的、化学的サブグループにより一般名により薬剤を同定できる。 | KEGG SIDER | 5770 (48,504) |
| USP | 米国薬局方により開発される薬品の分類であり約50のカテゴリから構成される。 | KEGG | 1459 (7567) |
| SIDER | 医薬品の有害事象報告をその頻度と共に集計したもの。薬剤は化合物の単位で表され、有害事象はMedDraの Preferred termとして表される。 | ATC | 997 (7,848,862) |
| KEGG DRUG | 日米欧の医薬品データを化学構造と成分の観点から集約したもので、薬剤ターゲット、代謝酵素、相互作用、化学構造など多くの情報が付随する。他のリソースとの対応付けが多くなされており、リソースのハブとしても利用する。 | ATC USP MEDIS | 5780 (109,976) |
| MEDIS 医薬品マスタ | 本邦の標準であり、HOTコードを管理番号として、薬価基準収載医薬品コード、YJコード、レセプト電算処理コード、JANコードとの対応が付いている。KEGG（YJコード）をSS-MIX2で用いられるHOTコードに対応付けるために利用する。 | KEGG | 26126 (387,319) |

表 1.本研究で用いた医薬品 DB の概要

(3) 診療データの RDF 化手法の開発

本邦の標準的診療データのストレージである SS-MIX2 とそこに含まれる HL7 v2.5 メッセージを対象として RDF データとするための方法を開発した。HL7 v2.5 の標準シリアライズ形式はパイプ記号とハット記号によりデータを区切るものであるため、各データがどのような種類の情報であるかを示すメタ情報はメッセージ内に明示されない。一方、HL7 メッセージは XML によるシリアライズ形式も仕様上規定されており、例えば処方オーダ（OMP-01）は、最上位にメッセージ型を表す<RDE_011>タグが位置し、その子要素として患者情報<PATIENT>やオーダ情報<ORDER>などの各セグメントを表すタグが位置する。同様に各セグメントの下位にはタイムスタンプ<TS>やコード化文字列を表す<CWR>などデータの型を表すタグと共にデータの値が表される。XML 要素のタグをメタ情報として利用し、XML を再帰的に走査することで RDF データを生成する方法を取った。このことにより、HL7 v2.5 メッセージの各セグメントや要素に対するメタ情報を手動で付与することなく RDF データを機械的に生成することができた（図１）。

また、RDF リソースの URI は一意である必要があるが、次のように SS-MIX2 ストレージのディレクトリと XML の構造を利用することでこれを決定することが出来る。SSMIX2 ストレージは、患者 ID、診療年月日、データ種別で階層化されたディレクトリに、オーダ番号とオーダ発生日時をファイル名称とする HL7 メッセージが生成される。オーダ発生日時はミリ秒までの粒度であるため、同じディレクトリに同じ名称のファイルは作成されない。そのため、URI の上位には施設を表すドメイン名に続き患者 ID からファイルの名称までをパスに含めることで、ファイルまでの一意性を確保できる。URI の下位は XML 化した HL7 メッセージのタグの名称によりパスを構成する。ここで、ORDER や RESULT を始めとして、XML の同一階層に繰り返し出現するセグメントの場合には、その回数を URI に含めることで重複を避ける必要がある。

```
@prefix hl7v25: <http://hl7.org/v25#>.
@prefix ssmix2: <http://ssmix.org/v2#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

<http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1>
  hl7v25:PATIENT  <http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/PATIENT>;
  hl7v25:ORDER    <http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1>.

<http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1>
  hl7v25:ORC  <http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1/ORC>;
  hl7v25:RXE  <http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1/RXE>.

<http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1/RXE>
  hl7v25:RXE.2   <http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1/RXE/RXE.2>.

<http://m.u-tokyo.ac.jp/0123456789/20130518/OMP-01/.../RDE_O11/1/ORDER/1/RXE/RXE.2>
  hl7v25:CE.1  "103022801"^^xsd:string;
  hl7v25:CE.2  "BLOPRESS 4MG"^^xsd:string;
  hl7v25:CE.3  "HOT9"^^xsd:string.
```

図１． RDF 化した HL7 メッセージの例

(4) 有害事象検出をユースケースとした SPARQL 表現の開発

開発した Linked Drug Database と RDF 化した HL7 メッセージを用いることで可能となるクエリ表現を検討した。その一つとして、既知の有害事象をもとにした医薬品の同定とその処方期間に生じた有害事象の検索例を挙げる（図２）。このクエリは SIDER において白血球減少をきたす医薬品を同定し、これら医薬品が処方された症例のうち、各処方期間中に白血球数の低下を伴った症例を検索する。より具体的には、SPARQL の Federation クエリの機能を用い、SERVICE 節内で SIDER において 30％以上の頻度で”leukopenia”もしくは”neutropenia”を生じる医薬品を同定し、ATC、KEGG、MEDIS 医薬品マスタを利用して個々の医薬品コードに解決する。続くトリプルのパターンマッチでは、処方オーダの医薬品コード、処方量、処方日、処方日数と、白血球数とその検査日を各変数にバインドし、処方オーダ日と処方日数からなる処方期間中に白血球数が 3000 以下である症例を検索する。

```
PREFIX …
SELECT DISTINCT ?patient  ?drug_code  ?dose_per_day
 ?prescription_date ?duration
WHERE {
# ① SIDER で定義される白血球減少をきたす薬剤を HOT コードに解決
 SERVICE <http://location-1:8890/sparql> {
   GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdf/sider> {
     ?sider    sider:link_atc        ?atc;
               sider:ae              ?ae.
     ?ae       rdfs:label            ?label;
               sider:lower_bound  ?lb.
     FILTER (regex(?label, 'leukopenia', 'i') || regex(?label, 'neutropenia', 'i'))
     FILTER (?lb > 0.3)
   }
   GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdf/atc> {
     ?atc      atc:link_kegg         ?kegg.
   }
   GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdf/kegg> {
     ?kegg     kegg:link_medis       ?medis.
   }
   GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdf/medis/drug> {
     ?medis    medisd:hot9_code  ?drug_code.
   }
}
# ② HOT コードをもとに処方オーダ(OMP-01)を検索
  ?patient   ssmix2:OMP-01  [hl7:RDE_O11    ?rdeo11].
  ?rdeo11    hl7:ORDER      [hl7:RXE    ?rxe;  hl7:TIMING   ?timing].
  ?rxe       hl7:RXE.2      [hl7:CE.1   ?drug_code];
             hl7:RXE.19     [hl7:CQ.1   ?dose_per_day]
  ?timing    hl7:TQ1        [hl7:TQ1.7   [hl7:TS.1  ?prescription_date]];
             hl7:TQ1        [hl7:TQ1.6   [hl7:CQ.1  ?duration]].
# ③処方期間中の白血球数で 3000 以下のものを検索
  ?patient   ssmix2:OML-11  [hl7:OUL_R22 ?oulr22].
  ?oulr22    hl7:SPECIMEN   [hl7:ORDER   [hl7:RESULT [hl7:OBX  ?obx]]].
  ?obx       hl7:OBX.3      [hl7:CE.1 '2A990000001992052'^^xsd:string];
             hl7:OBX.5      ?lab_value;
             hl7:OBX.14     [hl7:TS.1   ?lab_date].
  FILTER (?prescription_date < ?lab_date && ?lab_date <
          bif:dateadd('day', ?duration, ?prescription_date)).
  FILTER (?lab_value < 3.0).
}
```

図２.有害事象として白血球の減少をきたす薬剤の投与後に白血球数の減少を生じた症例を検索する SPARQL クエリ

(5) SPARQL クエリの実行可能性の検証

有害事象をユースケースとした３種類の SPARQL クエリの実行可能性を検証した。東京大学医学部附属病院の有する SS-MIX2 ストレージの 3 年分の処方オーダ(RDE^011)と検体検査結果(OUL^R22)の HL7 メッセージを対象として前述の方法によって RDF データを生成した。対象とした HL7 メッセージ数はそれぞれ 190 万と 210 万であり、生成された RDF トリプル数はそれぞれ6.5億､7.9億であった。各クエリの実行結果を表２に示す。クエリ 1 の「レニン・アンジオテンシン阻害剤」という表現は、Linked Drug Database により 476 種の異なる HOT コードに解決され、これら医薬品の全処方は 197366 件であった。同様にクエリ 2 は 130 種類、クエリ 3 は 78 種類の HOT コードに解決され、それぞれ 1171 件と 58 件の結果が得られた。

| クエリ | 検索条件の概要 | 医薬品リソース | HOT コード数 | 結果(件) |
|---|---|---|---|---|
| 1 | レニン・アンジオテンシン阻害剤に分類される医薬品が処方された症例とその全処方。 | ATC KEGG MEDIS | 476 | 197366 |
| 2 | 有害事象として白血球減少を有する医薬品の処方期間中に、白血球数3000以下を認めた症例と該当の処方。 | SIDER ATC KEGG MEDIS | 130 | 1171 |
| 3 | 非定型抗精神病薬のうち5HT2CもしくはH1受容体阻害作用を有する医薬品の初回処方と最終処方の期間中に境界型糖尿病の基準を満たす症例とその処方薬、処方期間。除外条件として、初回処方日の過去の60日以内に前述の基準を満たさないこと。 | USP KEGG MEDIS | 78 | 58 |

表２．３種類の SPARQL クエリの実行結果

(6) SPARQL クエリの実行速度の評価

本研究の手法による診療データの検索が現実的な応答速度であることを示すために、クエリの応答時間を計測した。診療データの増加によるクエリの応答時間を比較するために、前述の 3 年分の処方オーダと検体検査結果の HL7 メッセージを 10 等分し、10%ずつ増加させた際の応答時間を計測した。また、本研究の提案手法である、SPARQL の Federation 機能により Linked Drug Database にアクセスして医薬品コードを解決した場合と、従来的な方法である医薬品コードをクエリに直接記述した場合とで応答時間の比較を行った。前者の比較に関して、最も単純なクエリ１の応答時間は診療データの増加に伴い対数増加を示し、クエリ 2 と 3 は線形増加を示し、いずれも発散することはなかった。また、後者の比較に関して、提案手法の応答時間は従来手法の約 1.5 倍を要したが、これは、Linked Drug Database にアクセスして医薬品コードを解決する際のオーバーヘッドに起因するものであった。

(7) 研究成果の総括

本研究は、標準的診療データの更なる活用を目指し、医薬品リソースに記述される詳細情報をクエリ拡張の知識として用い、本邦の標準的な診療データである HL7 メッセージの検索が可能であることを示した。提案手法は知識リソースと診療データとを分離したクエリ表現が実行可能であることを示すものであり、このことは知識、つまりオントロジをより充実させることで、診療データの利用性を向上させる可能性を示唆するものである。

本研究のもう一つのチャレンジは、HL7 メッセージの標準構造に変更を加えることなく、一般的なアルゴリズムにより RDF データとした場合でも SPARQL による検索が行えることを示した点にある。このことは、SS-MIX2 ストレージを有する施設においての適用可能性を示唆する。

研究当初に予定していた、臨床試験の適格基準のクエリ表現に関しては、診療録や各種レポート類など自然言語で記載された診療情報から非常に多岐にわたる項目を抽出し、それら項目を表すオントロジと構造的なデータモデルを開発する必要があり、これを行うためのオントロジの開発と診療情報とのマッピングに時間を要することから、本研究期間中に十分に達成することが出来なかった。

５．主な発表論文等

〔雑誌論文〕（計１件）国際英文（査読有り）

Kawazoe Y, Imai T, Ohe K, A Querying Method over RDF-ized Health Level Seven v2.5 Messages Using Life Science Knowledge Resources, JMIR Med Inform 2016;4(2):e12.

〔学会発表〕（計１件）国内学会（査読有り）

河添 悦昌，今井 健，大江和彦．Linked Data による SS-MIX2 標準化ストレージの活用に向けて−薬剤の有害事象検出を目的とした federation query の実行可能性の検討−．医療情報学 34(Suppl.)，pp.328-331，2014．
2014 年 11 月 6 日 幕張メッセ国際会議場(千葉県・幕張市)

〔その他〕

1) 研究代表者紹介
http://www.m.u-tokyo.ac.jp/medinfo/?page_id=182

2) 研究成果公開（Linked Drug Database）
https://github.com/linked-drug-data/publish

６．研究組織

(1)研究代表者
河添 悦昌（KAWAZOE YOSHIMASA）
東京大学・医学部附属病院・助教
研究者番号：10621477
(2)研究分担者：なし
(3)連携研究者：なし