ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2019, Vol. 53, No. 5, pp. 288–294. © Allerton Press, Inc., 2019.
Russian Text © The Author(s), 2019, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2: Informatsionnye Protsessy i Sistemy, 2019, No. 10, pp. 00000–00000.

THE JSM METHOD OF AUTOMATED RESEARCH SUPPORT AND ITS APPLICATION IN INTELLIGENT SYSTEMS FOR MEDICINE

An Intelligent System for Diagnostics of Pancreatic Diseases

O. P. Shesternikova[a], *, V. K. Finn[b], **, L. V. Vinokurova[c], *, K. A. Les'ko[c], ****, G. G. Varvarina[c], *****, and E. Yu. Tyulyaeva[c], ********

[a] *LLC Progtek, Moscow, Russia*
[b] *Informatics and Management Federal Research Center, Russian Academy of Sciences, Moscow, Russia*
[c] *Loginov Medical Clinical Scientific Center, Moscow Health Protection Department, Moscow, Russia*
**e-mail: oshesternikova@gmail.com*
***e-mail: ira.finn@gmail.com*
****e-mail: vinokurova52@mail.ru*
*****e-mail: k_lesko@mail.ru*
******e-mail: varvarina@mail.ru*
*******e-mail: elena_tyulyaeva16@mail.ru*

Received July 24, 2019

Abstract—This paper describes an intelligent system that performs the JSM automated research support method, which is designed to diagnose pancreatic diseases, that is, chronic pancreatitis and pancreatic cancer. A preliminary study is presented; further trends for the development of the system are listed.

Keywords: JSM ARS method, *intelligent system*, *pancreatic cancer*, *chronic pancreatitis*
DOI: 10.3103/S000510551905008X

INTRODUCTION

The JSM automated research support method has been successfully used to analyze gastroenterological data in the task of predicting the development of diabetes in patients with chronic pancreatitis [1, 2]. In the course of these studies, new methods have been developed and tested to find empirical patterns and an intelligent system has been created, which consists of a fact base, a solver, and a user interface. Another task in the field of gastroenterology is the differentiation of pancreatic diseases, that is, chronic pancreatitis and pancreatic cancer, which meets the conditions of applicability of the JSM method [3]. For this task, a new intelligent system that uses the Solver of the previous system has been created.

THE SUBJECT OF THE RESEARCH

Acute and chronic pancreatitis (CP) and pancreatic cancer (PC) are the main diseases of the pancreas. Clinical, laboratory, instrumental, and intraoperative data, including morphological data, do not always make it possible to accurately determine the nature of a pancreas disease.

The greatest difficulties arise in a situation where pancreatic cancer and chronic pancreatitis coexist in one patient, when a tumor develops as a complication of long-lasting CP, or when PC is accompanied by severe inflammation.

The prognosis for patients with pancreatic cancer is unfavorable: the 5-year survival remains at the level of 1–4%, while median survival is 4–6 months [4]. Chronic pancreatitis is a long-term process, which masks small neoplasms and complicates the early diagnostics of pancreatic cancer. CP with a duration of more than 10 years is a risk factor for the development of PC [5]. Thus, the differential diagnostics of PC and CP is an extremely difficult task, which is often not solvable using standard diagnostic techniques.

The need for early diagnostics of socially significant diseases, such as pancreatic cancer, necessitates the development and application of the JSM ARS method for the recognition of pancreatic diseases.

DESCRIPTION OF THE DIAGNOSTIC SYSTEM

The system we have developed is based on the JSM automated research support method [3], which contains the main components of an intelligent system (IS): a fact base (FB), a solver, and a user interface.

The system is a client application, i.e., a program that runs on the local computer and is designed for interactive interaction with the user. The data that are necessary are stored in the file system of the local computer. The development language is *C#* 5.0; the application environment is the *CLR Microsoft.Net Framework 4.5+*.

The system consists of four main modules, which are presented below in the figure:

(1) a module for the graphical interface of the system;

(2) a module that provides the basic infrastructure of the system;

(3) a module for input/output (working with the file system) and data preparation (it correlates with the FB IS);

(4) a module for the heuristics of the JSM method (it correlates with the IS Solver).

We will describe the functionality of each module in more detail, its structure and the third-party libraries and platforms.

The module that forms the application infrastructure includes the following:

• Components of support for the *MVVM design pattern* (*Model-View-ViewModel*), which is integrated with the *Windows Presentation Foundation platform*;

• Components for creating a *DI* container that performs *dependency injection*;

• Components for logging errors and messages that occur during system operation.

The Graphical User Interface (*GUI*) is based on *the Windows Presentation Foundation* platform, which is part of the standard class library (*FCL*) within the *.Net Framework*. The advantages of the platform are as follows: presentation of *UI* (*User Interface*) in a special *XAML* markup language, which allows the form description to be separated from the *code behind*; using *DirectX Application Programming Interfaces* (*API*) to increase productivity.

The system interface provides functionality for user interaction with source data and procedures performed in the Solver: entering the parameters of the procedure, executing procedures, viewing, and saving results.

The input–output and data preprocessing module performs the following functions:

• interaction with the file system;

• navigation in data;

• preparation of a fact base;

• export of system data to a specific list of formats.

We will name the main stages for the preparation of a fact base (FB) from source data:

(1) Definition of the *similarity* operation for the examples of source data;

(2) Selection of examples with the studied effect and examples with the absence of the studied effect in the initial array (in the terminology of the JSM method, we say that we attribute the estimates of the truth values: "+1" for the examples of the first group, and thus we obtain (+)-examples, and "–1" are attributed to the examples of the second group, and thus we obtain (–)-examples; the true estimate "0" was not used in the above studies);

(3) Selection of the examples about which the data about the effect are unknown and formation of a control group for prediction (we attribute the estimate "τ").

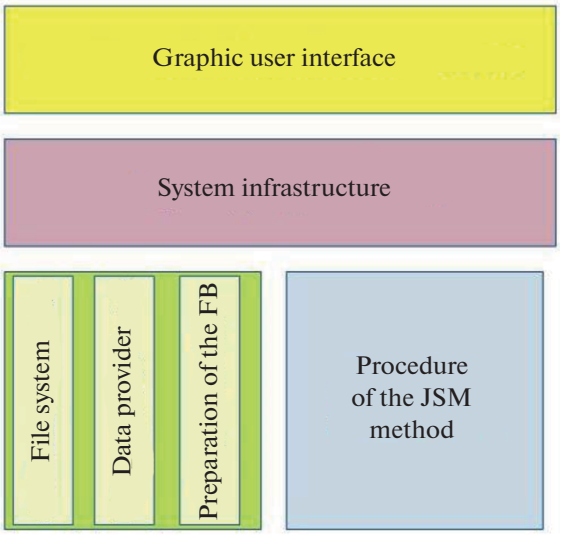

Fig. 1. The architecture of the intelligent system (modules).

In our studies, the data consist of signs (these are tuples or a "flat table" with sign columns and example rows). In this case, the result of the similarity operation for two examples was a tuple from the results of the sequential application of the similarity operation to the elements of the original tuples.

For an intelligent system, an approach was applied in which the source data exist independently of the internal data types of the system that are used for the JSM method. An *Excel* table (*xlsx* format), whose columns correspond to signs, while the rows correspond to examples, is used as the data source. The basic requirements for Table are as follows: the presence of an identifier column (values must be unique); the cells contain only values (no links and formulas).

This approach involves the main stages of preparation of the FB: (1) the conversion of the source data into the internal types; (2) the attribution of estimates of truth values. The settings for these steps are performed in the configuration file. The conversion is adjusted by matching the columns of Table with the signs of examples from the FB; the attribution of the estimates of truth values is performed using formulas that return Boolean values. The configuration file can

Table 1. A table of correct predictions and errors for the strategies*

	l_0		A		b		c	
	+	–	+	–	+	–	+	–
2,4,10,12		5	1		1	2		
6,8,14,16	1	4	1	2				1
1,3,9,11		3	1		4	1		
5,7,15, 13	1	2	1	2		2		1

* The strategies are denoted by numbers from Listing 1 and grouped according to correct predictions and errors.

be edited either manually or using a special utility. Thus, the preparation of the FB is distinguished into a separate independent stage.

For the Solver, the system uses a ready-made module that performs the heuristics of the JSM method developed for the gastroenterological data analysis system [1] and does not provide the user with access to all possible procedures from the module. The procedures of the JSM study that work with exploring fact bases have not been displayed as yet.

The diagnostic system uses the procedures performed in the Solver:

1. The atomic JSM method (the studied effect consists of one sign), including the following:

(1) Similarity predicates: simple similarity (denoted by the symbol a), similarity with the prohibition of a counterexample (denoted by the symbol b), a simplified similarity-difference method (denoted by the symbol d_0);

(2) 16 strategies, each of which is determined by a pair of similarity predicates: for (+)-examples (M^+-predicate) and for (−)-examples (M^--predicate) $Str_{x,y}$, $x, y \in \{a, (ab), (ad0), (ad0b)\}$ [6] (a complete list of strategies with numbers is given below (see *Listing 1*));

(3) Restrictions on the construction of obtained hypotheses about causes:

(a) Restrictions on the number of examples involved in generating similarity ("parents" of a hypothesis): "greater than or equal to," "less than or equal to";

(b) Filters according to the number and values of signs included in the hypothesis;

(c) The value of explainability of the initial array (the ratio of the number of the examples that are "parents" of the obtained hypotheses about causes to the total number of examples: in fractions of 1).

2. Moving control (cross-checking), where each example from the source data is used sequentially as a control group for subsequent comparison with actual values.

Listing 1. The list of strategies used in the system is (1) $Str_{a,a}$, (2) $Str_{ab,a}$, (3) $Str_{ad0,a}$, (4) $Str_{ad0b,a}$, (5) $Str_{a,ab}$, (6) $Str_{ab,ab}$, (7) $Str_{ad0,ab}$, (8) $Str_{ad0b,ab}$, (9) $Str_{a,ad0}$, (10) $Str_{ab,ad0}$, (11) $Str_{ad0,ad0}$, (12) $Str_{ad0b,ad0}$, (13) $Str_{a,ad0b}$, (14) $Str_{ab,ad0b}$, (15) $Str_{ad0,ad0b}$, (16) $Str_{ad0b,ad0b}$.

The user interface enables:

- downloading and viewing data from the file system;
- creating configurations to adjust the fact base;
- adjusting and conducting JSM reasoning;
- viewing the results of the JSM reasoning: hypotheses of causes, hypotheses of predictions and values of explainability;
- exporting the source examples and obtained hypotheses that are displayed on the user's screen to PDF, RTF, XPS, and HTML formats;
- exporting the results to a structured format defined by the system (*xml*);
- viewing the contents of the export results file;
- performing the moving control and viewing the results.

DESCRIPTION OF THE FACT BASE

The fact database that is formed for the task includes the case histories of patients diagnosed with chronic pancreatitis, some of whom were diagnosed with pancreatic cancer. As signs of examples from the FB, 53 signs were selected that belong to the following groups: general clinical data (gender, age, body mass index, alcohol and tobacco dependence, and the duration of the disease); laboratory data (biochemical parameters, general blood tests, as well as tumor markers); the presence of signs of CP and PC obtained by ultrasound examination; and the values of intravenous contrasting using computed tomography. A complete list of signs and their data types are given in *Appendix* 1.

The substantiation for choosing these signs is as follows.

The occurrence of PC is highly dependent on age. Smoking is the cause of both CP and PC. The recent onset of diabetes can also be an early symptom of PC. The clinical pictures of CP and PC have a set of similar manifestations: pain, weight loss, jaundice, and the development of diabetes. The significance of serum tumor markers in the diagnostics of pancreatic cancer is currently assessed as auxiliary. The results can only be used in combination with other methods to accurately determine the diagnosis. The following markers are most often used in the diagnostics of pancreatic cancer: *CA* 19-9, *CEA*, and *CA* 242. However, these indicators have limitations.

The diagnostics of pancreatic diseases are performed using almost all methods of radiation diagnostics. In our country, the most common methods are the ultrasound method (US) and computed tomography (CT). The latter allows evaluation of the degree of attenuation of x-ray radiation during the passage through tissues and expresses it in units of the linear attenuation scale of radiation, that is, Hounsfield units (*HU*), which reflect the x-ray density of tissues. The use of intravenous contrasting with multiphase examination and subsequent densitometric analysis by CT makes it possible to assess the state of tissues and the nature of their blood supply, which plays an extremely important role in the diagnostics of pancreatic cancer.

Each of the selected signs refers to one of three types:

(1) The enumerated type $a_i \in \{a_{i1}, \ldots, a_{in}\}$, $1 \le i \le 53$ (the similarity operation was defined as the value of a sign if the values of the sign coincided in the examples, or the lack of similarity in the opposite case);

(2) The logical type, which can be reduced to the type from point 1 with an enumeration of two distinguished values that correspond to Yes and No;

(3) The numerical type in which the numerical characteristics of two examples are considered identical if they are separated by a distance of no more than a given value. This data type was first applied to the systems that implement the JSM method.

The studied effect was the diagnosis of pancreatic cancer (PC). Accordingly, the group with the presence of the effect ((+)-examples) included examples with the diagnoses of PC and CP, and the group with the absence of the effect ((–)-the examples) included examples with the diagnosis of CP, but without PC.

There are a total of 73 examples in the array, of which 22 are (+)-examples and 51 are (–)-examples ($|FB| = 73$, $|FB^+| = 22$, $|FB^-| = 51$); the control group consisted of nine examples ($|FB^\tau| = 9$).

PRELIMINARY RESEARCH AND FURTHER DEVELOPMENT OF THE SYSTEM

For the analyzed array, JSM reasoning was carried out, and a table, whose structure had been developed in the previous studies, was complied for assessing the accuracy of predictions. [1]. This table is based on a comparison of the predicted values of the examples from the control group and actual values: the group of correct predictions (denoted by l_0), critical errors (denoted by a, (+)- examples are predicted as (–)-examples and vice versa), non-critical errors (denoted by b, the examples are predicted as (0)), and prediction failures (c) are distinguished. The data on the number of examples in the groups are given in Table 1 for the 16 strategies separately for (+)- and (–)-examples.

The table also reflects the division of the set of strategies with regard to the relationship of equivalence of the strategies (the same predictions and the same errors).

The strategies with a strong M^+ predicate and weak M^- predicate showed the best result: they predict (–)-examples that are more numerous in the control group better. One of the cases of correct prediction is given in *Appendix* 2. It should also be noted that there are gross errors (type **a**) in all strategies: a gross error related to a (+)-example is present in all strategies, and two more prediction errors of (–)-examples will appear in the strategies that have a strong M^--predicate. These three prediction errors correspond to the most complex examples from the control group for making a clinical diagnosis. To eliminate errors, revision of the original FB and a change in the set of signs of the examples included in the FB are planned.

Thus, at the moment, the system is to be developed in two directions: (1) analysis of the occurrence of errors, improvement of prediction parameters and (2) expansion of the original fact base and application of the methods for detecting empirical patterns, which were developed in [2].

CONCLUSIONS

Pancreatic cancer is one of the most serious oncological diseases with relatively low survival rates; effective treatment is possible only if it is detected at an early stage. Pancreatic cancer is often combined with chronic pancreatitis; however, early diagnostics of cancer in the presence of chronic pancreatitis is a poorly resolvable task for public health. The use of the JSM automated research support method enables early differential diagnostics between chronic pancreatitis and pancreatic cancer and, therefore, makes it possible to conduct effective treatment.

An important feature of the application of the JSM ARS method is that it uses as facts the results of standard methods of clinical laboratory and radiation diagnostics, which are widespread in modern medicine.

The preliminary results showed the effectiveness of the use of the JSM ARS method in patients with chronic pancreatitis for the differential diagnostics of pancreatic cancer.

The systems of the JSM automated research support method with unified and updated fact bases for the early diagnostics of pancreatic cancer must be used in clinical and telemedicine practice to reduce the mortality rate of the Russian population from oncological diseases.

ACKNOWLEDGMENTS

The study was performed with partial support of the Russian Foundation for Basic Research (project no. 18-29-03063MK).

REFERENCES

1. Shesternikova, O.P. and Pankratova, E.S., An intelligent system for detecting patterns in gastroenterological data, *Trudy Pyatnadtsatoi natsional'noi konferentsii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2016 (3–7 oktyabrya 2016 g., g. Smolensk, Rossiya)* (Proc. Fifteenth National Conference on Artificial Intelligence with International Participation KII-2016 (October 3–7, 2016, Smolensk, Russia)), Smolensk, 2016, vol. 1, pp. 396–404.
2. Finn, V.K. and Shesternikova, O.P., The heuristics of detection of empirical regularities by JSM reasoning, *Autom. Doc. Math. Linguist.,* 2018, vol. 52, no. 5, pp. 215–247.
3. *DSM-metod avtomaticheskogo porozhdeniya gipotez: Logicheskie i epistemologicheskie osnovaniya,* (The JSM Method for Automatic Hypothesis Generation: Logical and Epistemological Foundations), Anshakov, O.M., Ed., Moscow: LIBROKOM, 2009.
4. Russo, A., Rosell, R., and Rolfo, C., *Targeted Therapies for Solid Tumors: A Handbook for Moving Toward New Frontiers in Cancer Treatment,* Humana Press, 2015.
5. Pinho, A.V., Chantrill, L., and Rooman, I., Chronic pancreatitis: A path to pancreatic cancer, *Cancer Lett.,* 2013, vol. 345, no. 2, pp. 203–209.
6. Finn, V.K., Distributive lattices of inductive JSM procedures, *Autom. Doc. Math. Linguist.,* 2014, vol. 48, no. 6, pp. 265–295.
7. Finn, V.K., About heuristics of JSM studies (additions to articles), *Autom. Doc. Math. Linguist.,* 2019.

APPENDIX 1

The list of features used in the fact base and their data types

№	Sign name	Data type
	1. Clinical data	
1	1.1 Gender	Enumeration
2	1.2 Age	Integer
3	1.3 Body mass index	Number with two decimal digits
4	1.4 Duration of the disease	Number with two decimal digits
5	1.5 Presence of alcohol addiction	Binary type
6	1.6 Availability of tobacco addiction	Binary type
7	1.7 Development of diabetes	Binary type
8	1.8 Pancreatic cancer	Binary type
	2. Laboratory data	
	2.1 Biochemistry	
9	2.1.1 Total bilirubin	Number with two decimal digits
10	2.1.2 Direct bilirubin	Number with two decimal digits
11	2.1.3 Indirect bilirubin	Number with two decimal digits
12	2.1.4 Gamma-glutamyltranspeptidase (GGTP)	Number with two decimal digits
13	2.1.6 Glucose	Number with two decimal digits
14	2.1.5 Total protein	Number with two decimal digits
15	2.1.7 c-peptide	Number with two decimal digits
16	2.1.8 Fecal elastase	Number with two decimal digits
	2.2 General blood test	
17	2.2.1 Hemoglobin	Number with two decimal digits
18	2.2.2 White blood cells	Number with two decimal digits
19	2.2.3 ESR	Number with two decimal digits
	2.3 Oncomarkers	
20	2.3.1 CA 19-9	Number with two decimal digits
21	2.3.2 CA 242	Number with two decimal digits
22	2.3.3 CEA	Number with two decimal digits
	3. Ultrasound examination	
	3.1 Reliable signs of pancreatic cancer (PC)	
23	3.1.1 Identification of a volumetric neoplasm (more often solid), hypo and isoechoic identification	Binary type
	3.2 Indirect signs of PC	
24	3.2.1 Uniform dilatation of the main pancreatic duct (MPD) without pronounced wall compaction	Binary type
	3.3 Direct signs of chronic pancreatitis	
25	3.3.1 Calcifications	Binary type
	3.4 Signs of chronic pancreatitis	
26	3.4.1 Hyperechoic structure of the gland	Binary type
27	3.4.2 Uneven dilatation of the MPD, compaction of its walls	Binary type
	4. Computed tomography (CT)	
28	4.1 Neoplasms in the structure of the pancreas	Binary type
29	4.2 Biliary hypertension	Binary type
	4.3 Dilatation of the MPD	
30	4.3.1 No	Binary type
31	4.3.2 Yes, regular	Binary type
32	4.3.3 Yes, regular	Binary type
	4.4 Densitometric characteristics for pancreatic cancer in phases, *HU*	
	4.4.1 Native	
33	4.4.1.1 min	Integer
34	4.4.1.1 max	Integer
	4.4.2 Arterial	
35	4.4.2.1 min	Integer
36	4.4.2.1 max	Integer
	4.4.3 Venous	
37	4.4.3.1 min	Integer
38	4.4.3.1 max	Integer
	4.4.4 Delayed	
39	4.4.4.1 min	Integer

№	Sign name	Data type
40	4.4.4.1 max	Integer
	4.5 Density gradient between tumor and unchanged tissue	
	4.5.1 Native	
41	4.5.1.1 min	Integer
42	4.5.1.1 max	Integer
	4.5.2 Arterial	
43	4.5.2.1 min	Integer
44	4.5.2.1 max	Integer
	4.5.3 Venous	
45	4.5.3.1 min	Integer
46	4.5.3.1 max	Integer
	4.5.4 Delayed	
47	4.5.4.1 min	Integer
48	4.5.4.1 max	Integer
	4.6 Gradient average	
49	4.6.1 Native	Number with two decimal digits
50	4.6.2 Arterial	Number with two decimal digits
51	4.6.3 Venous	Number with two decimal digits
52	4.6.4 Delayed	Number with two decimal digits

APPENDIX 2

An example of a correct prediction

The source example for prediction:

0 – Id	73
1 – Gender	M
2 – Age	72
3 – BMI	26.4
4 – Duration of the disease	4
5 – Alcohol	Yes
6 – Smoking	Yes
7 – ID	No
8 – Pancreatic cancer	
9 – Total bilirubin	12.9
10 – Direct bilirubin	2.9
11– Indirect bilirubin	10
12– GGTP	202
13 – Total protein	70.7
14 – Glucose	5.9
15 – C-peptide	0.6
16 – Fecal elastase	296
17 – Hemoglobin	127
18 – White blood cells	6.3
19 – ESR	33
20 – CaA19-9	974
21 – CA 242	150
22 – CEA	4.5
23 – Detection of a voluminous neoplasm (more often solid), hypo and isoechoic detection	Yes
24 – (US) Regular dilatation of the MPD without pronounced compaction of its walls	Yes
25 – (US) Regular dilatation of the MPD without pronounced compaction of its walls	No
26 – Hyperechoic structure of the gland	Yes
27 – (US) Irregular dilatation of the MPD, compaction of its walls	No
28 – Neoplasms in the structure of the pancreas	Yes
29 – Biliary hypertension	Yes

30 – (CT) There is a dilatation of the MPD	No
31 – (CT) Regular dilatation of the MPD	Yes
32 – (CT) Irregular dilatation of the MPD	No
33 – Densitometry (native, min)	20
34 – Densitometry (native, max)	77
35 – Densitometry (arterial, min)	16
36 – Densitometry (arterial, max)	106
37 – Densitometry (venous, min)	24
38 – Densitometry (venous, max)	121
39 – Densitometry (delayed, min)	37
40 – Densitometry (delayed, max)	137
41– Gradient (native, min)	24
42 – Gradient (native, max)	8
43 – Gradient (arterial, min)	8
44 – Gradient (arterial, max)	6
45 – Gradient (venous, min)	43
46 – Gradient (venous, max)	23
47 – Gradient (delayed, min)	22
48 – Gradient (delayed, max)	29
49 – Gradient average (native)	16
50 – Gradient average (arterial)	7
51 – Gradient average (venous)	33
52 – Gradient average (delayed)	25.5

In this example, the system diagnosed PC using the following hypotheses (the signs that have no values in the hypothesis are omitted):

Hypothesis 1

22 – CEA	3.1–10.2
25 – (US) Regular dilatation of the MPD without pronounced compaction of walls	No
26 – Hyperechoic structure of the gland	Yes
28 – Neoplasms in the structure of the pancreas	Yes
49 – Gradient average (native)	9–16

Hypothesis 2

4 – Duration of the disease	0.45–4
14 – Glucose	5.13–6.27
23 – Detection of a voluminous neoplasm (more often solid), hypo and isoechoic detection	Yes
25 – US) Regular dilatation of the MPD without pronounced compaction of walls	No
26 – Hyperechoic structure of the gland	Yes

Translated by L. A. Solovyova

SPELL: OK