Ontological Features of Electronic Health Records Reveal Distinct Association Patterns in Liver Cancer

Lawrence WC Chan, SC Cesar Wong
Department of Health Technology and Informatics
Hong Kong Polytechnic University
Hong Kong
wing.chi.chan@polyu.edu.hk

Keith WH Chiu
Department of Diagnostic Radiology
University of Hong Kong
Hong Kong

***Abstract*— Electronic Health Record (EHR) system is not only aimed to provide a digital and structural form of patient records but also support the clinical decision, patient care and patient advice. The EHR database is still an under-explored big data resource that has hosted a large number of cases with complete recovery, good prognosis, reliable diagnostic tests and effective treatments. A set of 112 abdominal computed tomography imaging examination reports, consisting of 59 cases of hepatocellular carcinoma (HCC) or liver metastases (so called HCC group for simplicity) and 53 cases with no abnormality detected (NAD group), was collected from four hospitals in Hong Kong. We extracted terms related to liver cancer from the reports and mapped them to ontological features using Systematized Nomenclature of Medicine (SNOMED) Clinical Terms (CT). Each feature value was further weighted using a systematic PubMed search method. Association levels between every two features in HCC and NAD groups were quantified using Pearson's correlation coefficient. The distribution of association levels in HCC group was compared with that in NAD group. HCC group reveals a distinct association pattern that signifies liver cancer and provides clinical decision support for suspected cases.**

Keywords—Ontology; Electronic Health Record; SNOMED; Hepatocellular Carcinoma; Cancer Signature; Clinical Decision Support

I. INTRODUCTION

Sheer amount of clinical data hosted by the electronic health record (EHR) system facilitates the exploration of disease signatures and potentiates the relevant clinical decision support functions [1,2]. Feature vector model has been developed for converting the clinical texts and image patterns of an EHR into an array of numerical values [3-6].

The support of a medical ontology is required to map textual information, such as image findings in a diagnostic report, to a feature vector [5,6]. Systematized Nomenclature of Medicine (SNOMED) Clinical Terms (CT) is an ontological standard of clinical terms, which are organized as concepts and linked with "is-a" or inverse "is-a" relationships [7-10]. In such hierarchical structure, concepts at a particular level could be chosen as the feature concepts. The semantic distance between a clinical term in EHR and a feature concept can be quantified by counting the edges along the path connecting them in the "is-a" hierarchy [3-5,11,12]. Aggregating all the semantic distances to the feature concepts generates an ontological feature vector that characterizes an EHR with its disease context. We hypothesize that the feature association patterns derived from the EHRs can uniquely distinguish a disease group from the non-disease group. The identified patterns can be used to develop a clinical decision support function for new clinical cases.

II. METHODS

A. Data Collection

We collected retrospectively 112 image reports of abdominal computed tomography examinations from the Radiology Departments of four local hospitals in Hong Kong. HCC or liver metastases were found in 59 cases (called HCC group for simplicity) and the other 53 cases had no abnormality detected (NAD group). Before the data collection, third party clinical personnel has removed the patient name, identity card number, telephone number and address from the reports and assigned a randomly generated unique ID to each case. We have obtained Human Subject Ethics Approval from the Hong Kong Polytechnic University (HSEARS20140710002).

B. Ontological Feature Extraction

The HCC-related clinical terms were extracted manually from the image reports according to SNOMED CT curated in the Unified Medical Language System (UMLS; license code: NLM-0315126310). UMLS organizes clinical terms in concepts and SNOMED CT defines the relationship between concepts using the "is-a" hierarchical tree. The extracted terms were projected to the feature concepts at a particular level to ensure consistent comparison between reports. Due to the optimal classification granularity, level-4 concepts were considered as feature concepts in this work [12].

For each report, a feature vector, given by $[a_1, a_2, \ldots, a_m]$, was generated using edge counting approach and the vector element a_i is given by the following formula.

$$a_i = \frac{\sqrt{p_i}}{1+\min_{j=1\ldots n} s_{ij}} \quad (1)$$

where p_i is the conditional probability of the i^{th} feature concept given the occurrence of liver cancer and s_{ij} represents the edge count between the i^{th} feature concept and the j^{th} clinical term extracted from a report. With the value between 0 and 1, a_i indicates the relevance between the i^{th} feature concept and a clinical term in a report. Such relevance can be modulated by the conditional probability, p_i, which is estimated by the specific term weighting approach [13].

C. Ontological Association Patterns

The association level between two feature concepts was denoted by $C_d(i, j)$ for the HCC group, and $C_n(i, j)$ for the NAD group, as given by the following formulas.

$$C_d(i,j) = | r(x_{di}, x_{dj}) | \quad (2)$$

$$C_n(i,j) = | r(x_{ni}, x_{nj}) | \quad (3)$$

where x_{di} and x_{dj} represent the numerical array weighting the i^{th} and j^{th} feature concepts across the HCC group; x_{ni} and x_{nj} represent the numerical array weighting the i^{th} and j^{th} feature concepts across the NAD group; $r(x_i,x_j)$ is Pearson correlation coefficient between two arrays. Two sets of correlation coefficients, C_d and C_n, in the HCC and NAD groups formed two cumulative distributions, F_d and F_n, which were compared using two-sample Kolmogorov-Smirnov (KS). To test the significance of difference, the maximum deviation between two cumulative distributions, D value, was compared with its critical value, D_α, which is derived based on our developed method [14] and given by following equations. A correlation threshold, at which F_d and F_n were extremely deviated can be identified and used to characterize the perturbed ontological association pattern.

$$D = \max_C | F_d(C) - F_n(C) | \quad (4)$$

$$F_d(C) = Prob(C_d \le C) \quad (5)$$

$$F_n(C) = Prob(C_n \le C) \quad (6)$$

$$D_\alpha = \gamma(\alpha)\sqrt{\frac{4}{k(k-1)}} \quad (7)$$

where α is the significance level, i.e. 0.05, $\gamma(0.05)$=3.1 and k=30 in this study. The critical value of D is 0.2102.

III. Results

A. Extracted Features

From 59 and 53 image reports of respective HCC and NAD groups, 38 clinical terms were extracted and mapped to 38 unique concepts in UMLS. These terms were then projected to 30 feature concepts at level-4 of SNOMED CT "is-a" hierarchy (Table I). After counting the edges and estimating the conditional probabilities of these concepts, their weightings were calculated and formed 30×59 and 30×53 matrices for HCC and NAD groups.

B. Ontological Association Patterns

The association level between every two feature concepts was calculated. We generated 435 association levels for each of HCC and NAD groups. Figure 1 shows the cumulative distributions of association levels for the two groups and their difference. The maximum deviation, D = 0.333, was found at C=0.03 and greater than its critical value. Therefore, the two ontological association patterns are significantly different.

TABLE I. Feature Concepts

Abdominal organ finding	Growth alteration
Blood vessel finding	Imaging result abnormal (Imaging Procedure)
Disorder of body cavity	Mechanical abnormality
Disorder of body system	Finding of biliary tract
Disorder of cardiovascular system	Hemorrhage into peritoneal cavity
Disorder of digestive system	Disorder of connective tissue
Disorder of soft tissue	Degenerative abnormality
Disorder of trunk	Traumatic and/or non-traumatic injury of anatomical site
Finding of trunk structure	Abnormal radiologic density, diffuse
Liver finding	Imaging result abnormal (Evaluation Procedure)
Radiologic finding	Abnormal radiologic density, irregular
Cyst of abdomen	Abnormal radiologic density, nodular
Mass of body region	Abnormal radiologic density, small area
Mass of digestive structure	Multiple lesions
Neoplastic disease	Finding of number of lesions

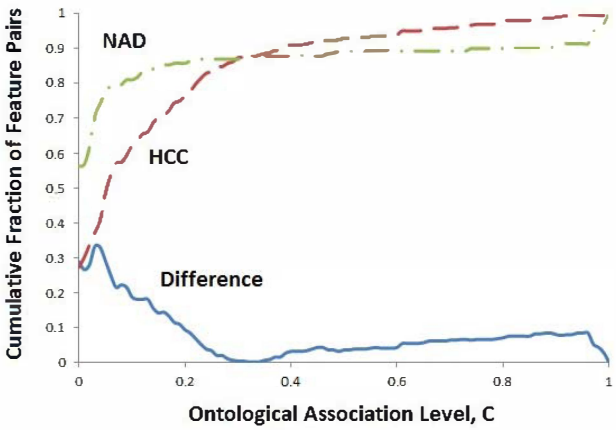

Fig. 1. Two distinct ontological association patterns. Cumulative distributions of ontological association levels across NAD and HCC groups are indicated by dash-dotted and dash lines respectively. Solid line represents the difference between these two cumulative distributions.

IV. Discussion

This study illustrated an approach for characterizing textual image reports by numerical values that weight the alignment of report contents with the ontological standard. The mapped feature concepts exhibited distinct association patterns, which are significantly different between HCC and NAD groups. The concept pairs connected in the association patterns can be considered as a signature of liver cancer. For suspected cases, this signature can be used to assist the clinical decision when associations of those pairs are observed. It is worth noting that the discovered signature should be validated with independent data before its clinical applications.

An alternative application of the identified association patterns is the detection of inaccurate medical coding. When a disease is diagnosed, the "co-activated" feature concepts can be obtained and checked against the pairs in the disease-specific patterns. Potential inaccurate coding can be detected and the clinicians will be alerted. On a public health level, systematic failure in appropriate medical coding may result in over and under or over adjustment to case-mix measurements when assessing quality of care [15]. In some healthcare models, this will also affect billing, reimbursement and insurance claims [16].

REFERENCES

[1] Peter BJ, Lars JJ, Søren B: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 2012, 13(6):395-405

[2] Ceuster W, Smith B: Strategies for referent tracking in electronic health records. Journal of Biomedical Informatics 2006, 39:362-378

[3] Chan LWC, Benzie IFF, Liu Y, et al.: Is the inter-patient coincidence of a subclinical disorder related to EHR similarity? 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services 2011:177-180

[4] Sánchez D, Batet M, Isern D, et al.: Ontology-based semantic similarity: A new feature-based approach. Expert Systems With Applications 2012, 39(9):7718-7728

[5] Batet M, Sánchez D, Aida V: An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 2011, 44:118-125

[6] Richesson RL, Andrew JE, Krischer JP, et al.: Use of SNOMD CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. Journal of the American Medical Informatics Association 2006, 13(5):536-546

[7] Melton GB, Parsons S, Morrison FP, et al.: Inter-patient distance metrics using SNOMED CT defining relationships. Journal of Biomedical Informatics 2006, 39(6):697-705

[8] Pedersen T, Pakhomov SVS, Patwardhan S, et al,: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 2007, 40(3):288-299

[9] Wasserman H, Wang J: An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Symposium 2003:699-703

[10] Lieberman MI, Ricciardi TN, Masarie FE, et al.: The use of SNOMED CT simplifies querying of a clinical data warehouse. AMIA Symposium 2003:910

[11] Lord PW, Stevens RD, Brass A, et al.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003, 19(10):1275-1283

[12] Chan LWC, Liu Y, Shyu CR, et al.: A SNOMED supported ontological vector model for subclinical disorder detection using EHR similarity. Engineering Applications of Artificial Intelligence 2011, 24:1398-1409

[13] Chan LWC, Liu Y, Chan T, Law HKW, Wong SCC, Yeung APH, Lo KF, Yeung SW, Kwok KY, Chan WYL, Lau TYH, Shyu CR. PubMed-supported Clinical Term Weighting Approach for Improving Inter-Patient Similarity Measure in Diagnosis Prediction. BMC Medical Informatics and Decision Making 2015, 15:43. (doi: 10.1186/s12911-015-0166-2)

[14] Chan LWC, Lin X, Yung G, Lui T, Chiu YM, Wang F, Tsui NBY, Cho WCS, Yip SP, Siu PM, Wong SCC, Yung BYM. Novel structural co-expression analysis linking the NPM1-associated ribosomal biogenesis network to chronic myelogenous leukemia. Scientific Reports 2015, 5 (10973). (doi: 10.1038/srep10973)

[15] Case-mix measurement and assessing quality of hospital care. Health Care Financ Rev. 1987(Suppl):39-48.

[16] Slight M, Belley SE, Shrader W, Sr., Endicott M. ICD-10: Targeting reporting and reimbursement. Health Manag Technol. 2016, 37(3):8-11.