Hanfei Bao: The Research on the Informatics Attributes of Medical Data, Information And Knowledge(I), Chinese J. of Medical Treatment, 2(1): 4-8, 2003

The Research on the Informatics Attributes of Medical Data, Information and Knowledge(I)

Hanfei Bao
The Laboratory of TCM-Informatics and TCM-Standardization
Shanghai University of Traditional Chinese Medicine

Abstract The paper philosophically discussed the following questions: ①the inborn relationship between natural language and human being and its possible effect on the acceptance of the formatted data in CPR by doctors and nurses; ②the axis-dependence in the data, information and knowledge expression and the time-interrupted nature of the data in CPR , which might be complemented by the imagination of doctors based on their unexpressed medical knowledge base in their brains; ③the various temporal attributes in medical data and their examples; ④ some basic temporal expressions and their notation; ⑤symbolization of some original examples of time expression in medical records.

Keywords Computer-based patient record(CPR) Medical artificial intelligence Medical Informatics Time expression

As the views of the author, there are three new research fields in Medical Informatics nowadays which will influence deeply on the future medicine and will make the "good friend" relationship between Medicine and Medical Informatics into "honey moon" relationship. These three research fields are (1)medical ontology[1-3] ,which will involve widely the standardized portion of medical knowledge; (2)computer-based patient record(CPR)[4-5]; (3)computer-based calculation of biological mass data[6]。Although the latter is usually called Bioinformatics, it will be closely related with the observation, understanding, diagnosis and even treatment of diseases, and sometimes people use the name Biomedical Informatics to unite Medical and Biological Informatics. The development of these three fields would be important also for the long-term (perhaps innocent)dream of the author: the large-scale integration of medical knowledge. [7-10]。

From viewpoint of the future medicine, the research of CPR is far beyond a development of an electronic informational product. it involves the basic stones for the storage, expression, processing, operation, exploitation and integration of medical data, information and knowledge. Therefore it will impact profoundly the future medicine. The mass data with the increment of thousands upon thousands daily from medical practices raises a big challenge to the traditional and individual operation by doctors. It can surely be imagined that at least some parts of the work on the analysis, synthesis and mining of masses of the transfield, or heterogeneous medical data are beyond the abilities of doctor's brain. How the data of patient records would be formulated and computed and how the new information would be mined and the new knowledge would be discovered from those large-scale raw data by computer? It can be clearly foreseen that these questions would be increasingly important in medicine and life sciences.

This series of the investigations by the author will mainly involve the problems of the informatics attributes of the patient record (traditional or electronic) data expression.

Ⅰ.Natural Language: a inborn relationship with human being

Language or language system is some kind of the visualization, audiblization and materialisation of our brain's actions. Language system is such a system which is so close to us as our hands, feet and clothes, that makes us mistaking that we know it very much. Whereas because we know scarcely about the actions of our mysterious mind, and consequently we know scarcely on the language system, either. It is a barrier on the way to reach our expected goal of medical artifitial intelligence. But if we expect to achieve the substantial progress in artificial intelligence, particularly in medical artificial intelligence which might be the most complicated one, the barrier has to be gone beyond.

So called the language of patient record in this article generally refers to all of the data recorded in patient's history. Like any other languages, it can be philosophically divided into the three portions: (1)The reflection of the world in the mind; (2) The reflection of the world in the reality; (3) The reflection of the world in the interaction between the previous two. The one of the most challenged currently of PCR and its efforts for the formalization of the medical data is the acceptance of doctors and nurses. We try to discuss the problem more profoundly in following text. So called natural language or free text is unformatted language contrasting to the formatted data, information and knowledge. In our daily conversation, we even don't need to follow strictly the grammar. The only "gold standard" is that there is no misunderstanding between people who are communicating. But as the reusable or common-sharable medical resources, especially expecting to be powerfully supported by computer, medical records need formatted data rather than free text. It seems, however, that the problem is far from being simple. It is quite possible that the free text reflects the "freedom desire" of human being by instinct in the actions of observation, expression, recognition and alteration of the world. That is there might be a inborn relationship between natural language and human being. And because natural language has been habitually accepted by human being generation by generation, the expression load and the cognition load for us by natural lanuage should be much less than by formatted data, although this conclusion needs to be confirmed by experiment. Therefore when we talk about many merits and huge potentialities of formatted PCR, we should not forget many merits of natural language and its inborn relationship with people. That is may be one reason why doctors are aware clearly of the functional shortage of natural language in CPR, but at the same time reluctant to part from it. That might be explanatory why when we do our best to format the patient data, we always should not forget to reserve an area for free expression.

Ⅱ. Reference-axis-dependence of language:

Human being faces an indefinite world. We can never end our interesting and desire of observation, description, recognition and change(or control) of the world which include ourselves. When we observe, describe and recognize and change the world, we always, obviously or unobviously, consciously or unconsciously, choose a reference axis(or dimension) first. We may frequently call axis other terms, for instances, viewpoint, viewangle, parameters, variables or attributes etc. according to particular situations. Any our actions, no matter physical, mental, linguistic, etc depend on these axes, otherwise we hardly say any words or any sentences. For example, when we talk about a disease, we can't do without physical(size, weight, shape, cardio-electricity, myo-electricity, encephalo-electricity, X-ray, etc), chemical(blood sugar, urine sugar, serum potassium, serum total cholesterol, etc), anatomical(symmetry, deformity, developmental condition, etc), physiological, immunological (antibody, antigen, T-cell, K-cell, macrophage, etc), psychological and genetic parameters, features or symptoms. For any words, phrases, sentences, paragraphs, no exceptions exist. Therefore whatever we recognize and express, no matter by

character, number, graph, image, audio frequency, video frequency, etc are actually a variety of axis-dependent models of the original world. The reference-axis-dependence is a very important pragmatic aspect for any kind of data and knowledge sources: rule-, frame-, object-oriented expression, including my three-set(subject set, object set and condition set) relationship of knowledge expression[9-10].

Facing to the ten-thousand-axis world, there is no way to limit those axes and their combination to a finite set. Additionally we usually are not fully aware of the axes' existence, in other wards, they are usually potential. These reference axes and their combinations as the factors of backgrounds or pragmatics are among the common or tacit cognition of people who are communicating and understanding each other. Unfortunately for the communication between computers nowadays, every thing should be expressed explicitly, nothing could be held back. That might be one reason why the developments of the artificial intelligence are often not as ideal as we expected.

Ⅲ.Patient history "montage"

Like any other history record, patient history could be considered as the function of time. Of course, the nature of patient history is continuous. Whereas we can usually record it at special time points or durations. Therefore in fact patient history is always interrupted, constructed by the fixed data at certain time points or durations. The data gaps between the time points or durations are finally bridged by the imagination of doctors based on the unexpressed knowledge base(UKB)[10] in their brains. Thus the history of a particular patient is "continuous" or "dynamic" for the doctor, similar to the montage of movie. It is also simply similar to palaeozoology. What the palaeozoologists saw is the pieces of the fossil bones of prehistoric dinosaur, what we see in the movie <Jurassic Park> , however, are the lively creatures with muscle, skin, limbs, trunk, hunting for food, drinking water and even expressing emotions. They are completed mostly by the "trained imagination" of the palaeozoologists.

The "imagination" assisted by the unexpressed(or potential) knowledge base, and perhaps, experience base(UEB), picture base(UPB) in our brain plays a very important role in our mental activities. It is sometimes so powerful that the smaller or larger gaps of knowledge or information in certain domain can be leaped over. We often do this way when we try to understand the relations between the knowledge of, for instance, the basic medical sciences and the clinical sciences. As we known, for the computer nowadays those gaps are still a big challenge.

Ⅳ.The temporal attributes of data:

In this paper, so called "event" means any recorded medical semantic unit under discussion, without making further the differences between event, action and component or element, like in the book <Handbook of Medical Informatics>, edited by JH van Bemmel and MA Musen[11]. As we known, time is an intrinsic and essential attribute of data, information and knowledge. Imaginablly, the semantics related with time of the events in medical record is vital for data mining, knowledge discovery, decision making, etc and is a virgin land which is waiting for exploitation. Many scientists of Medical Informatics repeatedly pointed out that these functions are possible only when the data and their times are formatedly represented, because the free represented data and their time are confused for computer. Then what is the situation of data time represented in free text? The time representations in free text appear quite "free" indeed and are usually determined by three factors:①the particular and innate temporal character of the events under consideration; ②the semantic requirement which we need to fulfil and ③ the style which the writers prefer.

Modern cosmology declared that the birth of the universe is the first beginning of the general time. That means from the viewpoint of the universe evolution the general time is like a forward arrow. Whereas from

viewpoint of our daily life(or from midscopic viewpoint rather than from macroscopic or microscopic viewpoint), the general time is more like a two-direction arrow(Fig. 1), moving both backward and forward endless.

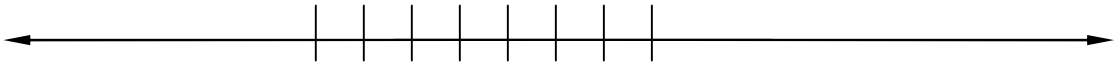

Fig. 1 the general time from midscopic viewpoint

Time of an event generally has the following intrinsic attributes or parameters: starting point（S）, ending point(E), duration(D) or continuity, openedness-closedness, orderedness, certainty, regularity, periodity, etc. And there are also other attributes such as the representation features(absolutely, relatively, positively, negatively represented) and the semantic attributes etc. These attributes affect unavoidably medical artificial intelligence.

(1)The temporal attributes for single event:

Essentially the time of a single event consists of three basic parameters: S, E, D. Among them, any one can be derived from the other two. Therefore we may get the expression t = f(S,D), here t is the time of an single event.

①Continuity or duration: Continuity or duration is defined as the length of time. If D = 0, then the time is called the point time(PT) e.g. "at 7: 30 in the morning taking sample of venous blood" ,and if D > 0, the non-point time, or period time ,e.g. "had been taken the drug for one week", "has a chill for 3 days, accompanied by chest pain and cough"., "has been taken piperazine phosphate for 7 days from the 20th, September, total dose was 4.2 g" .

② Certainty : Any temporal attributes such as starting point, ending point, duration, continuity, openedness-closedness, orderedness, etc might be certain or uncertain.

The point time which is certain is named certain point time, e.g. "having lunch at 12:00", "admitted at 15th of March, 1995" , otherwise it is called uncertain point time，"abdominal pain sometimes". Some times it can be partially certain. e.g. "has suffered fever since yesterday"（without knowing when it stops）。

The certainty usually depends on the semantics of the events, for instances, in clinical care monitoring, the patient data each minute or even each second should be taken account of, whereas for many chronic disease, patients might be examined monthly or yearly.

③openedness-closedness: If both starting point and ending point are certain, the time is called the closed time, otherwise it is the opened time or arrow(AR) time. The arrow time can be further subdivided into three types theoretically, i.e. the forward arrow(FAR) time and the backward arrow(BAR) time and two-direction arrow(TDAR) time, if the ending point, starting point and both the ending and starting points of time are uncertain, respectively. Here are some examples: "starting smoking in 1977" and "went into a coma 10 minutes ago" for FAR and "stopped smoking in 1974" and "having been in good health before the attack" for BAR.

(2)The temporal attributes for multiple event:

①regularity: It refers to the temporal relationship between a repeatedly occurring event and another event. e.g. "epigastralgia in starvation". We call this type of time regular time.

②periodity: A multiple event occurs at regular intervals. e.g. "take the pills every four hours".

③orderedness: A series of the dynamic or sequential actions occur in order of time. A typical example is an

surgical operation. Each former action usually is the prerequisite for the latter, whereas the latter is the trend of the former.

(3)representation attributes:

①absolute time: Strictly speaking, we only have time in the relative sense. Any time has its reference time point or event, otherwise there is no time at all. So called absolute time such as "4: 58 PM, the First of May , 2002 " has its reference event of birth of Christ.

②relative time: The expression of relative time is by certain reference event. For examples, "one week latter", "The patient has low fever, dizzy and headache reactions when taking the medicine."

The most common used reference time is the data-acquiring time(T). It can be expressed as before , after, directly before or directly after, indirectly before or indirectly after data-acquiring time.

The second common used reference time is age(AGE), which can be calculated from the birth time(BT).

③positive (or right)time(POST): The positive time here means the time when the events interested in being going on, e.g. "I am in my office from 8:00 morning to 5:00 afternoon ","chronic constipation for one year."

④negative time or complementary time(COMT):The expression of time might be also the interval or absence of things described, e.g. " I am not at home from 8:00 morning to 5:00 afternoon", "giving up smoking for ten years", "no history of palpitation, short breath, cyanosis, nocturnal paroxysmal dyspea", etc.

(4)Semantic attributes:

Many medical concepts, terms or events themselves carry the temporal attributes. For examples, diabetes usually means "lifelong disease", "hepatic function test HbsAg/HbeAg positive" usually does not mean an instant event. "peptic ulcer" implies "regular" epigastralgia。"epilepsy" means an "aperiodic" attack, "measles" signifies "the course of disease about 10 days" and "lifelong immunity afterwards".

Ⅴ. The symbolized expression of time

As mentioned above, the times in medical records are expressed in quite different forms depending on the temporal nature of the event, the purpose to treat the event or just the expression custom of the doctors. In order to investigate the possibility to formalize the various types of the time expressions and further to process them automatically by computer, the following parts of the paper will discuss the symbolization of these time types.

In table 1 we list the some basic temporal expressions of time and their symbolization with arbitrarily determined notation.

The following shows some examples for a number of different time expressions common found in patient record and their symbolized forms. Let's assume data acquiring time(T) be "the third, October, 2000". In order to distinguish from the temporal background, the discussed portion of time is italic and underlined in symbolized forms, and the discussed events in free text are double underlined.

"frequent epigastralgia for two years" . The symbolized form: |..*..*..|^, S=T-D= the third, October, 1998,D= two years。

"no history of micturition frequency, micturition urgency and urodynia" , The symbolized form: |.....^, S=BT,D=AGE+。

"half an year ago, the hepatic function tests done by our outpatient service show: total bilirubin 10.26umol/L, ALT 86u, HbsAg\AbeAg positive，anti-HAVIgM and anti-HCV negative，then admitted to an infectious disease hospital as the patient with acute non-icteric type B virus hepatitis" . The symbolized form: |…..|^,S=T-00.06.00=

the third, April, 2000,,D=0;

“half an year ago, the hepatic function tests done by our outpatient service show: total bilirubin 10.26umol/L, ALT 86u, HbsAg\AbeAg positive，anti-HAVIgM and anti-HCV negative，then admitted to an infectious disease hospital as the patient with acute non-icteric type B virus hepatitis”. The symbolized form based on the medical semantics:|------^,S=T-00.06.00= the third, April, 2000,D=? and D > 0;

“half an year ago, patient was admitted to an infectious disease hospital and had been treated with Silibinin, etc for one month ”, The symbolized form for the event: |----|…..|^, S= the third, April, 2000, D=one month;

“half an year ago, patient had been treated with Silibinin, etc for one month, afterwards ALT restored normal and HbeAg remained active ”, The symbolized form for the event: |-----|-----^, S= the third, May, 2000,D=?;

“eructation, acid regurgitation some times”， The symbolized form for the event: ..*..*..^, S=?,D=?;

“barium meal examination of digestive tract was done in the hospital and diagnosis of antrum gastrititis was made”, The symbolized form for the action of diagnosis: ..*..^, S=?,D=0; The symbolized form for the medical semantics of the diagnosis: ------^, S=?,D>0;

“epigastralgia and midgastralgia *regularly*.” The symbolized form: ..&..&..^ ,S=?,D=?;

“Because repeat tonsillitis, tonsillectomy was done in 1980” , The symbolized form for the action: ^^…..|..*..|, S=1980,D=0;

“Because repeated tonsillitis, tonsillectomy was done in 1980”. The medical semantics of tonsillectomy is that tonsil no more exists, any problem due to tonsillitis should not be considered afterwards, The symbolized form based on the semantics: ^^…..|..*.., S=? ,D>0;

“(patient) had been in Shanghai during his childhood” , The symbolized form: |-----|…..|^, S=BT,D=the length of childhood;

“The patient has lived in the city all the time”, The symbolized form: |-----^ ,S=BT,D=AGE+。

basic temporal expressions symbolization	
certain point time	!
uncertain point time	*
regular and **uncertain point time**	&
periodic **point time**	~
single **uncertain point time**	..*..
multiple **uncertain point time**	..*..*..
multiple regular and **uncertain point time**	..&..&..
multiple periodic and **uncertain point time**	..~..~..
data acquiring point time	^
the Christian Era	^^
other event as reference time	^^^
the end on a period of time	|
the period during action(positive period)	|-----|
positive forward arrow(PFAR)	|-----
positive backward arrow(PBAR)	-----|
positive two-direction arrow(PTDAR)	-----
the period during the interval of action(negative period)	|…..|
negative forward arrow(NFAR)	|……
negative backward arrow(NBAR)	……|
negative two-direction arrow(NTDAR)	……
uncertain point time in a negative period	| * |

Table 1 Some basic temporal expressions and their symbolization

References

1. C. Rosse et al:Motivational and organizational prinicples for anatomical knowledge representation: the Digital Anatomist Symbolic Knowledge Base", Journal of the American Medical Informatics Association (JAMIA), 5(1):17-40,1998
2. Myeng-Ki Kim, Recent Progress in Ontology-based Medical Computing in Korea,Proceedings of the fourth China-Japan-Korea Joint Symposium on Medical Informatics(CJKMI'02),278,2002
3. AL Rector, JE Rogers, P Pole: The GALEN High Level Ontology, http://www.cs.man.ac.uk/mig/galen/

4. AM van Ginneken , M de Wilde. A new Approach to Structured Data Collection. In: Waegemann CP, ed. Proceedings of TEPR 2000, May 8-11 2000, San Francisco, Ca. 2000:627-35.

5.H Takeda, Y Matsumura and H Nakano: Recent Topics on Electronic Patient Record System in Japan, Proceedings of the fourth China-Japan-Korea Joint Symposium on Medical Informatics(CJKMI'02),3-7,2002

6.BL Hao. and SY Zhang:《Biological Informatics》, Shanghai: Shanghai Science-Technology Press, 2000

7.HF Bao: The structure characteristics of the new research QMSOC and its relevant operators, J Tongji Med Univ, 1989, 9(4):235-238

8. HF Bao, JH Geng and ZF Su: Pansystems Methodology(PM) and a new research on large-scale integration of biomedicine—An introduction of QMSOC and its recent progresses , Acta of Jiansu Industrial College, Journal of Jiangsu Institute of Technology, 1991, 4 (2): 69-75

9. HF Bao: HCSL:A Human-Computer Commonly Understandable and Communicatable Medical Language,《Proceedings of The First China-Japan-Korea Joint Symposium on Medical Informatics (CJKMI'99)》, p177-181,1999

10. HF Bao. Q Liu: To Develop the Human-Computer Complex Pansystem in Medical Knowledge Engineering:"Computerman", International Journal of Famous Doctor, 4(4):1-5,2002

11.JH. van Bemmel, MA Musen: 《Handbook of Medical Informatics》, Houten:Bohn Stafleu Van Loghum, 1997