Similarity Computation based on Formal Concept Analysis for Colorectal Cancer Patients

Jing Xiang[1,2*]
1.Department of Biostatistics
School of Public Health
Shandong University
Jinan,China
Email:xiang_jing@163.com

Hanbing Xu[*]
2.School of Public Health
and Management
Binzhou Medical University
Yantai,China
Email: ice_xu0109@163.com

Suresh Pokharel
School of Information Technology and
Electrical Engineering
The University of Queensland
Brisbane, Australia
email: s.pokharel@uq.edu.au

Jiqing Li
Department of Biostatistics
School of Public Health
Shandong University
Jinan,China
Email:15554136938@163.com

Fuzhong Xue[△]
Department of Biostatistics
School of Public Health
Shandong University
Jinan,China
Email: xuefzh@sdu.edu.cn

Ping Zhang[△]
Menzies Health Institute
Griffith University
Gold Coast, Australia
Email: p.zhang@griffith.edu.au

***Abstract*— Colorectal cancer is a heterogeneous disease. Its response to targeted therapies is associated with various factors, and the treatment effect differ significantly between individuals. Personalize medical treatment (PMT), which takes into consideration of individual patient characteristics, is the most effective way to deal with this issue. Patient similarity and clustering analysis is an important part in PMT. Earlier works mainly focused on similarity computation among the patients but overlook to preserve relationships. This paper presents a formal concept analysis-based approach for computing the similarity between colorectal cancer patients. The approach not only does the clustering of patients based on their similarity but also can preserve the relations between clusters in hierarchical structural form. This would allow us to build a knowledge base which is helpful for clinicians to take fast and effective decision for treatment and care of colorectal cancer patient.**

Keywords—colorectal cancer; patient similarity; formal concept analysis

I. INTRODUCTION

Colorectal cancer (CRC) is a common malignant, immunogenic, and immune targeted tumor disease in the digestive tract. It is the second most common cancer in women and the third in men all around the world which causes more than 500000 death every year [1,2]. Chemotherapy, radiotherapy and surgery are the common treatments for colorectal cancer. Individual heterogeneity of sensitivity and toxicity within and between tumors makes treatment effect differ significantly, even with the same treatment on the same site of the tumor. Personalize medical treatment (PMT) is one of the effective solutions for this problem. PMT is the tailoring of treatments based on individual patient characteristic (biological features and environmental factors). In the case of CRC, the aim of PMT is to have effective treatment by avoiding potential adverse effects or delaying in seeking better alternative treatments [3].

Cancer staging is the process of finding out how much cancer is in a person's body and where it's located. Along with the type of cancer a person has, the stage of the cancer is one of the most important factors for the doctors to determine a patient's prognosis and treatment. A tumor staging system, named TNM [4], is mostly used for prognosis of colorectal cancer (CRC). In the TNM system, the overall stage is determined after the cancer is assigned a letter or number categories, to describe the original tumor (T), the spread to the lymph node (N), and whether the cancer has spread/metastasized (M). Doctors assign the stages (eg. Stage I, IIA, IIB etc) of the cancer by combining the T, N, and M classifications to help for patient treatment planning. Pathological staging of CRC has important clinical significance for choosing proper patient treatments and corresponding follow up plans.

However, even patients with the same stage and pathological type can have different responses to the same radial or chemotherapy regimen, which indicates the individual differences among the patients. Similarity analysis of colorectal cancer patients based on their characteristics may help categorize patients to different groups and recommend the relevant treatment scheme towards personalized treatment [5]. One of the fundamental challenges for PMT is to group patients in significant subsets (small cluster of patients) based on their similarity in addition with capturing the relationships between them. The key of creating meaningful patient clusters is to capture the right patient characteristics (features) and to apply the suitable algorithms. Two important tasks in the process are (i) similarity computation: it measures the

This work was supported by China Postdoctoral Science Foundation (No.2017M612295) and China Statistical Science Foundation (No.2020LY079).

* These authors contributed equally to this work.
△ Corresponding author.

similarity between the CRC patients based on their characteristics such as cancer stage, vital signs, symptoms and other clinical signs. (ii) capturing relationships: the relationships can be captured in patient level as well as within the subsets level based on their common characteristics. This enable us to build a knowledge base which helps make clinical applications.

In the past, many methods have been proposed for patient similarity computation, utilizing cancer staging levels and other prognostic factors. Huang et al. [6] proposed a patient similarity measurement to build predictive models for diabetes status. The similarity calculation method was based on Euclidean distance and Jaccard distance. Pai et al. [7] proposed a novel supervised patient classification framework based on patient similarity networks--netDx, the similarity metric used was Euclidean distance and Pearson correlation. Similarly, Pokharel et al. [8] proposed ontology-based method for calculating patient similarity for intensive care unit (ICU) patients. Later, Pokharel et al. [9, 10] proposed temporal tree with sequential pattern mining to capture inherent relationships between the clinical events due their co-occurrence. However, all of the above patient similarity-based method only focuses on computing the similarity between the patient but don't consider to preserve the relationships into the patient level and within the subsets level.

To address the above listed problems, we proposed the formal concept analysis (FCA) based similarity computation approach. FCA expresses knowledge bases in a hierarchical structure, which not only reflects the hierarchical structure but also the relations between concepts or clustered groups [11]. Each concept in the hierarchy represents the objects sharing some set of properties, and each sub-concept in the hierarchy represents a subset of the objects (as well as a superset of the properties) in the concepts above it. This research aims to apply FCA for colorectal cancer patients to build a knowledge base and to utilize the concept lattices as references for individual patient treatment planning.

II. Materials and Methods

A. Data Sets

The data used for this study were collected from hosptal information system (HIS) of Shandong provincial hospital, China from November 2010 and July 2016 [12]. The main inclusion criteria for selecting the patient cohort are follows: (i) diagnosed as CRC according to histology or cytology, (ii) at least 18 years old, (iii) had not received neoadjuvant therapy (iv) had radical surgery, (v) available for follow-up data, (vi) preoperative biochemical test data can be obtained. Also, the following are the exclusion criteria: (i) emergency and untreatable due to widespread metastasis, (ii) combined with other cancers and (iii) receiving preoperative chemotherapy or immunotherapy. With the inclusion and exclusion criteria, we get 2442 eligible patients which are diagnosed with CRC.

Besides cancer stages, patient age, cancer degree of differentiation, histological type, number of sample lymph nodes, carcinoembryonic antigen (CEA), cancer antigen 19-9 (CA19-9), lymph node ratio (LNR) and lymph vascular invasion have been reported as prognostic factors of cancer patients [12-16]. All these values were recorded for each patient in the collected data set, and were used for this study to build a knowledge base which can be used as references of choosing the most appropriate treatment or care plans for CRC patients. Seven cancer stage groups were formed based on the TNM system with the recorded original tumor (T), lymph node (N) and metastasized (M) classifications. They are named as stage I, IIA, IIB, IIC, IIIA, IIIB and IIIC, representing different levels of cancer spreading into nearby tissues or lymph nodes [17].

The study was approved by the Ethics Committee of School of Public Health, Shandong University and conducted in accordance with the ethical guidelines of the Declaration of Helsinki of the World Medical Association

B. Formal concept analysis

FCA as a term was first introduced by Wille in 1981 [18]. It provides a conceptual framework for structuring, analyzing and visualizing data, to make them more understandable [11]. The method defines a concept as a unit of thought comprising of a set of objects and a set of their shared attributes. It finds all concepts and their dependencies from the tabular input data which is defined as a formal context. Two sets of output data are resulted from the analysis performed on the formal context. The first set gives a hierarchical relationship of all the established concepts in the form of line diagram called a concept lattice. In the concept lattice, each concept is represented as a node, called concept node. The second set gives a list of all found interdependencies among attributes in the formal context [19]. In FCA, every concept in the concept lattice (a node in the line diagram) consists of two parts: connotation (attribute set) and extension (object set). Concept lattice is to explain the relationship between connotation and extension, and it is the unity of concept connotation and extension.

1）Formal context

A formal context can be represented as a triple $K=(G,M,I)$, where G and M are two sets of elements, which are objects and attributes respectively. I is the binary relationship between G and M. In order to express the relationship I between object g and attribute m, we can write it as gIm or $(g,m)\in I$ and interpret it as "object g has attribute m". In formal context $K=(G,M,I)$, given a set of object subsets $A \subseteq G$ and a set of attribute subsets $B \subseteq M$, then a set of dual operators A'and B' are defined as follows:

$$A' = \{m \in M \mid \forall g \in A, gIm\}$$
$$B' = \{g \in G \mid \forall m \in B, gIm\} \quad (1)$$

Where A' is a collection of all the attributes that are shared by all objects in A. B' is a collection of all objects, all of which have all the attributes in B. If A and B satisfy $A'= B$, $B' = A$, then (A, B) is called a concept of formal context, where A is the extension of the concept and B is the connotation of the concept [18].

In this study with the colorectal cancer data, each patient is an object, and the corresponding features of patients are attributes. We establish formal context according to the

relations between objects and attributes, thus forming a hierarchical structure between concepts.

2) Concept similarity in FCA

Concept similarity can be measured by the distances between the concepts in the hierarchy of concept lattice. At present, there are many methods to calculate concept similarity, such as concept similarity based on concept instances, concept similarity based on attributes and concept similarity based on concept relations. This paper mainly introduces the concept similarity based on attributes, which is used for this study

Concept similarity indicates the degree of that two concepts share the same attributes. If two entity concepts have more identical attributes, the higher the concept similarity. The similarity calculation of objects and attributes between concept nodes can be measured by distance. Larger distance between two concept nodes indicates less number of the same objects and attributes between the two concept nodes, that is, the concept similarity is lower. Concept similarity between two concept nodes (A_1, B_1 and A_2, B_2) can be calculated as follows:

$$\begin{aligned}&\mathrm{Sim}((A_1,B_1),(A_2,B_2))\\&=\left(\frac{|A_1\cap A_2|}{m}\right)*\alpha+\left(\frac{|B_1\cap B_2|}{n}\right)*\beta\end{aligned}\qquad(2)$$

Where m=max ($|A_1|$, $|A_2|$), n=max ($|B_1|$, $|B_2|$). According to the duality principle of concept lattice, the objects and attributes of concept nodes have the same status, so $\alpha=\beta=0.5$.

Tadrat et al. [20] proposed a similarity measure in formal concept analysis for case-based reasoning. They used the hierarchical structure of FCA to build the case database and proposed a new concept similarity method. Suppose $C_P = (E_P, I_P)$ is a concept in the formal context (*G, M, I)*, where E_P represents the object set of a formal concept case and I_P represents the attribute set of this source concept Given a $C_N = (E_N, I_N)$ is the concept of the target concept, E_N and I_N represent the object set and attribute set of the target concept respectively.

The similarity of C_P and C_N is calculated as follows:

$$\begin{aligned}&\mathrm{Sim}(C_P,N_P)\\&=\frac{1}{2}\left(\frac{\sum_{u\in I_N\cap I_P}(\log\frac{N}{Fa_u})^2}{\left[\sum_{i\in I_N}(\log\frac{N}{Fa_i})^2\sum_{j\in I_P}(\log\frac{N}{Fa_j})^2\right]^{1/2}}+\left(\frac{\sum_{v\in E_N\cap E_P}(\log\frac{N}{Fc_v})^2}{\left[\sum_{k\in E_N}(\log\frac{N}{Fc_k})^2\sum_{i\in E_P}(\log\frac{N}{Fc_i})^2\right]^{1/2}}\right)\right)\end{aligned}\qquad(3)$$

where N is a total number of formal concepts, Fa_u, Fa_i and Fa_j are a frequency of attributes u, i and j, respectively, $\{u,i,j\}\in M$, and Fc_v,Fc_k and Fc_l are frequencies of cases v, k and l, respectively, $\{v,k,l\}\in G$.

TABLE I BASELINE STATISTICAL DESCRIPTION OF PATIENTS WITH CRC

Characteristic	N (%)
Age (years)	
<60	1108 (45.4)
60-70	824 (33.7)
≥70	510 (20.9)
Differentiation	
moderate	425 (17.4)
poor	1878 (76.9)
well	139 (5.7)
Histological type	
Mucinous adenocarcinoma	327 (13.4)
Adenocarcinoma	2115 (86.6)
Sample lymph nodes	
<10	224 (9.2)
≥10	2218 (90.8)
CEA (μg/ml)	
<5	533 (60.1)
≥5	367 (39.1)
CA19-9 (U/ml)	
<37	2100 (86.0)
≥37	342 (14.0)
Lymphovascular invasion	
Yes	109 (4.5)
No	2333 (95.5)
LNR	
<0.13	1779 (72.9)
≥0.13	663 (27.1)

TABLE II DATA CODING FOR FORMAL CONTEXT CONSTRUCTION

Objects	Attributes								
	age <60	Age >70	Differentiation: Poor	CEA <5	…	IIC	IIIA	IIIB	IIIC
patient1	1	0	0	……	0	1	0	0	0
patient2	0	1	0	……	0	1	0	0	0
patient3	0	0	0	……	0	1	0	0	0
patient4	1	0	0	……	0	1	0	0	0
patient5	1	0	0	……	0	1	0	0	0
patient6	1	0	0	……	0	1	0	0	0
patient7	1	0	1	……	0	1	0	0	0
……	……	……	……	……	……	……	……	……	……
patient2441	1	0	0		0	0	1	0	0
Patient2442	0	0	0		0	0	1	0	0

III. RESULTS

A. Description of demographic characteristics

The samples were demographically described (Table I), in three age groups. Cancer degree of differentiation, histological type, number of sample lymph nodes, carcinoembryonic antigen (CEA), cancer antigen 19-9 (CA19-9), lymph node

ratio (LNR) and lymph vascular invasion are categorized following the clinical instruction.

B. Formal Context of Colorectal Cancer Patients

The formal context of the 2442 patients was constructed with fcaR package in R [21]. Attributes (features) of patients were coded as binary values before being used as the input of the package. Part of the input file is shown in table II. A total of 1155 concepts were obtained from the data set. Table III shows the concept values in the formal context. Each row in the table shows one concept in the context, with its corresponding attributes and the number of objects included in the concept. The concepts at the top of the hierarchy of the concept lattice include more individual patients who share the small sets of attributes, while the concepts at the lower levels contains less patients who share more specific attributes. The lattice with all the patients and attributes constructed can be used as a knowledge base. Patients share the same concept indicates the potential of similar effect with same clinical procedures, for example the treatment applied to.

IV. DISCUSSION

In this study, FCA is introduced for patient similarity analysis which not only presents a clear hierarchy, but also reflects the relations between concepts. The formal context built from this study can be used as the reference knowledge base for choosing alternative treatment plans for the patients who did not receive the effective treatment. For example, patient A and patient B share one concept, knowing patient A had effective treatment and patient B had another treatment procedure but not as effective as patient A had. Doctors can review the treatment plan for patient A and adjust the treatment for patient B accordingly.

In another way, a similarity measure between a new patient to the patients included in the knowledge base can be used to help finding the patients who share the most attributes with the new patient. The similarity between patients or the similarity between a patient and the concept included in the formal context can be calculated in different existing methods for example Euclidean distance or the method proposed by Tadrat et al. [20]. Patients have high similarity or within the set of concepts which are close to each other are recommended to be considered for sharing certain clinical values. Different clustering may be applied for different clinical practices.

V. CONCLUSION

In this study, the concept lattice theory was applied to patient similarity analysis. A concept lattice of CRC patients that was built with extra patient attributes besides the cancer stages can be used as a knowledge base for alternative clinical advices to existing or new patients. Further clustering based on the concepts presented in the conceptual structure can be applied to various clinical applications, corresponding to different risk groupings. It adds values in health services towards personalized cancer care or treatment.

TABLE III CONCEPT LATTICE OF COLORECTAL CANCER PATIENTS

Concept	Attribute set	Number Of Objects
C1	Stage: IIIC	357
C2	Stage: IIIB	530
C3	Stage: IIIA	66
C4	LNR: <0.13	1779
C5	LNR: <0.13 Stage: IIIB	252
C6	LNR: 0.13 Stag: IIIA	38
C7	LNR: 0.13 Stag: IIC	17
C8	LNR: <0.13 Stage: IIB	648
C9	LNR: <0.13 Stage: IIA	396
C10	Lymphovascular invasion: Yes	109
……	……	……
C1151	age<60 Differentiation: poor Histological type:Mucinous Adenocarcinoma CEA<5 ca199<37 LNR<0.13 Stage: IIIA	2
C1152	age<60 Differentiation: poor Histological type:Mucinous Adenocarcinoma ca199<37 Sample_lymph_nodes: <10	3
C1153	age<60 Differentiation: poor Histological type:Mucinous Adenocarcinoma ca199<37 Sample_lymph_nodes: <10 Stage: IIC	1
C1154	Age: <60 Differentiation: poor Histological type: Mucinous Adenocarcinoma ca199: <37 Sample_lymph_nodes: <10 CEA: <5 ca199: <3.7 LNR: <0.13	2
C1155	Age: <60 Differentiation: poor Histological type:Mucinous Adenocarcinoma ca199: <37 Sample_lymph_nodes: <10 CEA: <5 ca199: <3.7 LNR: <0.13 Stage: IIB	1

REFERENCES

[1] R. L. Siegel, Kimberly D. Miller, Stacey A. Fedewa et al., “Colorectal cancer statistics, 2017,” CA: A Cancer Journal for Clinicians, vol . 67, no. 3, pp. 177-193, 2017.

[2] Ferlay J, Soerjomataram I, Dikshit R, et al., Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359-86.

[3] A. Guglielmo, N. Staropoli, M. Giancotti and M. Mauro, “Personalized medicine in colorectal cancer diagnosis and treatment: a systematic review of health economic evaluations”, Cost Effectiveness and Resource Allocation, 2018, 16:2 https://doi.org/10.1186/s12962-018-0085-z.

[4] F. Moccia et al., "Lymph Node Ratio Versus TNM System As Prognostic Factor in Colorectal Cancer Staging. a Single Center Experience," Open Med (Wars), vol. 14, pp. 523-531, 2019.

[5] S. A. Brown, "Patient Similarity: Emerging Concepts in Systems and Precision Medicine," (in English), Frontiers in Physiology, Editorial Material vol. 7, p. 6, Nov 2016, Art. no. 561.

[6] Y. Huang et al., "Study on Patient Similarity Measurement Based on Electronic Medical Records," (in eng), Stud Health Technol Inform, vol. 264, pp. 1484-1485, Aug 21 2019.

[7] S. Pai, S. Hui, R. Isserlin, M. A. Shah, H. Kaka, and G. D. Bader, "netDx: interpretable patient classification using integrated patient similarity networks," Molecular Systems Biology, vol. 15, no. 3, Mar 2019, Art. no. e8497.

[8] Pokharel, S., Li, X., Zhao, X., Adhikari, A., & Li, Y. (2018, January). Similarity Computing on Electronic Health Records. In PACIS (p. 198).

[9] Pokharel, S., Zuccon, G., Li, X., Utomo, C. P., & Li, Y. (2020). Tempora Tree Representation for Similarity Computation between Medical Patients. Artificial Intelligence in Medicine, 101900.

[10] Pokharel, S., Zuccon, G., & Li, Y. (2020). Representing EHRs with Temporal Tree and Sequential Pattern Mining for Similarity Computing, International Conference on Advanced Data Mining and Applications (ADMA), 2020

[11] A. Formica, "Concept similarity in Formal Concept Analysis: An information content approach," Knowledge-Based Systems, vol. 21, no. 1, pp. 80-87, 2008.

[12] J. Li, X. Li, J. Gu, X. Ma, and F. Xue, "A competing-risks nomogram for predicting probability of death from CRC in Chinese Han patients with Stage I-III CRC," Jpn J Clin Oncol, vol. 48, no. 12, pp. 1088-1095, Dec 1 2018.

[13] C. Compton, “Colorectal carcinoma: diagnostic, prognostic, and molecular features”, Mod Pathol 2003;16:376–88.

[14] K. Sun, S. Chen, J. Xu, G. Li, Y. He, “The prognostic significance of the prognostic nutritional index in cancer: a systematic review and meta-analysis”, J Cancer Res Clin Oncol 2014;140:1537–49.

[15] C. Zheng, W. Zhan, J. Zhao et al., “The prognostic value of preoperative serum levels of CEA, CA19-9 and CA72-4 in patients with colorectal cancer”, World J Gastroenterol 2001;7:431–4.

[16] M. May, E. Herrmann E, C. Bolenz et al., “Association between the number of dissected lymph nodes during pelvic lymphadenectomy and cancer-specific survival in patients with lymph node-negative urothelial carcinoma of the bladder undergoing radical cystectomy”, Ann Surg Oncol 2011;18:2018–25.

[17] Cancer.Net, Colorectal Cancer: Stages, https://www.cancer.net/cancer-types/colorectal-cancer/stages, accessed on 6 Nov, 2020.

[18] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts". Proceedings of the NATO Advanced Study Institute held at Banff, Canada, August 28 to September 12, 1981.

[19] F. Škopljanac-Mačina, B. Blašković, Formal Concept Analysis - Overview and Applications, Procedia Engineering, Volume 69, 2014, Pages 1258-1267.

[20] J. Tadrat, V. Boonjing, and P. Pattaraintakorn, "A new similarity measure in formal concept analysis for case-based reasoning," Expert Systems with Applications, vol. 39, no. 1, pp. 967-972, 2012.

[21] D. L. Rodriguez, A. Mora, J. Dominguez and A.Villalon, “fcaR: Formal Concept Analysis”, R package version 1.0.