

Temporal Pattern 3: Identifying Adverse Events During a Period Between the Initial and Final Medications

This pattern assumes that the impact of a certain drug extends from its initial medication date to its final medication date, and it is to identify the adverse events during this period. The drug administration period estimated in this pattern could include extended time periods during which the drug had not been administered, and thus, it is possible that the defined drug administration period significantly deviates from the actual drug administration period. However, because the effects of some drugs could continue for an extended time period after drug administration has ended, this pattern identifies adverse events of these types of drugs whose effects extend beyond the end of the medication period.

Temporal Pattern 4: Excluding Adverse Events Immediately Before an Initial Medication

This pattern is to increase the degree of certainty of a causal relationship between a drug and an adverse event by excluding the adverse events immediately before initial medication of the drug.

Experiment Settings

In the next section, we first show a summary of created RDF data to use in this experiment. To ensure the impartiality of the benchmark results, all segments of the HL7 messages were converted to RDF data, rather than arbitrarily deleting unnecessary segments. We then explain three types of query in use cases for detecting ADEs, which are available in our proposed method and show the execution results of searches using these queries. The goal of our experiments was not to investigate specific adverse events, but rather show that it is possible to search over the RDF-ized HL7 messages using SPARQL queries that combine external knowledge and temporal patterns. So, we finally present results of a benchmark that measures the execution time to show that the searches over RDF-ized HL7 messages through SPARQL provide a feasible response speed.

To show the relationships between the execution time of the query and the amount of data, we divided whole HL7 messages equally into 10 subdatasets in which the HL7 messages were arranged in ascending order of the date of administration. Then, we measured the execution time of the queries issued five times at each point by increasing every subdatasets. We tested two types of query expressions for each three query in order to compare our proposed query expression with a conventional one. The proposed query uses the Linked Drug Data dynamically by SPARQL federation function in the manner as shown in Figures 5 to 7 below. The conventional query enumerates the

individual drug codes in SPARQL filter keyword in advance, which were obtained from the Linked Drug Data separately. Thus, the execution time of the proposed query included, (1) a time to search for individual drug codes from an expression like "renin angiotensin inhibitors" by accessing to Linked Drug Data and (2) a time to search for medication records of RDF-ized HL7 data based on the searched drug codes. On the other hand, the execution time of the conventional query did not include a time to search for the individual drug codes because they are enumerated in advance.

We measured the execution time after relaunching the RDF store and clearing the cache each time a query was executed. Therefore, the execution time included the time it takes to load the data to memory, execute the query, and display the execution results. As we observed that the execution speed dropped drastically when SPARQL queries were not completely optimized through automatic optimization, we manually optimized the execution sequences and then locked them using functionality available in Virtuoso. With regard to the environment for executing queries, the RDF-ized HL7 messages and the Linked Drug Data were stored in two different SPARQL endpoints on a secure network. For the RDF-ized HL7 messages, we used hardware with Intel Xeon 2.60 GHz processors and 256 GB random access memory (RAM). For the Linked Drug Data, we used hardware with Intel Xeon 2.20 GHz processors and 128 GB RAM. Both pieces of hardware ran the CentOS6.5 operating system, and Virtuoso Open-Source Edition 7.1.0 was used as the RDF store.

Results

Converted RDF Data

The University of Tokyo Hospital is an educational hospital with more than 1100 beds and 760,000 visits annually. Since 2011, the hospital has been collecting data in the form of HL7 messages in a SS-MIX2 storage. From these collected data, we used the medication orders and laboratory test results during the 3-year period from January 1, 2011 to December 31, 2013. There were approximately 148,000 unique patients, and the number of HL7 messages included was 1.9 million for RDE^O11 (medication orders) and 2.1 million for OUL^R22 (laboratory test results). We then converted them into RDF using the method explained earlier. Approximately 650 million RDF triples for RDE^O11 and 790 million RDF triples for OUL^R22 were converted, and the average number of triples in one message was 360. It was also that the approximate time to convert HL7 messages into RDF were 17 hours and 30 minutes for RDE^O11 and 25 hours 10 minutes for OUL^R22 when we used single CPU (Table 2).

Table 2. Summary of the RDF-ized HL7 messages.

Type of HL7 message	Information content	Number of HL7 messages (million)	Number of RDF triples (million)	Triples in a message	Time to convert HL7 messages into RDF
RDE^O11	Medication order	1.9	650	342	17 hours 30 minutes
OUL^R22	Laboratory test result	2.1	790	376	25 hours 10 minutes
Total	-	4.0	1440	360	42 hours 40 minutes

SPARQL Expressions for Searching Adverse Events

Query 1: Identifying Drugs Based on Pharmaceutical Classification and Searching For All Relevant Medication Orders

This is the most basic query searching medication orders that are classified in a certain pharmaceutical category. The query (Figure 5) searches for all medication orders for drugs classified as renin angiotensin inhibitors. The SERVICE clause that follows the WHERE clause queries the Linked Drug Data stored

at a SPARQL endpoint, identifies all ATC subclasses of renin angiotensin inhibitors, and resolves their individual drug codes through the KEGG and MEDIS DRUG. When this finishes, triple pattern matching identifies the patients who were prescribed drugs with the code that the SERVICE clause resolved, and binds the dosage amount, medication date, and number of medication days to their corresponding variables of the patients. Query results are returned in a table with the column names described in variables of the SELECT statement. This query does not consider the temporal relationship with other events; thus, it is for Temporal Pattern 1.

Figure 5. SPARQL expression of Query 1. This query searches all medication orders for drugs classified as renin angiotensin inhibitors.

```

SELECT DISTINCT ?patient ?drug_code ?dose_per_day
               ?prescription_date ?duration
WHERE {
  SERVICE <http://location-1:8890/sparql> {
    GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdf/atc> {
      ?atc      rdfs:label      ?atc_name.
      FILTER regex(?atc_name, 'AGENTS ACTING ON
                        THE RENIN-ANGIOTENSIN SYSTEM', 'i').
      ?atc_sub  rdfs:subClassOf ?atc.
      ?atc_sub  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    medis:hot9_code  ?drug_code.
    }
  }
  ?patient    ssmix2:OMP-01 [hl7:RDE_O11 ?rdeo11].
  ?rdeo11     hl7:ORDER      [hl7:RXE      ?rx;
                              hl7:TIMING    ?timing].
  ?rx         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]].
}

```

Query 2: Identifying Drugs Based on Known Adverse Events and Searching for Adverse Events During the Relevant Medication Periods

This query identifies drugs from known adverse events registered in SIDER 2 and searches for clinical cases that may include adverse events resulting from the identified drugs. Specifically, we consider a query (Figure 6) to identify drugs that cause leukopenia or neutropenia as adverse events in SIDER 2 and to search for the clinical cases where the identified drugs were prescribed and a drop in the leukocyte counts was observed during each medication period. Similar to Query 1, the

SERVICE clause identifies the drugs that cause leukopenia or neutropenia at a frequency of 30% or higher in SIDER 2 and resolves their individual drug codes through the ATC, KEGG, and MEDIS DRUG. When this finishes, triple pattern matching binds the drug codes of the prescribed drugs, dosage amounts, medication dates, duration of each medication, leukocyte counts, and its examination date to their corresponding variables, and then searches for clinical cases where the leukocyte counts was 3000 or less during the medication period (defined as the period starting on the day of the medication order and continues for the number of prescribed days). Because this query searches for adverse events during each medication period, it is for Temporal Pattern 2.

Figure 6. SPARQL expression of Query 2. This query searches all cases for which a leukocyte count of 3000 or less was observed during the medication period of drug types having leukopenia or neutropenia as adverse events.

```

SELECT DISTINCT ?patient ?drug_code ?dose_per_day
               ?prescription_date ?duration
WHERE {
  SERVICE <http://location-1:8890/sparql>{
    GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdf/sider>{
      ?sider sider:link_atc      ?atc;
            sider:a             ?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 medis:hot9_code   ?drug_code.
    }
  }}
?patient  ssmix2:OMP-01 [hl7:RDE_O11?rdeo11].
?rdeo11  hl7:ORDER      [hl7:RXE ?rx; hl7:TIMING ?timing].
?rx      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]].

?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).
}

```

Query 3: Identifying Drugs Based on Pharmaceutical Classification and Their Targets, and Searching for Adverse Events During the Relevant Drug Medication Periods

This query illustrates that when multiple drug data resources are used, drugs can be identified with more detailed characteristics. In clinical backgrounds, atypical antipsychotic drugs are known to have a tendency to trigger diabetes. It is hypothesized that these drugs cause chronic bulimia by blocking 5HT_{2C} and H₁ receptors and bring about obesity and hyperinsulinemia, thereby inducing diabetes [34]. This query may help examine this hypothesis through identifying the drugs that demonstrate these characteristics and extracting clinical cases that satisfy the criteria for diabetes during the medication

period. As mentioned above, this query (Figure 7) first narrows down drugs classified as atypical antipsychotic drugs in the USP classification to those in KEGG having an inhibitory effect on 5HT_{2C} or H₁ receptors, and then resolves individual drug codes through MEDIS DRUG. It then uses a filter operation to derive the initial and final medication dates for each patient from the medication orders of the drugs with the resolved drug codes, and extracts clinical cases where the HbA_{1c} value or the serum glucose satisfies the criteria for impaired glucose tolerance during the medication period. Note that as the HbA_{1c} value changes gradually, we used the period between the initial and final medications, rather than using each medication period. We also added a condition to exclude clinical cases satisfying the same criteria within 60 days of the initial medication. Therefore, this query is for a combined temporal pattern of Temporal Patterns 3 and 4.

Figure 7. SPARQL expression of Query 3. This query searches all cases satisfying the criteria for impaired glucose tolerance during a period between the initial and final medications of atypical antipsychotic drugs that have a 5HT2C or H1 receptor inhibitory effect. The clinical cases that satisfy the above criteria within 60 days of the initial medication are excluded. In this query, two subqueries are used. In subquery 1, the cases having the period of initial and final medications of the atypical antipsychotic are identified. In subquery 2, the cases satisfying the criteria for impaired glucose tolerance during the period are identified.

```

SELECT DISTINCT ?patient ?hot_code ?first ?last ?lab_date ?lab_value
WHERE {
  FILTER NOT EXISTS {
    ?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 ?lab_code];
        hl7:OBX.5 ?lab_value;
        hl7:OBX.14 [hl7:TS.1 ?lab_date].
    FILTER ( (?lab_code = '0170700_84' && ?lab_value > 6.5) ||
            (?lab_code = '3D045000001920402' && ?lab_value > 6.2)).
    FILTER (bif:dateadd('day', -60, ?first) < ?lab_date && ?lab_date < ?first)
  }
  { # SUBQUERY 2
    SELECT ?patient ?drug_code ?first ?last ?lab_date ?lab_value
    WHERE {
      ?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 ?lab_code];
          hl7:OBX.5 ?lab_value;
          hl7:OBX.14 [hl7:TS.1 ?lab_date].
      FILTER ( (?lab_code = '0170700_84' && ?lab_value > 6.5) ||
              (?lab_code = '3D045000001920402' && ?lab_value > 6.2)).
      FILTER (?first < ?lab_date && ?lab_date < ?last) .
    }
    { # SUBQUERY 1
      SELECT ?patient ?drug_code (MIN(?p_date) AS ?first)
                                   (MAX(?p_date) AS ?last)
      WHERE {
        SERVICE <http://location-1.8890/sparql> {
          GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdi/medis/drug> {
            ?medis medis:hot9_code ?drug_code;
                  medis:kokuji_maisho ?brand_name;
          }
          GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdi/kegg> {
            ?kegg kegg:link_medis ?medis;
                  kegg:target ?target.
            FILTER (regex(?target, 'H1-receptor antagonist', 'i') ||
                    regex(?target, '5-HT2C-receptor antagonist', 'i')).
          }
          GRAPH <http://www.m.u-tokyo.ac.jp/medinfo/rdi/usp> {
            ?usp rdfs:label ?label.
            FILTER (regex(?label, '2nd Generation/Atypical', 'i')).
            ?usp_sub rdfs:subClassOf ?usp;
                    usp:link_kegg ?kegg.
          }
        }
        ?patient ssmix2:OMP-01 [hl7:RDE_O11 ?rdeo11].
        ?rdeo11 hl7:ORDER [hl7:RXE ?rx; hl7:TIMING ?timing].
        ?rx hl7:RXE.2 [hl7:CE.1 ?drug_code].
        ?timing hl7:TQ1 [hl7:TQ1.7 [hl7:TS.1 ?p_date]].
      }
    }
  }
}

```

Execution Results of Each Query

Table 3 shows the results of executing each query over the RDF-ized HL7 messages for a 3-year period. The Query 1 expression of "drugs classified as renin angiotensin inhibitors" yielded 476 different types of drug codes by the Linked Drug Data, and there were a total of 197,366 medication orders found for these drugs. Similarly, the Query 2 expression of "drug types

having leukopenia as adverse events" yielded 131 types of drug codes using SIDER 2, and the Query 3 expression of "of the atypical antipsychotic drugs, those having a 5HT2C or H1 inhibitory effect" yielded 78 drug types, with Queries 2 and 3 obtaining 1171 and 58 results, respectively.

Query Execution Performance

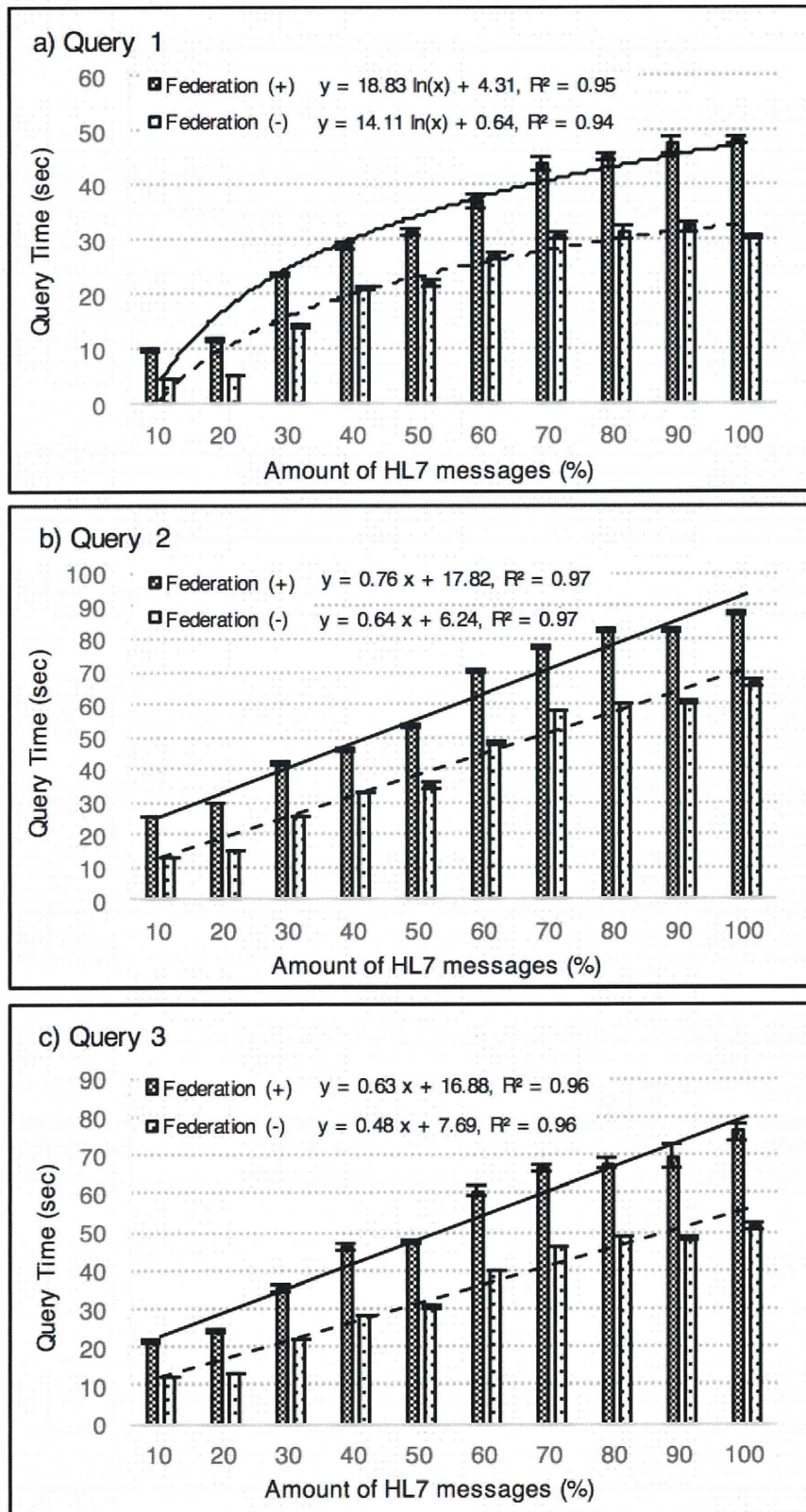
Figure 8 shows, for each three query, the average measured execution times of the two types of query expression (ie, our proposed query that use Linked Drug Data dynamically with SPARQL federation function and a conventional query in which the individual drug codes are enumerated in SPARQL filter

keyword). The average execution time of the proposed queries were significantly longer than the conventional one, and these were 49% longer in Query 1, 43% in Query 2, and 51% in Query 3, in total. It was also that the execution time of the Query 1 showed logarithmic growth, and the Query 2 and the Query 3 showed linear growth. The coefficient of determination in these regressions ranged from 0.94 to 0.97.

Table 3. Summary of each query and the respective execution results.

No.	Summary of query condition	Drug data sources to resolve the expression	No. of resolved drug codes	Results
1	Cases for which drug types classified as renin angiotensin inhibitors were prescribed, and all medications of such drugs.	ATC KEGG MEDIS DRUG	476	197,366
2	Cases for which a leukocyte counts of 3000 or less was observed during the medication period of drug types having leukopenia or neutropenia as adverse events, and all corresponding medications.	SIDER 2 ATC KEGG MEDIS DRUG	131	1171
3	Cases satisfying the criteria for impaired glucose tolerance during a period between the initial and final medications of atypical antipsychotic drugs that have a 5HT _{2C} or H ₁ receptor inhibitory effect. Clinical cases that satisfy the above criteria within 60 days of the initial medication are excluded.	USP KEGG MEDIS DRUG	78	58

Figure 8. The average measured execution times of Queries 1, 2, and 3 obtained through the experiments are shown in a), b), c), respectively. In each subfigure, bar graphs represent the average measured execution times of the two types of query expression with standard errors, and solid or dashed line represent the approximate average execution times.



Discussion

Primary Findings

To further improve the usability of EMRs, EMRs need to be integrated with external data sources that serve as knowledge resources. Currently, clinical specialists provide and interpret the knowledge used in making clinical data inquiries and, in many cases, manually translate the knowledge into codes of terminology and describe them in queries. This not only requires time and increases the number of errors [13,15] but also leads to the possibility of differing interpretations of coding, resulting in incompatibility among query results. Semantic Web technology provides a framework for integrating heterogeneous data sets using RDF and enables the extraction of data from multiple endpoints on a network using queries in a uniform format and standard Web protocols. This makes it possibly not only to integrate heterogeneous knowledge resources but also to share publicly available resources as knowledge sources and to handle highly confidential clinical data without compromising their confidentiality.

We converted drug databases to the Linked Drug Data, used them as the knowledge for query expansions, and searched over the RDF-ized HL7 messages. We showed three queries illustrated by the queries for drugs including renin angiotensin inhibitors, as well as more advanced expressions for drugs that cause leukopenia and also for atypical antipsychotic drugs. We only show three queries, although, we believe that wider ranges of queries are possible by combining four temporal patterns and various search expressions to identify drugs. These query expressions require clinical knowledge, and such knowledge must be supplied from external knowledge sources, as clinical data do not contain such knowledge. Our query expression used knowledge of drugs separate from clinical data that exist at a different endpoint on a network through SPARQL's federation query. This suggests that enhancing knowledge resources would improve the search usability of clinical data and the possibility to search over clinical data on a shared knowledge basis.

The Query 1 example resulted in 476 drug code types for renin angiotensin inhibitors. However, in reality, it is unlikely that one hospital adopts all types of renin angiotensin inhibitors, and only a few types are actually adopted by any one hospital. Because different hospitals may adopt different drug types, the drug codes listed for one hospital may not apply to another. The proposed method dynamically resolves the expression like "renin angiotensin inhibitors" using external knowledge resources, enabling clinical data searches using expressions at a level close to the knowledge without considering specific types of drugs that different hospitals may adopt. This not only improves the usability of query expressions for specialists but also suggests the possibility of reusing queries (ie, using the same query at multiple hospitals) [34,35].

The Query 2 example showed a use case for ADEs, which used SIDER 2 to search drugs that potentially cause leukopenia. As for the database of ADEs itself, there is another publicly available database named ADEpedia 2.0 that use RxNorm codes for medications and SNOMED CT or MedDRA codes for phenotypes related to ADEs [36]. In this database, the

relationships between the drugs and ADEs are represented by predicates such as 'contraindicated_drug' for information of contraindications and 'causative_agent_of' for adverse drug effects. Although SIDER 2 and ADEpedia 2.0 is useful to search known relationships between the drugs and ADEs, they are not necessarily enough for a use case to investigate unknown ADEs that may be discovered from EMR. To enable this, we needed to complement them by using the different type of drug database. We showed Query 3 example that make use of the information of drug class and type of receptor, which are enabled by linking USP and KEGG. Although this query shows a limited example, increasing variation of the search expression by using multiple drug database is assumed to be useful for investigating ADEs, and in order to do so, it is primarily important that these databases can be linked each other.

We showed a method for converting RDF data not by selecting arbitrary elements contained in the HL7 message but by using all the elements as they are. The reason why is because it was difficult to specify which elements are necessary for a clinical study in advance. As a trade-off, the SPARQL query we showed may be difficult to describe unless we are familiar with the specifications of the HL7 message. The difficulty of describing this SPARQL query will be summarized in the following three points. First, when describing the pattern matching of SPARQL, nesting up to reaching the necessary elements would be considerably deep. For example, until reaching the drug code, it is necessary to pass through five nodes: RDE_O11, ORDER, RXE, RXE.2, and CE.1. For this reason, the user must be familiar with the structure of the HL7 message. In order to solve this problem, it is conceivable to select the elements that are required for a clinical study from the HL7 message, reconstructing a simpler model of RDF data composed of only its elements. To do this, a guideline for which elements should be converted might be useful, and to make such a guideline, it is desirable that Health Level Seven and some associations related to clinical research discuss and select the required elements necessary for clinical researches in general. Second, the vocabulary that is reusable to represent the RDF resource is not used. Some properties such as "patient ID," "birthdate," and "gender" shown in Figure 2 might be good to associate them with the existing vocabulary that is defined in the ontology such as foaf and vCard. However, the vocabulary corresponding to almost all other HL7 elements, including the drug code, medication dose, unit of the dose, and so on did not exist as far as we know. Therefore, in this study, we gave greater importance to keeping the consistency of the method of converting the HL7 to RDF by using the names of the tags obtained when converting the HL7 to XML as the vocabulary rather than reusing only those few vocabulary. Finally, temporal reasoning is important for investigation for ADEs, although, it might be difficult to write it against our RDF-ized HL7 data with SPARQL. We used filter-based solution in Queries 2 and 3 to compare the date of laboratory test results and the date of the medications in order to be able to consider the causal relationships between them. We also used subquery solution in Query 3 to identify the first and the last time of medications of atypical antipsychotic drugs in order to identify diabetes that occurred or not occurred during time frames based on the two time points. Although we showed these queries as possible as simple, they might be typically

verbose and difficult to write. It is conceivable that using Allen's temporal predicates such as "before," "after," and "during" in the pattern matching of the query [14] is useful to avoid the SPARQL filter-based comparison of the time. In order to do that, an interval-based temporal information should be given to the comparable events and they should be connected according to their relationships when the RDF data are created. It might be also that giving a mark to specific time events such as the first and the last time of medications is useful to identify them without the subquery solution. These methods make the description of the query more concise at the expense of computational complexity at the time of creating RDF data. In this study, we did not apply these methods because we focused on using all elements in HL7 message as they are, it would be worth to consider to make the expression of temporal reasoning concise.

Regarding the query execution time, we tested two types of query expression for each three query to show the difference of the execution time between our proposed query expression and a conventional one. As for the conventional expression, the number of the drug codes enumerated in each query were 476, 131, and 78, respectively, as shown in Table 3. The advantage of the proposed query is that the expression is concise and human readable in comparison to the conventional one, and that allows identification of drugs based on the detailed information rather than the drug codes can be listed. On the other hand, the disadvantages are that it is inferior in execution time, it takes approximately 40% to 50% more time than conventional one. It was also that what kind of drug code will be searched is unknown until the query is run. These comparative aspects indicate a trade-off between simplicity of the query expression and the execution time of the query as well as search reliability. In particular, as it is necessary to separately consider the reliability of the drug code obtained by the Linked Drug Data, this can be noted as one of the limitations of this study.

The result of the experiment also showed that the average execution time of the Query 1 showed logarithmic growth, and the time of the Queries 2 and 3 showed linear growth with the coefficient of determination ranged from 0.94 to 0.97. This indicates that these regressions approximated the query execution time well. These results might be counterintuitive especially in the logarithmic growth in Query 1, although, it was assumed to be possible that the logarithmic growth is consistent with computational complexity of B-Tree indices is $O(\log n)$, which are used in the RDF database we used. Although

the result will not be generalized because an execution time of a query depends on various settings, such as amount of data, the content of the query, and the kind of the database system, the execution time of these queries increased with the amount of data without diverging in our experiments.

Limitations

We converted HL7 messages to RDF data automatically without changing the HL7 message structures. This suggests that the proposed method can be applied not only at the University of Tokyo Hospital that has adopted SS-MIX2 storage but also at numerous other hospitals that use HL7 messages. To demonstrate this, future research is required to verify the applicability of the proposed method at multiple hospitals. In addition, we considered adverse events cases in our research, and thus, it was medication orders and laboratory test results that were converted to RDF data. However, HL7 messages contain other types of clinical data such as patient demographics, diagnostic disease, and some kind of order information. When these types of clinical data are converted to RDF data, a wider variety of query expressions are required to search over the converted RDF data, and future research should examine such query expressions. We have not verified the drugs identified through our query expansions, nor verified extracted clinical data against the gold standard, and these are the limitations of the research.

Conclusions

This study applied Semantic Web technology to use publicly available drug databases as the knowledge for query expansions and demonstrated clinical data searches through SPARQL. The proposed method executed queries with knowledge resources separate from clinical data, suggesting that enhancing knowledge resources would improve the usability of clinical data. This study also converted HL7 messages to RDF data using an automatic way without modifying the HL7 message structures and demonstrated searches over the converted RDF data using SPARQL. This suggests that the proposed method can be applied not only at the University of Tokyo Hospital that has adopted SS-MIX2 storage but also at numerous other hospitals that use HL7 messages. We have not verified the drugs identified through query expansions, nor verified extracted clinical data; such verifications will be performed in future research. Future research also includes applying the proposed method at other hospitals and supporting a wider variety of HL7 messages.

Acknowledgments

This research was supported by the JSPS KAKENHI Grant Number 25870156, the Cabinet Office, Government of Japan and the JSPS through the Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program), and the Center of Innovation Program from Japan Science and Technology Agency, JST.

Conflicts of Interest

None declared.

References

1. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, MIT Critical Data Conference 2014 Organizing Committee. Making big data useful for health care: a summary of the inaugural mit critical data conference. *JMIR Med Inform* 2014;2:e22 [FREE Full text] [doi: [10.2196/medinform.3447](https://doi.org/10.2196/medinform.3447)] [Medline: [25600172](https://pubmed.ncbi.nlm.nih.gov/25600172/)]
2. Moseley ET, Hsu DJ, Stone DJ, Celi LA. Beyond open big data: addressing unreliable research. *J Med Internet Res* 2014;16:e259 [FREE Full text] [doi: [10.2196/jmir.3871](https://doi.org/10.2196/jmir.3871)] [Medline: [25405277](https://pubmed.ncbi.nlm.nih.gov/25405277/)]
3. Bizer C, Heath T, Berners-Lee T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems* 2009;5:1-22. [doi: [10.4018/jswis.2009081901](https://doi.org/10.4018/jswis.2009081901)]
4. Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;41:706-716 [FREE Full text] [doi: [10.1016/j.jbi.2008.03.004](https://doi.org/10.1016/j.jbi.2008.03.004)] [Medline: [18472304](https://pubmed.ncbi.nlm.nih.gov/18472304/)]
5. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27-30 [FREE Full text] [Medline: [10592173](https://pubmed.ncbi.nlm.nih.gov/10592173/)]
6. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;38(Database issue):D355-D360 [FREE Full text] [doi: [10.1093/nar/gkp896](https://doi.org/10.1093/nar/gkp896)] [Medline: [19880382](https://pubmed.ncbi.nlm.nih.gov/19880382/)]
7. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25:1251-1255 [FREE Full text] [doi: [10.1038/nbt1346](https://doi.org/10.1038/nbt1346)] [Medline: [17989687](https://pubmed.ncbi.nlm.nih.gov/17989687/)]
8. UniProt C. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2007 Jan;35(Database issue):D193-D197 [FREE Full text] [doi: [10.1093/nar/gk1929](https://doi.org/10.1093/nar/gk1929)] [Medline: [17142230](https://pubmed.ncbi.nlm.nih.gov/17142230/)]
9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25-29 [FREE Full text] [doi: [10.1038/75556](https://doi.org/10.1038/75556)] [Medline: [10802651](https://pubmed.ncbi.nlm.nih.gov/10802651/)]
10. Katayama T, Wilkinson MD, Micklem G, Kawashima S, Yamaguchi A, Nakao M, et al. The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J Biomed Semantics* 2013;4:6 [FREE Full text] [doi: [10.1186/2041-1480-4-6](https://doi.org/10.1186/2041-1480-4-6)] [Medline: [23398680](https://pubmed.ncbi.nlm.nih.gov/23398680/)]
11. Katayama T, Wilkinson MD, Aoki-Kinoshita KF, Kawashima S, Yamamoto Y, Yamaguchi A, et al. BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains. *J Biomed Semantics* 2014;5:5 [FREE Full text] [doi: [10.1186/2041-1480-5-5](https://doi.org/10.1186/2041-1480-5-5)] [Medline: [24495517](https://pubmed.ncbi.nlm.nih.gov/24495517/)]
12. Teodoro D, Pasche E, Gobeill J, Emonet S, Ruch P, Lovis C. Building a transnational biosurveillance network using semantic web technologies: requirements, design, and preliminary evaluation. *J Med Internet Res* 2012;14:e73 [FREE Full text] [doi: [10.2196/jmir.2043](https://doi.org/10.2196/jmir.2043)] [Medline: [22642960](https://pubmed.ncbi.nlm.nih.gov/22642960/)]
13. Assélé KA, Choquet R, Mels G, Daniel C, Charlet J, Jaulent M. An ontological approach for the exploitation of clinical data. *Stud Health Technol Inform* 2013;192:142-146. [Medline: [23920532](https://pubmed.ncbi.nlm.nih.gov/23920532/)]
14. Riazanov A, Klein A, Shaban-Nejad A, Rose GW, Forster AJ, Buckeridge DL, et al. Semantic querying of relational data for clinical intelligence: a semantic web services-based approach. *J Biomed Semantics* 2013;4:9 [FREE Full text] [doi: [10.1186/2041-1480-4-9](https://doi.org/10.1186/2041-1480-4-9)] [Medline: [23497556](https://pubmed.ncbi.nlm.nih.gov/23497556/)]
15. Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. *J Biomed Semantics* 2012;3:10 [FREE Full text] [doi: [10.1186/2041-1480-3-10](https://doi.org/10.1186/2041-1480-3-10)] [Medline: [23244446](https://pubmed.ncbi.nlm.nih.gov/23244446/)]
16. Pathak J, Kiefer RC, Chute CG. Using linked data for mining drug-drug interactions in electronic health records. *Stud Health Technol Inform* 2013;192:682-686 [FREE Full text] [Medline: [23920643](https://pubmed.ncbi.nlm.nih.gov/23920643/)]
17. Health Level Seven International. URL: <http://www.hl7.org/about/index.cfm> [accessed 2016-03-07] [WebCite Cache ID [6FcscLBTW](https://www.webcitation.org/6FcscLBTW)]
18. W3C Recommendation. RDF Schema 1.1 URL: <http://www.w3.org/TR/rdf-schema/> [accessed 2016-03-06] [WebCite Cache ID [50iWdofwS](https://www.webcitation.org/50iWdofwS)]
19. W3C Recommendation. SPARQL 1.1 Overview URL: <http://www.w3.org/TR/sparql11-overview/> [accessed 2016-03-06] [WebCite Cache ID [6cVGtJo3N](https://www.webcitation.org/6cVGtJo3N)]
20. Kimura M, Nakayasu K, Ohshima Y, Fujita N, Nakashima N, Jozaki H, et al. SS-MIX: a ministry project to promote standardized healthcare information exchange. *Methods Inf Med* 2011;50:131-139. [doi: [10.3414/ME10-01-0015](https://doi.org/10.3414/ME10-01-0015)] [Medline: [21206962](https://pubmed.ncbi.nlm.nih.gov/21206962/)]
21. Japan Association for Medical Informatics. SS-MIX2 Standardized Storage Explanation of Structure and Guidelines for Implementation Ver.1.2 URL: <https://www.jami.jp/english/about/index.html> [accessed 2016-03-07] [WebCite Cache ID [6eIRBdzAt](https://www.webcitation.org/6eIRBdzAt)]
22. Medical Information System Development Center. URL: <http://www.medis.or.jp/> [accessed 2016-03-06] [WebCite Cache ID [6cVGURpW](https://www.webcitation.org/6cVGURpW)]
23. Japanese Association of Healthcare Information System Industry. URL: <http://www.jahis.jp/english/> [accessed 2016-03-05] [WebCite Cache ID [6cVGbAzh4](https://www.webcitation.org/6cVGbAzh4)]
24. Consortium for SS-MIX Dissemination and Promotion. URL: <http://www.ss-mix.org/consE/> [accessed 2016-03-06] [WebCite Cache ID [6eIYZwBmC](https://www.webcitation.org/6eIYZwBmC)]

25. Kimura M, Kanno T, Tani S, Satomura Y. Standardizations of clinical laboratory examinations in Japan. *Int J Med Inform* 1998;48:239-246. [Medline: [9600425](#)]
26. Bhattacharyya L, Cecil T, Dabbah R, Roll D, Schuber S, Sheinin EB, USP Council of Experts Executive Committee. The value of USP public standards for therapeutic products. *Pharm Res* 2004;21(10):1725-1731. [Medline: [15553215](#)]
27. HAPI - The Open Source HL7 API for Java. URL: <http://hl7api.sourceforge.net/> [accessed 2016-03-07] [WebCite Cache ID [6cVGx7kTz](#)]
28. Prasser F, Kohlmayer F, Kemper A, Kuhn K. A generic transformation of HL7 messages into the resource description framework data model. 2012 Presented at: Proceedings of GI-Jahrestagung; 2012; Germany p. 1559-1564.
29. Breitling F. A standard transformation from XML to RDF via XSLT. *Astron Nachr* 2009;330:755-760. [doi: [10.1002/asna.200811233](#)]
30. W3C Interest Group Note. Cool URIs for the Semantic Web URL: <http://www.w3.org/TR/2008/NOTE-cooluris-20080331/> [accessed 2016-03-05] [WebCite Cache ID [6cVH0n4zB](#)]
31. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2015:D1075-D1079 [FREE Full text] [doi: [10.1093/nar/gkv1075](#)] [Medline: [26481350](#)]
32. STITCH: Chemical-Protein Interactions. URL: <http://stitch.embl.de/> [accessed 2016-03-07] [WebCite Cache ID [6eOXG4IqU](#)]
33. Linked Drug Data. URL: <https://github.com/linked-drug-data/publish> [accessed 2016-03-07] [WebCite Cache ID [6fpyluNOV](#)]
34. Stahl SM, Mignon L, Meyer JM. Which comes first: atypical antipsychotic treatment or cardiometabolic risk? *Acta Psychiatr Scand* 2009;119:171-179. [doi: [10.1111/j.1600-0447.2008.01334.x](#)] [Medline: [19178394](#)]
35. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144-151 [FREE Full text] [doi: [10.1136/amiajn1-2011-000681](#)] [Medline: [22733976](#)]
36. Jiang G, Liu H, Solbrig HR, Chute CG. ADEpedia 2.0: Integration of Normalized Adverse Drug Events (ADEs) Knowledge from the UMLS. *AMIA Jt Summits Transl Sci Proc* 2013;2013:100-104 [FREE Full text] [Medline: [24303245](#)]

Abbreviations

- 5HT2C**: serotonin 2C
- ADE**: adverse drug event
- ATC**: Anatomical Therapeutic Chemical Classification System
- EMR**: electronic medical record
- H1**: histamine antagonist of the H1 receptor
- HAPI**: Health Level Seven application programming interface
- HL7**: Health Level Seven
- ICD10**: International Classification of Diseases, and 10th Revision
- KEGG**: Kyoto Encyclopedia of Genes and Genomes
- MEDIS-DC**: Medical Information System Development Center
- MSH**: message header
- ORC**: order-related information
- PID**: patient identification
- RAM**: random access memory
- RDF**: Resource Description Framework
- RDFS**: Resource Description Framework Schema
- RXE**: pharmacy
- SIDER 2**: SIdE Effect Resource
- SPARQL**: SPARQL Protocol and RDF Query Language
- SS-MIX2**: Standardized Structured Medical Record Information Exchange version 2
- TQ1**: timing and quantity
- URI**: uniform resource identifier
- USP**: United States Pharmacopeial Convention Classification System
- XML**: Extensible Markup Language.