Comparing structured and unstructured information content in EHR: application to coeliac disease

JB Escudie, B Rance, AS Jannot, A Burgun

Département informatique biomédical et santé publique, Hôpital Européen Georges Pompidou HEGP, AP-HP, Paris, France

INSERM, U-1138, team 22, Information Sciences to support Personalized Medicine

Paris Descartes University, Medical Faculty.

OBJECTIVE: Medical documents stores information not available in any other form. Therefore natural language processing (NLP) for extracting this information is complementary to structured data queries in a clinical data warehouse (CDW). In the specific context of studying phenotypes associated to celiac disease, we developed and evaluated a method to extract a large panel of targeted phenotypes in clinical documents.

METHODS: We identified a cohort of celiac patients (CD) in the i2b2 CDW implemented at French hospital HEGP. All discharges and letters concerning these patients constituted our corpus. On a training subset, we constituted a catalog of regular expressions to identify a large number of phenotypes in free texts through an iterative process. These regular expressions were then mapped to a terminology of phenotypes elaborated for the celiac disease study with experts of the disease and based on literature review. We extracted phenotype information on the whole corpus and a subset was used as evaluation corpus. The evaluation corpus was manually read by a trained physician to extract information on 17 concepts of our terminology to establish our gold standard. We also queried the CDW using structured data for the same phenotypes. Precision, recall and F-measure were computed on a document level and aggregated on a patient level for NLP and structured data extractions.

RESULTS: We identified 440 patients with diagnosed CD and retrieved 1401 documents. After 5 iterations on a total of 50 documents, the catalog had 493 regular expressions. Terminology contained 114 entries and 4 hierarchical levels. Evaluation set was constituted by 62 documents covering 20 patients' medical records. At document level F-measure ranged from 0.43 for anxiety and depression identification (recall and precision of 0.43) to 0.93 for CD specific anti-bodies negative dosage (recall of 1.00, precision of 0.88). At patient level, our NLP method performed better than structured data queries on every phenotype. F-measure ranged from 0.60 for villous atrophy (recall and precision of 0.60 vs 0.00 for structured data) to 1.00 for viral hepatitis (recall and precision of 1.00 vs 0.00 for structured data).

CONCLUSION: Our method successfully extracted a wide range of phenotypes that no current NLP tools based on UMLS or SNOMED could have extracted because of very specific phenotypes absent from these terminologies and because documents were written in French. Future extraction method could combine NLP and structured data to enhance performance.