Toward an automatic definition of the oncologic EHR data elements from NCIT

Annabel Bourdé¹, Marc Cuggia¹, Bruno Turlin², Sébastien Vincendeau², Valérie Bertaud¹, Catherine Bohec³, Régis Duvaufrier²

¹UMR 936 Inserm, Faculté de médecine de Rennes. France  
²CHU Pontchaillou, Rennes, France  
³Réseau ONCOBRETAGNE, Rennes -France
Introduction

• Defining a data model is a bottom-up process in two steps

STEP 1
Definition by Experts

STEP 2
Formalization into DATA ELEMENTS using Information models: HL7/OpenEHR
Reference terminology/ontology: Snomed, ICD, NCIT
Introduction

• **Key Idea : Top down approach**
  - From a domain ontology, extract automatically, from few **key concepts**, a **sub-ontology** that makes the definition of data elements and their value sets easier.

• **Related work :**
  - **LexValueSets**
  - Use only SNOMED CT
    ➔ Need to develop a generic method

• Tested on prostate cancer domain
• compared to oncologic electronic record (gold standard)
Methods : Extraction algorithm

Sub ontology extractor :

1. Input:
   - an ontology in OWL format (NCIT)
   - a list of key concepts from which extraction starts
   - The relationships to be followed
     • Subsumption : toward ascendants, toward descendants
     • Roles to be followed (i.e: has_finding).

2. Search for semantically related concepts

3. Put them in a new subontology

4. Process stops when roots and leaf concepts are reached

5. Output : sub-ontology domain centred
Prostate_Gland

Disease_Has_Associated_Anatomic_Site

Prostate_Neoplasm

Malignant_Prostate_Neoplasm

Prostate_Neuroendocrine_Neoplasm

Prostate_Lymphoma

Prostate_Myeloid_Sarcoma

Prostate_Carcinoid_Tumor

Prostate_Disorder

Male_Reproductive_System_Neoplasm

Male_Reproductive_System_Disorder

Reproductive_System_Neoplasm

Organ_System

Exocrine_Gland

Exocrine_System

Anatomic_Structure_Is_Physical_Part_Of

Methods : Extraction algorithm (2)
Methods : Subontology extraction

Automatic extraction of 2 subontologies :

• ➔ Subontology 1
  – Starting from one single key concept « Prostate_Neoplasm »

• ➔ Subontology 2
  – Starting from the concept « Prostate_Neoplasm »
  – and 5 key concepts given by experts
    • Prostate_Adenocarcinoma
    • Prostate_Cancer_TNM_Finding
    • Biopsy_of_Prostate
    • PSA_Assay
    • Total_Gleason_Score_for_Prostate_Cancer
Method: evaluation
Method: evaluation
Method: evaluation

Oncologic EHR

Manually encoded MDMRF concepts

Exact matching with A
Method: evaluation

Extracted Sub-ontology

- Manually encoded MDMRF concepts
- Exact matching with A
- Pertinent concepts for the MDMRF
- Pertinent concepts for an Oncologic EHR
- Non pertinent concepts for MDMRF and Oncologic EHR
Method: evaluation

- Manually encoded MDMRF concepts (A)
- Exact matching with A (B)
- Pertinent concepts for the MDMRF (B')
- Pertinent concepts for an Oncologic EHR (C)
- Non pertinent concepts for MDMRF and Oncologic EHR (D)
Method: evaluation

Oncologic EHR

MDMRF

A

Manually encoded MDMRF concepts

B

Exact matching with A

B'

Pertinent concepts for the MDMRF

C

Pertinent concepts for an Oncologic EHR

D

Non pertinent concepts for MDMRF and Oncologic EHR
Results

NCIT ontology > 83 000 concepts.

Sub-ontology 1
Extracted from 1 key concept
Contains 434 concepts
A : 82
B : 16
Recall B/A = 19,5%
Precision B/434=4%

Sub-ontology 2
Extracted from 1+5 concepts
Contains 483 concepts
A = 82
B = 42
Recall B/A = 51%
Precision B/483 = 9%

B’= 42+98=140
C = 34
D = 309
Results

Set B' : 42+98 concepts

recurrent_prostate_neoplasms instead of recurrent_disease

Missing concepts: Prostate_Adenosquamous_Carcinoma

Sub-ontology 2
Extracted from 1+5 concepts
Contains 483 concepts
A = 82
B = 42
Recall B/A = 51%
Precision B/483 = 9%

B' = 42+98=140
C = 34
D = 309
Results

SET C : 34 concepts

*Bone_Pain* or *Urinary_Retention*

Sub-ontology 2
Extracted from 1+5 concepts
Contain 483 concepts
A = 82
B = 42
Recall B/A = 51%
Precision B/483 = 9%

B’ = 42 + 98 = 140
C = 34
D = 309

Oncologic EHR
Results

SET D : 309
concepts structuring the sub-ontology

Disorder_by_Site

Sub-ontology 2
Extracted from 1+5 concepts
Contain 483 concepts
A = 82
B = 42
Recall B/A = 51%
Precision B/483 = 9%
B’= 42+98=140
C = 34
D = 309
Results

Set A minus set B (49%):
Medical history: Ischemic_Heart_Disease
Therapy: Adjuvant_Therapy

Sub-ontology 2
Extracted from 1+5 concepts
Contain 483 concepts
A = 82
B = 42
Recall B/A = 51%
Precision B/483 = 9%

B' = 42+98=140
C = 34
D = 309
Discussion and conclusion

• Limit: one single neoplasm type was tested.

• Most of the important concepts of the MDMRF were found
  – excepted concepts of medical history (not domain specific)

• MDMRF concepts are very specific and oriented toward therapeutic decision.
  – Better coverage considering a broader scope (oncologic EHR)
Discussion and conclusion

• Huge difference between subtontologies1 and 2:
  • NCIT: concept poorly connected or misclassified concepts
  • Adding new key concepts ➞ better recruitment

• Compared to LexValue Set: Coverage of relevant terms (51%) is higher than LexValueSets project (35%).

• Further works:
  – Improving the extractor performance
  – automatically generating data elements and value sets usable automatically in the EHR
Questions?
Method: evaluation

Step 1: Manually transform the paper version of the MDMRF into an electronic version

Multi Disciplinary Meeting Report Form (MDMRF) Paper version (defined by experts)

Data element definition and mapping to NCIT

Electronic version

MDMRF

age

stagging

Gleason

PSA

...