Information Discovery on Electronic Medical Records

Vagelis Hristidis, FIU
Fernando Farfán, FIU
Redmond P. Burke, MD Miami Children’s Hospital
Anthony F. Rossi, MD Miami Children’s Hospital
Jeffrey A. White, Teges Corporation

FLORIDA INTERNATIONAL UNIVERSITY
ELECTRONIC MEDICAL RECORDS (EMRs)

- Adoption of EMRs hard due to political reasons
  - No unique patient id
  - Confidentiality
  - HIPAA (Health Insurance Portability and Accountability Act)
- Move towards XML-based format.
- One of most promising: Health Level 7’s Clinical Document Architecture (CDA).
- EMRs pose new challenges for Computer Scientists
  - Confidentiality, authentication, secure exchange
  - Storage, Scalability
  - Dictionaries, terms disambiguation
  - Search for interesting patterns (Data Mining)
  - Data Integration, Schema mapping
  - Information Discovery
<component>
  <section>
    <code code="10160-0" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
    <title>Medications</title>
    <entry>
      <Observation>
        <code code="84100007" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" displayName="history taking (medication)"/>
        <value xsi:type="CD" code="195867001" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" displayName="Asthma"/>
        <originalText><reference value="ml"/></originalText>
      </value></Observation></entry>
    <entry>
      <Observation>
        <code code="84100007" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" displayName="history taking (medication)"/>
        <value xsi:type="CD" code="32398004" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" displayName="Bronchitis"/>
        <value xsi:type="CD" code="91143003" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" displayName="Albuterol"/>
      </value></Observation></entry>
    <entry>
      <SubstanceAdministration>
        <text><content ID="mi"">Theophylline</content> 20 mg every other day, alternating with 18 mg every other day. Stop if temperature is above 103F.</text>
        <consumable>
          <manufacturedProduct>
            <manufacturedLabeledDrug>
              <code code="66493003" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" displayName="Theophylline"/>
            </manufacturedLabeledDrug></manufacturedProduct></consumable>
          </manufacturedProduct></SubstanceAdministration></entry></section></component>
NEED FOR INFORMATION DISCOVERY ON EMRs

- EMRs become widespread
- Allow practitioners, healthcare stakeholders to locate relevant pieces of information
- E.g., a doctor checks possible conflicts between two drugs. Query: “drug-A drug-B death”
- “Google” on EMRs
LIMITATIONS OF Traditional IR

- Text-based search engines do not exploit the XML tags, hierarchical structure of XML.
- Whole XML document treated as single unit - unacceptable given the possibly large sizes of XML documents.
- Proximity in XML can also be measured in terms of containment edges.

LIMITATIONS OF General XML Search

- EMRs have known but complex, domain-specific semantics.
- General semantic rules (e.g., as in XSEarch) not always apply to EMR documents.
- Routine references in EMRs to external information sources like dictionaries and ontologies.
CHALLENGES OF INFORMATION DISCOVERY ON CDA DOCUMENTS
Challenge #1: Structure and Scope of Results

- HTML: Whole document as result
- CDA (XML): What is meaningful?
- Subtree as Result
- Path as Result
- Occasionally result schema specified by physician
- Scope: result across EMRs?
Challenge #2: Minimal Information Unit

- Granularity of a piece of information
  - Self-contained
  - Specific
- Issues related to personalization
Challenge #3: Semantics of Node and Edge Types

- Challenging to incorporate rich semantic information in the ranking model.
- Exploit CDA Object Model.
- Which result below is better?

(a) Path connecting "Substance" in line 49 and "Theophylline" in line 50
(b) Path connecting "Substance" in line 49 and "Theophylline" in line 54
Challenge #4: Access to Dictionaries and Ontologies

- References to external dictionaries and ontologies (LOINC, SNOMED).
- View Data Graph + Ontology Graph as merged graph.
- Keyword search on data graphs with authority flow semantics
- Performance: Size of Ontology Graph

• Query: “Asthma”.
• Documents D1, D2 refer to 79688008 and 955009 respectively.
• D1 or D2 is better?
Challenge #5: Different Types of Relationships in Ontology

- SNOMED has four different classes of relationships (Definitions, Qualifications, Historical, Other)
- Direction of edges:
  - IS-A
  - INVERSE-IS-A
Challenge #6: Arbitrary Levels of Nesting

- E.g., CDA document may contain path component/section/component/section/…
- Conflicts with XSearch rule of no repeating of same tag in result path.
**Challenge #7: Free Text Embedded in CDA Documents**

- Incorporate traditional text-based IR techniques.
- Combination of XML and IR ranking (as in [IR-style keyword search, VLDB03])

```xml
50 <text>
  <content ID="ml">Theophylline</content>
  20 mg every other day, alternating with 18 mg
every other day. Stop if temperature is above 103F.
</text>
```

(a) Free text occurrence of keywords on query "Temperature"

```xml
69 <th>Temperature</th>
```

(b) Embedded HTML fragment is the result of query "Temperature"
Challenge #8: Time and Location Attributes

- Time and Location are critical attributes in most queries for clinical domain.
- Time and Location can be used in metrics similar to idf.
  - Patient with asthma during Winter should be ranked higher for query ”Asthma”.
  - Patient who has the flu in town where no one else has it should be ranked higher for query “flu”
- Time, Location are also important in defining result’s schema
Challenge #9: EMR Document-as-Query

- Using a whole (or part of) EMR as the query, instead of a keyword query.
- Find similar CDA documents, based on:
  - History
  - Demographics
  - Medications/Vital Signs Sequences
  - Treatments, etc…
- Find PubMed articles relevant to an EMR.
Challenge #10: Handle Negative Statements

- Negative statements as important as positive ones in medicine.
- An EMR commonly lists the diagnoses that have been ruled out.
- E. g. "The patient does not have hypertension or diabetes".
- New ontological relationships to express negative findings.
Challenge #11: Personalization

- Different doctors rank entities differently.
- Ontology relationships may be viewed differently.
- Depending on the user role, edges and nodes may have different weight.
  - Medication more relevant than doctor’s name for a researcher.
CONCLUDING REMARKS

- Introduced the problem of Information Discovery on EMR’s.
- Challenges related to architecture, hierarchical structure of document, use of ontologies and dictionaries, special attributes like time and location.
THANK YOU - QUESTIONS?