FORMING MEDICAL DATABASE AND KNOWLEDGE FOR DIAGNOSTIC DISEASE

Oleksii Vasilenko, Oleksandr Kuzomin, Vladislav Shvets

Abstract: *The paper presents the system for intelligent analysis of clinical information. Authors describe methods implemented in the system for clinical information retrieval, intelligent diagnostics of chronic diseases, patient's features importance and for detection of hidden dependencies between features. Results of the experimental evaluation of these methods are also presented.* ***Background***: *Healthcare facilities generate a large flow of both structured and unstructured data which contain important information about patients. Test results are usually retained as structured data but some data is retained in the form of natural language texts (medical history, the results of physical examination, and the results of other examinations, such as ultrasound, ECG or X-ray studies). Many tasks arising in clinical practice can be automated applying methods for intelligent analysis of accumulated structured array and unstructured data that leads to improvement of the healthcare quality.*

Keywords: *data mining, knowledge search, intellectual system, databases.*

Introduction

In general, intelligent medical data processing systems have the following applications [Baranov A.A, 2016]: prediction, classification of clinical cases (diagnosis), search of similar clinical cases, observation of patients' condition.

One of the main areas of application of the intellectual system of medical data analysis is the differential diagnosis of the patient's condition: the discovery of the disease, its stage, the nature of the course of the disease. It is necessary to provide for a step-by-step diagnosis of a patient's illness with a precise diagnosis of the patient at each step.

Another important area of application is the prognosis of changes in the clinical condition of the patient in the application of different types of interventions and in the absence of interventions.

Another important area of application of the system is the tracking of dangerous - critical - changes in patient health indicators.

In addition to the main areas of application, the system will have to monitor and automatically evaluate medical personnel's actions.

Data analysis

For medical diagnostics, the methodology for using data analysis as a working tool is intensively developed during the creation of medical knowledge databases. This work is carried out in many research centers [Baranov A.A, 2016], while the ways of solving individual issues have been slightly modified, but the principled approach to building knowledge bases organized in the form of a hierarchical tree of concepts is maintained. Recently, in order to simplify the work of the doctor, a filling was added to the primary card (questionnaire) that was implemented on the computer screen. The doctor fills in a questionnaire with as simple features as possible, and the process of forming concepts is carried out using intelligent computer solutions (Fig. 1).

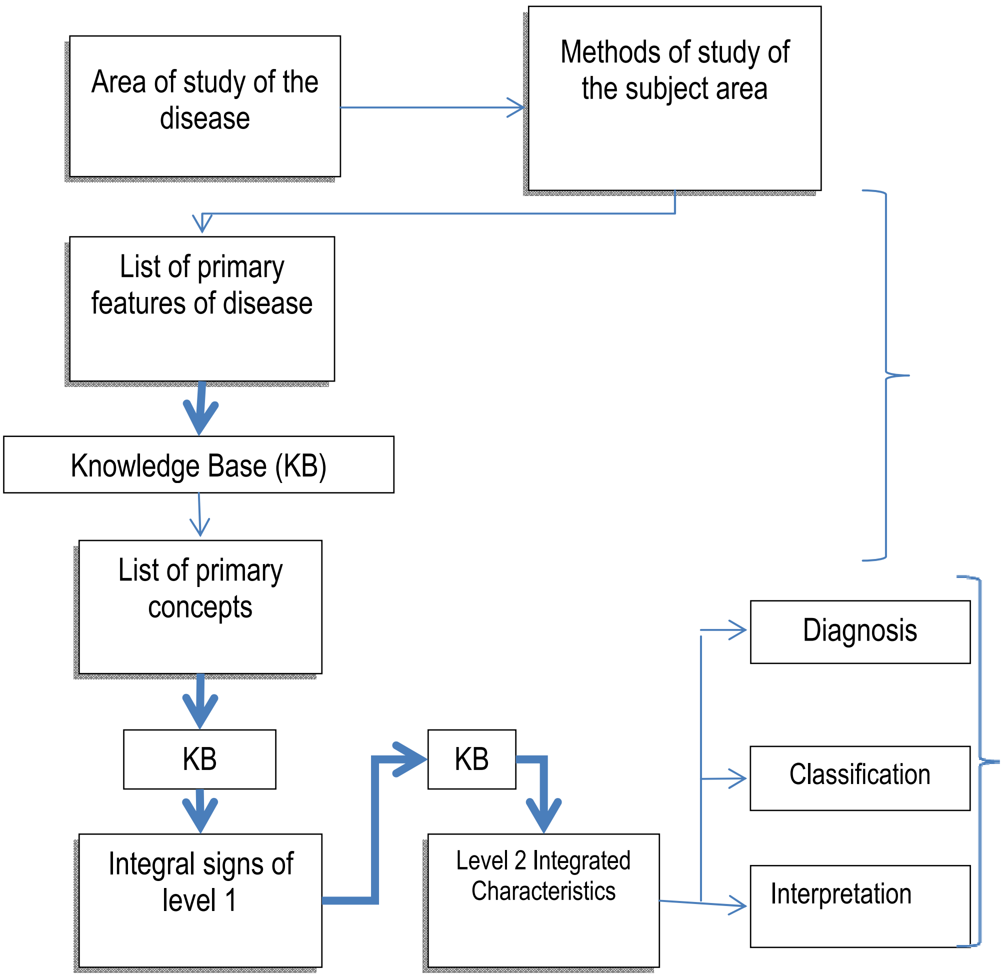
Area of study of the disease
Methods of study of the subject area
List of primary features of disease
Knowledge Base (KB)
List of primary concepts
KB
KB
Integral signs of level 1
Level 2 Integrated Characteristics
Diagnosis
Classification
Interpretation

Figure 1 - The basic scheme of the work of a doctor and a cassette recorder

At the first stage of knowledge discovery, the area and direction of the study, within or outside which is expected to be further developed, that is, the "Field of study of the disease," is determined. This can be a study of only one process or mechanism, for example, a diagnosis of a disease or a group of diseases, one description of the history of the disease or the study of the possibilities of one method.

At the next stage, it is recommended [Kuzomin and Vasylenko, 2014] to analyze internal knowledge blocks, that is, to identify "Methods or aggregates of techniques that examine one side of the investigated phenomenon, process, disease."

This procedure is performed expertly on the basis of objective possibilities and subjective assessments. Next, for each block of the method of research or a set of techniques that were intended to clarify one process, a list (list) of signs of "integral features of the 1st level" is being compiled, as, for example, mechanical features are considered, for example, breathing when detecting sinusitis.

This list includes all the data from the history of the disease, which may subsequently be included in the computational or logical decision-making procedures selection of the diagnosis. Thus the space of the primary features of "Integral signs of the 1st level" is formed.

In general, the analysis of poorly structured medical data contained in the history of the disease allows us to come to the assertion that we have: all the diversity of numbers, curves, images, clinical findings in different forms, in different formats. First of all, it must be transformed into two main types: numbers and texts. A specialist working with images and curves provides a doctor with a conclusion on the appropriate level of medical technology.

The text, officially included in the history of the disease, is both a medical and a legal document. In principle, the database can be formed in any way. It can contain non-formalized source texts, output values, curves, images (if computing allows). In this case, the work of a cryptographic recorder with a specialist on the formation of the primary space of characters may be conducted through a dialogue through the display screen. During the dialog, you can call the screen images, curves, texts from the database that are needed for analysis. If the computer does not have such a database, the work is carried out directly with the texts of medical stories, curves and images on various material media (paper, film, photography).

The next step is to move from the list of signs to the primary concepts that will form the basis of the information matrix "medical object - a sign" and serve as the basis for building medical advisory systems. The process of transition from the signs to the primary concepts is complex and time-consuming. In the course of it it is necessary to do some iterations and apply different approaches.

Therefore, the task of this study is to bring numbers and texts into a single metric information. This problem is solved in this way. When working with numbers, all numerical sampling is based on the numerical scale from the minimum to the maximum values. Then on this scale, the limits of the norm are allocated. The task of determining the limits of the norm is one of the most difficult medical problems. Here you should refer to the materials of publications [Baranov A.A, 2016].

Within the norm, it is possible to allocate three basic gradations: a typical core and two boundary forms that correspond to the notions of the upper and lower limits of the norm. Some researchers [Kuzomin and Vasylenko, 2014] distinguish more gradations (minnorma, optonorma, maxinorm, etc.). It is most effective to use the so-called dependency functions, which corresponds to the fuzzy nature of the data [Kuzomin and Vasylenko, 2014].

Of course, here it is also necessary to use the terminology commonly accepted by clinicians. Naturally, the boundaries between the levels of concepts are rather conditional, but their allocation is important for the following reasons: among a number of well-known doctors, there is the opinion that, given the relativity of the concepts of "norm" and "pathology", it is necessary to consider these processes as unified [Kuzomin and Vasylenko, 2014].

The real difficulties in determining the limits of the norm and pathology are well known, but still confuse the concept of "norm" and "pathology" can not be. The art of the doctor in many respects lies in the assessment of these shaky boundaries. In their definition, an important role is played by a conscious or unconscious idea of the other aspects of the body's vital functions or manifestations of the disease. Such well-used cytochemical parameters of blood relative to the norm allowed to more clearly understand what is happening in pathology [Baranov A.A, 2016].

In the presence of sufficient material, the methods of data analysis help to investigate the limit values, identify them with subclinical forms of diseases. Therefore, in this study, taking into account the above-mentioned features of medical data, existing methods need to be modified in the direction of more reliable results.

The next procedure for analyzing medical data is that the numerical series is transformed into a conceptual series [Baranov A.A, 2016]. Expert way the whole scale is divided into intervals (graduations), each grade is given a meaningful description.

When working with texts, cognitive science is faced with the fact that the doctor determines the same phenomenon in different words, depending on a number of subjective and objective reasons. This discrepancy can be observed even within the same conclusion of the same doctor. Consequently, the cognitive scientist must, "having traveled along the trail" of the expert doctor, compile a dictionary of his definitions, and then try to find out from the physician their contents, build chains of definitions and rank them according to the increasing severity of the process.

For example, a cytologist, describing the sputum, can describe its color, note the presence of blood streaks, the presence of erythrocytes among other cells. In this case, describing the color as rusty, noting the presence of erythrocytes or blood streakscitologist actually describes the same phenomenon: the presence of pulmonary hemoptysis.

As a result, it is possible to come to the chain of concepts and transform the unformed non-rendered description into ranks of the ranked concepts (absence of hemoptysis, small hemoptysis, severe hemoptysis, etc. in an ordered conceptual series.) (Table 1) gives a number of examples of the construction of primary concepts.

Table 1 - A series of examples of constructing primary concepts for GDS

Primary (notions that form) signs	Primary notions	Rules for the formation of primary concepts
Number of neutrophils in the field of vision (NF)	The sputum is slippery Slime-purulent sputum Purulent-slippery scrotum The sputum is slippery	NF up to 30 NF from 30 to 100 NF from 100 to 200 NF 200 and more
Sputum color	Hemorrhage	Sputum is light, no streaks of blood, red blood cells are absent.
The presence of red blood cells	Small hemopus	Sputum is light, without streaks of blood, red blood cells.
The presence of streaks of blood	Intact hemopus	Blood prickles in sputum or rusty sputum
Residual volume of lungs (ZO), residual volume of lungs (ZEL), age	Sharp increase in non-inflammatory parts of the lungs	ZO / ZEL more than 0.5 at the age of 30 years ЗО / ЗЕЛ more than 0.6 at the age from 30 to 40 years ЗО / ЗЕЛ more than 0,63 at the age over 40 years

Thus, the concept of the primary concepts as a unit of knowledge, which in the context of this task is not subject to further specification or division, is formed. From the above, we have that the primary concepts are formed by the cogneologist in conjunction with the expert doctors on the basis of the primary features that are the conceptually constructive elements. In this case, the first part of the knowledge base contains the definition of primary concepts, their clinical interpretation.

So, in the end, after these steps, we can have information that can:

- Be oriented to a generally accepted interpretation without any changes;
- have a clarifying character with the addition of additional elements to achieve an appropriate level of accuracy;
- have a look of illustration with a demonstration of a typical painting. New terms may be introduced (especially when describing new or little developed research methods).

To move from primary features to primary concepts requires a certain knowledge base. It includes a set of logical rules of the form

"IF, TERMS OF DEFINITION"

and may include a set of easiest computing procedures. So, to determine the concept of "lengthening the period of filling the right ventricle" according to jugular phlebography, primary characteristics are required: the duration of the spacing of the spacecraft on the ECG and the spacing of the UA on the phlebogram. In addition, the knowledge base should contain a regression equation that describes the relationship between QC and CA in healthy people, the magnitude of the standard deviation relative to this regression line "b". Then the process of determining the concept of "extension of the period of filling the right ventricle" involves calculating the deviation of the observed UA from the proper at this CK in the particles "b". As a result, the primary concepts become formalized and form the basis of the table "object - a sign." The software support of this part of the knowledge base should provide ease in the modifications of the definitions of the primary concepts for setting the system to work conditions in a particular medical institution.

The second part of the knowledge base is formed as a result of constructing the concepts of higher levels of abstraction of integral features (classifications). Under the integral sign is the complex concept of a high level of abstraction, which is based on a set of primary concepts, constructed on a strictly formal basis, using multidimensional methods of automatic classification. In some cases, the concept may be a logical refinement of existing medical classifications that have had the character of typologies or may lead to a new classification [Kuzomin and Vasylenko, 2014].

Often the integral sign coincides in its content with the medical concept of clinical syndrome [Kuzomin and Vasylenko, 2010]. Thus, the proposed methodology in [Kuzomin and Vasylenko, 2014] considers the construction of knowledge bases of consulting or expert systems as a process of bringing the totality of data and knowledge in the selected subject domain into an ordered hierarchical structure. "Raw" data from the history of the disease through the knowledge base of the first level transform into the conceptual series. At the next stage, the concepts that are primary and considered as data and using the knowledge base of the next level are transformed into integral signs, actually closes the results and their interpretation within one block of research. At the next stage, these integral signs are considered as data and using a higher level knowledge base used to classify, construct diagnostic findings, predict the course of diseases, and so on.

This methodology facilitates the interpretation of known knowledge and pushes a specialist to heuristic decisions in terms of explaining pathophysiological mechanisms, processes and dependencies. As to the use of this methodology in our study, all the main stages of data analysis will be the same, but their essence will have development in the direction of intellectualization using neurons structural algorithms.

In addition, several important points to be borne in mind when solving the problems that were formulated in the works of leading researchers in the field of computer-aided diagnostics:

- First, during the care of each patient in the clinic, the growth of electronic medical records (EHR) [Baranov A.A, 2016] and clinicians increase access to large volumes of data that not all doctors have the right, timely and qualitative use of this information.
- Secondly, BD - the big data requires the organization of their management, ensuring the reliability of storage, constant updating, connection with search robots.
- Thirdly, CDS, which is a computer support for clinical decisions that can change the dynamics of information at the patient's bed and have the opportunity to improve the quality of treatment.
- At the same time, the growth of the use of electronic medical records (EHR), which allows clinicians to increase access to large volumes of data collected during the care of each patient. This combination has led to some overload of the physician with information that challenges the physician's ability to focus on the necessary information, adjust this information to clinical practice standards, and use a combination of clinical data and medical knowledge to assist the patient with the best available medical evidence.

A side effect of the EHR population with patient data is the creation of large data warehouses that contain the accumulation of various clinical data. An analysis of these sets of data collections can give an idea of the nature of the disease and may indicate which of the available diagnostic and therapeutic approaches are likely to yield the desired results. Moreover, it is the information that is needed to support the creation of computer-assisted clinical decision support (CDS), which can change the dynamics of information development at the patient's bedside.

In order to use this information to develop CDS applications, you need to increase resources for obtaining data from data warehouse, analyzing it, and creating decision support tools that can help maintain patient care. This usually requires the collaboration of clinicians, database analysts, statisticians / data miners and software developers. This commitment of large resources is a key obstacle to the widespread use of data stored in clinical repositories for the development of decision support applications.

The literature [Baranov A.A, 2016] describes data analysis environments that are considered for the use of ontological data and knowledge. These systems are intended to demonstrate the means by which it is possible to use ontologies in conjunction with specialized data analysis programs.

By the way, this can reduce the requirements for the resources required to develop CDS diagnostic software. In the references this environment is called: "ontological system of diagnostic modeling" (ODMS) [Baranov A.A, 2016] (Fig. 2). Considering what we said in our study should take into account the most successful decisions regarding the use of medical ontological data on the disease and the construction of data warehouse (EDW).

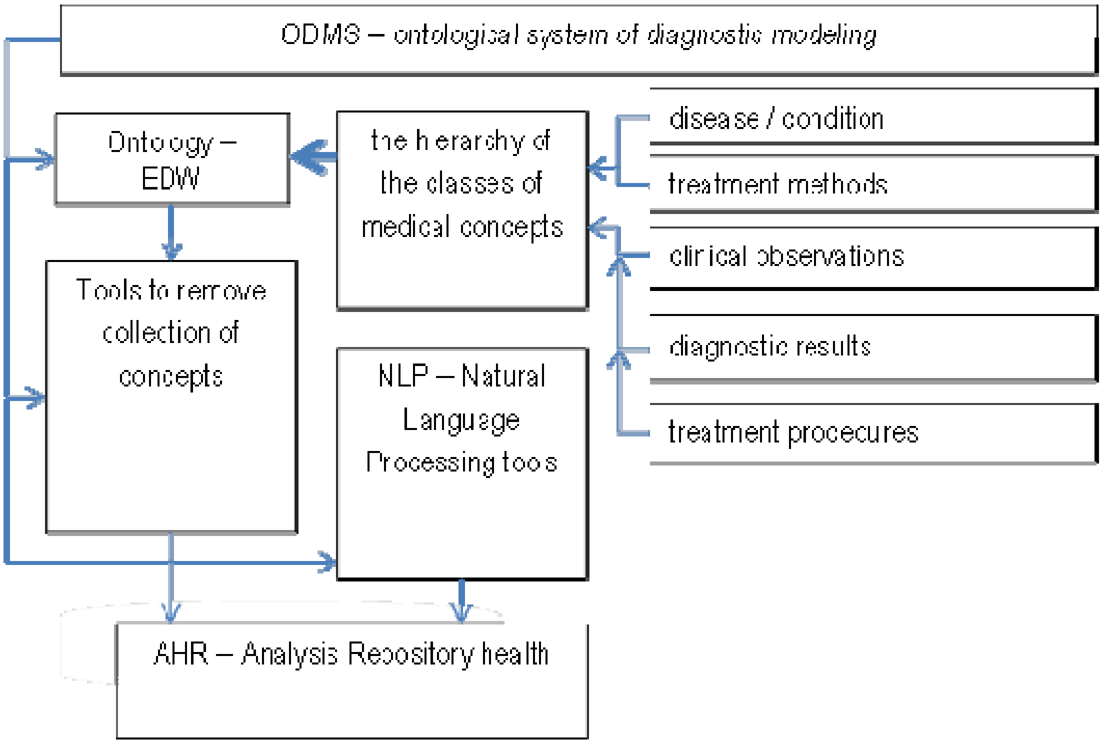
ODMS – ontological system of diagnostic modeling
Ontology – EDW
the hierarchy of the classes of medical concepts
disease / condition
treatment methods
clinical observations
diagnostic results
treatment procedures
Tools to remove collection of concepts
NLP – Natural Language Processing tools
AHR – Analysis Repository health

Figure 2 Ontological system of diagnostic modeling

Automation of the search for signs of illness

At present, the idea of automating the search for signs of a disease for symptoms has become relevant and have beneficial results [Baranov A.A, 2016]. However, when considering the problem of identifying symptoms, attention should be paid to the existence of uncertainty, subjectivity in assessing the patient's condition. Necessary understanding of the problem of identifying symptoms in contact with the patient's doctor. Some aspects of communication with the patient when setting up a preliminary survey are reflected on (Fig. 3).

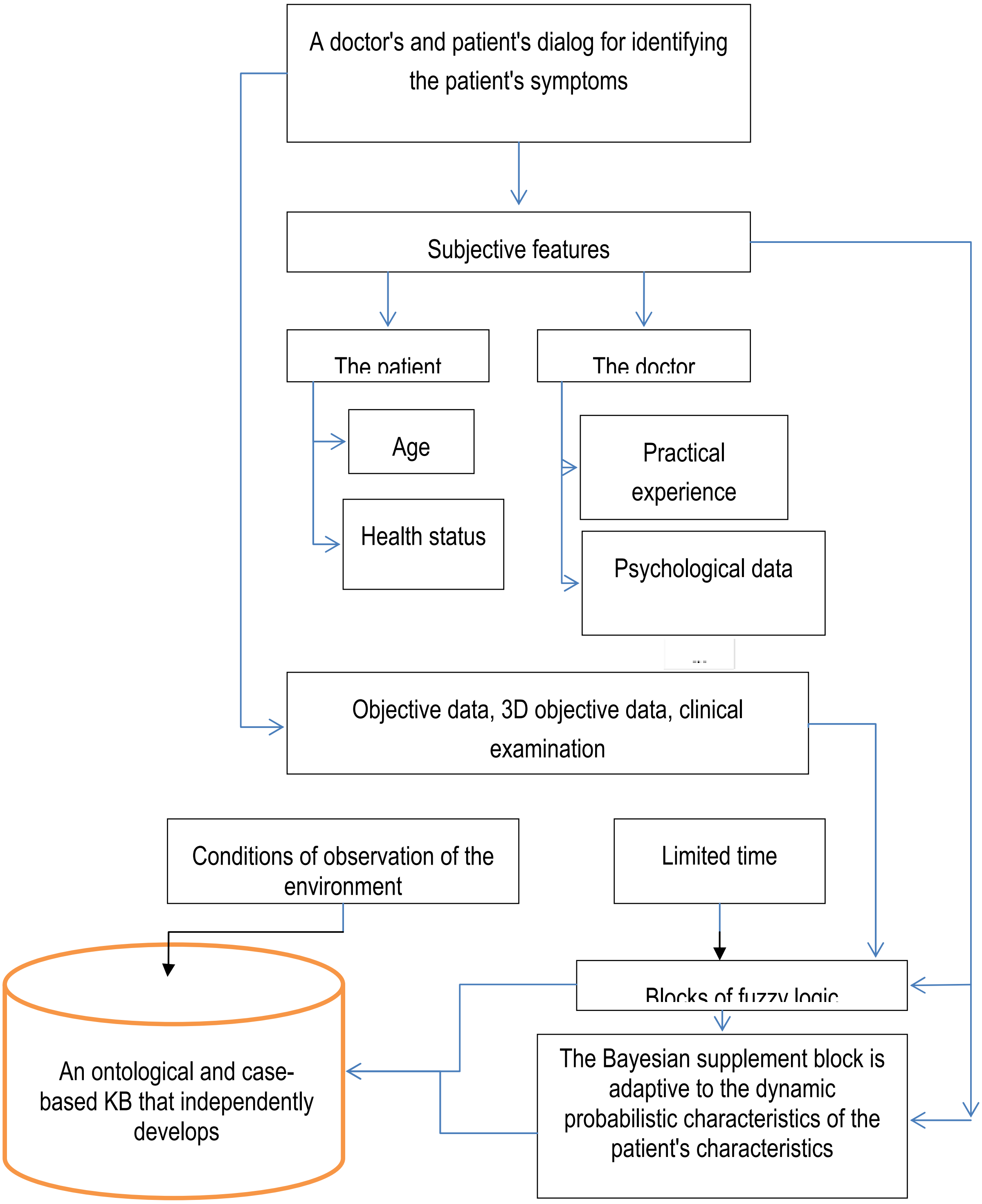
A doctor's and patient's dialog for identifying the patient's symptoms
Subjective features
The patient
The doctor
Age
Health status
Practical experience
Psychological data
Objective data, 3D objective data, clinical examination
Conditions of observation of the environment
Limited time
Blocks of fuzzy logic
The Bayesian supplement block is adaptive to the dynamic probabilistic characteristics of the patient's characteristics
An ontological and case-based KB that independently develops

Figure 3 - Formation of the basis of symptoms in the expert system of questioning and preliminary diagnosis

Conclusion

On the basis of the intellectual analysis of medical data and proposed typical directions of application, primary tasks were identified in developing an intellectual system for diagnosing diseases. Most significant of them:

1) Search for structured data upon request.

2) Search of hidden logic and statistical regularities in given sets of medical data.

3) Classification and prediction of signs of patients on the basis of revealed patterns for solving generalized problems of diagnosis and prognostication.

4) Structured Data Grouping.

5) Work with super massive data.

6) Linguistic analysis of text documents.

7) Search for similar text medical documents in the natural language.

Bibliography

[Baranov A.A, 2016] Технологии комплексного интеллектуального анализа клинических данных.// Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнёва Е.А., Антонова Е.В., Смирнов В.И./ *Вестник РАМН.* 2016;71(2):160–171. doi: 10.15690/vramn663)

[Kuzomin and Vasylenko, 2014] Kuzomin, O.Ya., Vasylenko, O., Obespechenie bezopasnosti ispolzovaniia baz dannyh v usloviiah chrezvychainyh situazii. International Journal "Information Technologies Knowledge", Vol. 8, Num. 2. 2014. pp. 173-187.

[Kuzomin and Vasylenko, 2010] Kuzomin, O.Ya., Vasylenko, O., Analiz estestvenno iazykovyh obiektov I predstavlenie znanii. Vostochno-Evropeiskii zhurnal peredovyh technologii, Vol. 6/2(48). 2010.

Authors' Information

Oleksii Vasylenko *– Aspirant of Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;*

e-mail: ichbierste@gmail.com tel.: +380 63 841 66 23

Major Fields of Scientific Research: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics.

***Prof. Dr.-hab. Oleksandr Kuzomin** – Informatics chair of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine;*

e-mail: kuzy@daad-alumni.de tel.: +38(057)7021515

Major Fields of Scientific Research: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.

***Vladislav Shvez**– Master student in Information and Communication technologies of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine;*

email: Vladyslav.shvets@nure.ua

Major Fields of Scientific Research: Big Data, Data Mining, Data Analyses.