Towards an information system for the collaborative investigation of neurodegenerative diseases

Matthias Virgin[1], Ilvio Bruder[2], Susanne Jürgensmann[2], Meike Klettke[2]

[1] DZNE e.V.
Location Rostock/ Greifswald
Germany
matthias.virgin@dzne.de

[2] Computer Science Institute
University of Rostock
Germany
{ ilvio.bruder,susanne.juergensmann,
meike.klettke } @uni-rostock.de

Abstract

In today's clinical research the importance of collaboration and the need to distribute research results increase. Integrated information systems for manipulating heterogeneous data become more and more important.

In this paper, we introduce a web-based information system for the management of clinical trials. Heterogeneous and distributed data of probands and clinical trials are stored and managed in a centralized system. After describing the main functionalities of the system, we will focus on heterogeneous data and retrieval concepts which enable the search for appropriate probands and clinical trials in the system.

Keywords: integrated information system, management of heterogeneous data, similarity function for data retrieval

1. Introduction

In the German Center for Neurodegenerative Diseases (DZNE), interdisciplinary scientists (from neurology, psychiatry, neuroepidemology, sociodemography, health economics and medical technology) explore the causes and risk factors that lead to a predisposition for neurodegeneration. Based on this, new approaches to identify and treat this type of disease will be developed.

For this purpose different clinical trials have to be conducted. In all areas of medicine these are necessary for the development, evaluation and improvement of medical treatments. With each clinical trial information will be collected about the probands and the achieved results. Thereby, large amounts of data are generated, including medical, demographical, economical and other kinds of information.

In the DZNE, clinical trials take place at various locations all over Germany. To organize and reuse the produced results from the associated studies different requirements have to be met [1].

1. The data that is generated in interdisciplinary studies is very heterogeneous. For this reason the data archives have to be integrated and stored in an efficient and optimized way.
2. A platform has to be realized that enables controlled access onto the collected data along with different possibilities for collaborative work.
3. Methods have to be developed that in the context of particular research questions support the search and retrieval of relevant information.

During the last year, a web-based information system was developed for integrating and managing the information of clinical trials at the DZNE. At the beginning the system was rather meant to support the administration of the probands' personal data. In the course of the usage a lot of medical data where added to the system. Now the system represents a special study management system for clinical trials in the field of neurodegenerative diseases. The system was developed due to the recommendations of the TMF e.V. Association (Technology, Methods, and Infrastructure for Networked Medical Research) [2]. The association offers guidelines about how to deal with personal and medical data in clinical trials [3].

Within the new system, heterogeneous multimodal data with a special interdisciplinary character are stored in a database. The data which originate from different sources are integrated, corrected, completed, preprocessed and presented in a well-arranged way. They comprise personal and medical information about the probands as well as facts about clinical and scientific trials [e.g. from 4, 5]. A reuse of this valuable statistical data makes sense for several reasons:

- Within the next years a large number of clinical trials at different locations of the DZNE will be accomplished. Hence an increased control to avoid double studies is necessary.
- The combination of existing data can potentially lead to new insights. Thereby not only the planning of future research work would be influenced but also so far unknown facts may be revealed.
- Clinical trials (and thus the generated data) are often very cost-intensive, in terms of time, money and effort. A reuse of data can save all of these resources.
- It is essential to make clinical trials as safe as possible. For that reason the prior approval of an ethics committee is usually required [6]. Still, a residual risk always remains. Therefore the less trials are conducted; the better it is; which in turn motivates a better usage of already existing data.

An existing integrated system and a method for finding similarities between probands and clinical trials are necessary for this task. The focus of this paper lies in the introduction of the integrated study management system and in the description of an approach for finding similar and supplemental clinical trials due to probands properties.

2. Objectives

Our main purpose is the assistance of collaborative work in the research area of neurodegenerative diseases. In order to achieve this, we developed an information system that provides access to all relevant data of probands, i.e. personal data and results from medical examinations. A calendar function for the management of appointments and examinations is integrated into the system as well.

The main objectives concerning the development of the platform were: an integrated system storing all information consistently and without changes in format, a user interface that is easy to handle, a component for checking the completeness of data entries as well as a possibility for managing the access control and thus the data protection.

Details about the implementation of this tool will be shown in the next section.

For using the information stored in the system, we need a query and retrieval interface. Our system already offers the possibility for querying proband data, search for probands with selected characteristics, and search for similarities between different probands.

These query function can be applied for finding and selecting probands for clinical trials.

We plan to extend the system by several methods for searching for similar or adequate studies. Retrieval facilities for complete studies are necessary to find information about relevant clinical trials as well as to avoid double studies or similar studies.

The next section describes these query options in detail.

3. Methods

Database Design: The information system for storing results of clinical trials is based on database technology. We use a MySQL database to store all relevant medical data and offer a web-based user interface for data access. Data from other sources (for instance Excel or CSV files) can as well be integrated into the system. For this purpose, techniques from federated database systems were applied to transform, validate and store the data in the database. The database inside the system for supporting clinical trials forms the basis for the similarity search. Figure 1 represents the main part of the database schema.

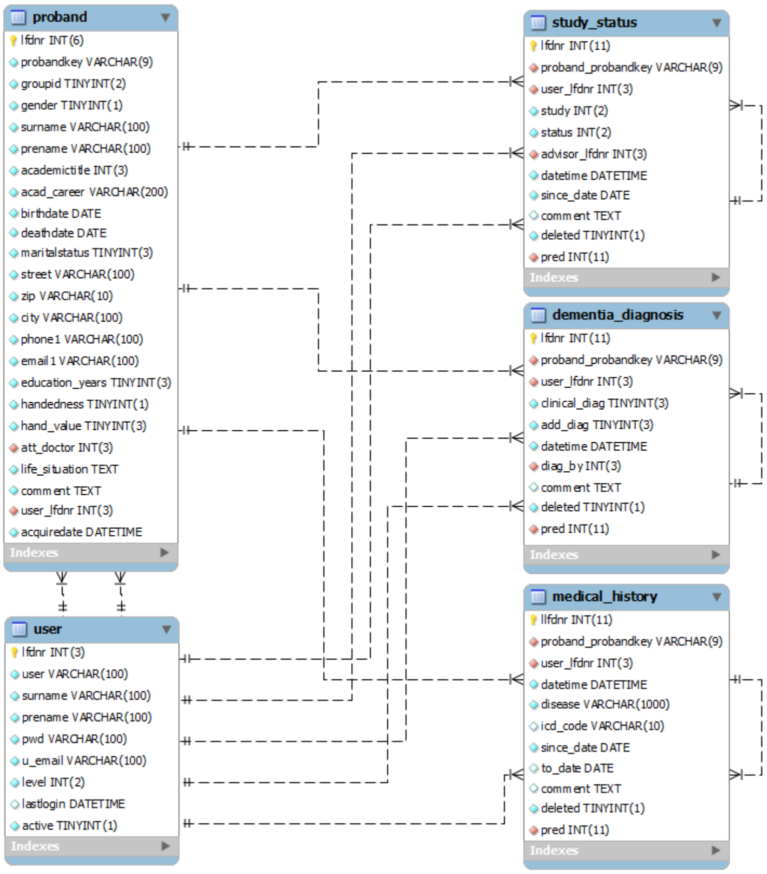
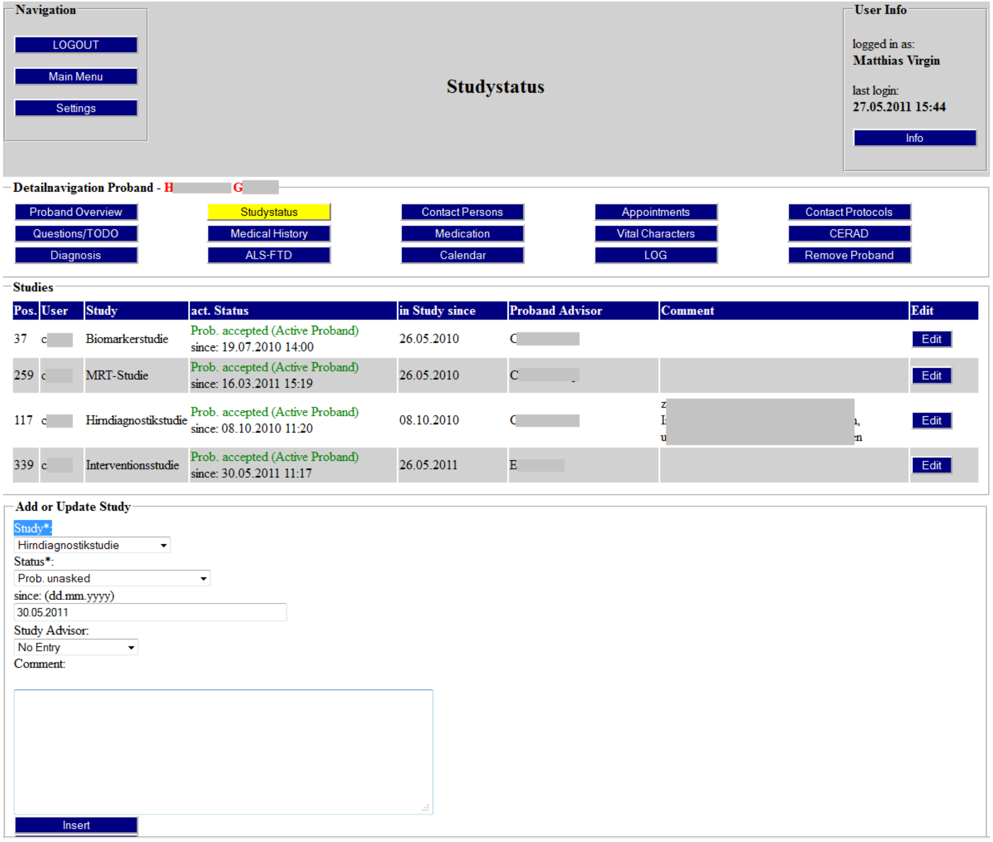


Figure 1: Main part of the database schema for the management of probands, personal data, and participation in clinical trials

User Interface: In order to access the data, insert new information, plan the timetables of medical trials, and evaluate the data, several user interfaces had been developed. Figure 2 shows a user interface for the management of clinical trials as an example.



Studystatus

logged in as:
Matthias Virgin

last login:
27.05.2011 15:44

Detailnavigation Proband - H [redacted] G [redacted]

Studies

Pos.	User	Study	act. Status	in Study since	Proband Advisor	Comment	Edit
37	c [redacted]	Biomarkerstudie	Prob. accepted (Active Proband) since: 19.07.2010 14:00	26.05.2010	C [redacted]		Edit
259	c [redacted]	MRT-Studie	Prob. accepted (Active Proband) since: 16.03.2011 15:19	26.05.2010	C [redacted]		Edit
117	c [redacted]	Hirndiagnostikstudie	Prob. accepted (Active Proband) since: 08.10.2010 11:20	08.10.2010	C [redacted]	z I [redacted] n, u n	Edit
339	c [redacted]	Interventionsstudie	Prob. accepted (Active Proband) since: 30.05.2011 11:17	26.05.2011	E [redacted]		Edit

Add or Update Study

Study*:
Hirndiagnostikstudie
Status*:
Prob. unasked
since: (dd.mm.yyyy)
30.05.2011
Study Advisor:
No Entry
Comment:

Figure 2: User interface for study coordinators

Query and Search Functions: In our work, we introduce an approach that determines similarities of probands feature characteristics for finding similar probands. Data mining technology like applying a function for calculating a distance between different vectors of features is used for this task (see also Han, Kamber, Pei [7]).

Search for Probands: In the first part, we concentrate on the search for probands. We support a query interface using database techniques to search for patients or probands with special characteristics. This operation bases on the data available for the probands, some of these data entries are enumerated in table I, II, III, and IV.

All probands information are stored as feature vectors (p_{i1} , p_{i2} , .., p_{in}) in the databases of our system. Because of data protection rights and recommendations in Germany, medical and personal data of probands are stored in separate database systems.

These vectors are a prerequisite for querying and retrieving data. In the following, the methods for using and adapting the similarity function on feature vectors (p_{i1}, p_{i2}, ..., p_{in}) are described:

1. Selecting attributes from the user profile:
   - Selecting a set of characteristics (p_{ia}, p_{ib}, p_{ic},...). From the list of all medical data, we can select those that are relevant for a medical trial or a special query.
   - For instance: We can decide that the characteristics *gender, age, medication, ICD-10 diagnosis*, and *results of the CERAD test* are important to find probands for a study and select these characteristics. All other characteristics are ignored. This operation corresponds to the projection in database queries.
2. Specify/Concretize a characteristic:
   - For all selected attributes, we can specify required values. This means we search for all profiles with a specific value, range of values or enumeration of values for one or several characteristics.

   - The following adjustments are possible:

Adjustment	***Example***
Exact value	gender = "male"
Range value	(age > 50) and (age <= 60)
negation of exact value	ICD-10 <> "I10.0"
negation of range value	(age < 18) or (age > 65)

This operation corresponds to the selection in databases.

3. Defining weights for each characteristic:
    - Some of the characteristics are more important than others. By weighting each characteristic, we can consider this fact in the search.
    - The weights are in the range (0..1] and represent the priority of a characteristic. A weight equal to 1 means that this element is very important for the following operations, especially for the similarity calculation. A weight less than 1 means that this characteristic has to be considered, but is less important than others.

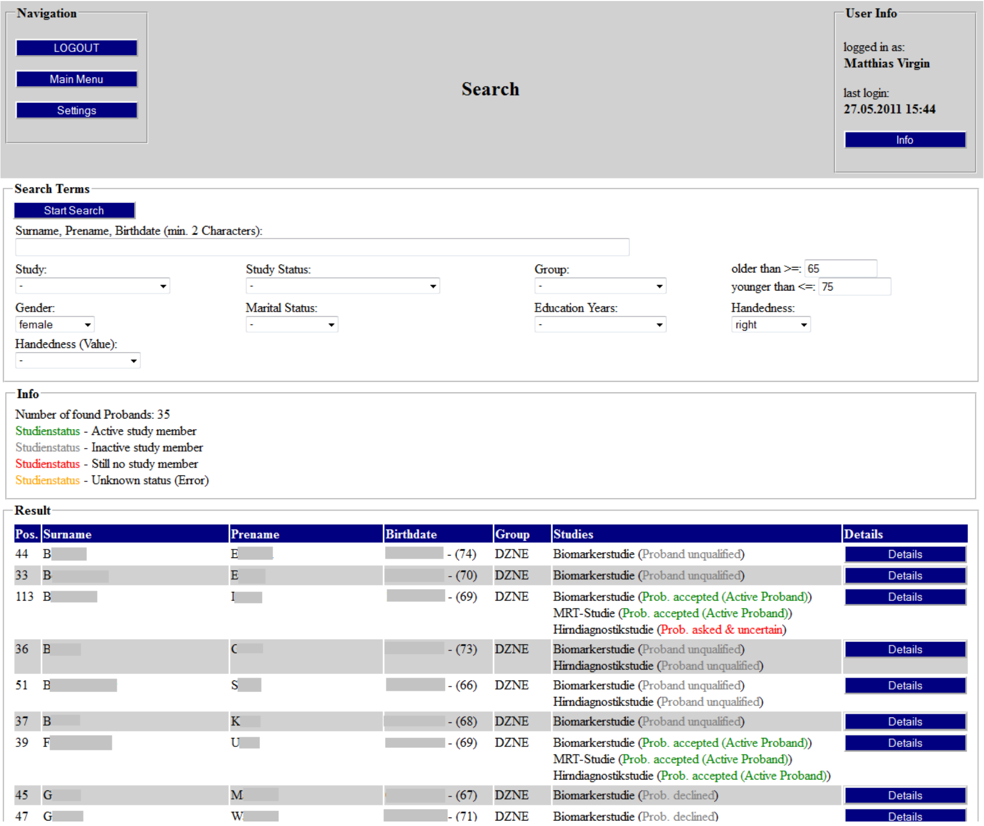
These three characteristics can orthogonally be combined for adapting a query as well as a similarity function.

4. Searching for proband groups:
    - The retrieval function has to consider dependencies between different proband profiles.
    - In some cases, we are looking for probands with a complementary profile. For instance, we are looking for probands of the same *age*, similar life situations, but complementary results in the *CERAD test*.
    - For other screening studies, we are looking for couples, both marriage partners with similar life situations and the same address, one partner diseased with dementia, the other partner not. Finding those couples of probands is possible

with our method because all the information is stored in databases. A database selfjoin operation generates a feature vector for a couple of probands. The above described query and retrieval methods can be applied and couples with the specified characteristics can be found.

Combining all these features, we can search for proband candidates with special characteristics. For instance, we can search for *female probands*, *between 65 and 75 years old*, and *handedness right*.

Figure 3 shows the user interface for this query offered to the study coordinator for querying the proband information in the system.



logged in as:
Matthias Virgin

last login:
27.05.2011 15:44

Surname, Prename, Birthdate (min. 2 Characters):

Study: -

Study Status: -

Group: -

older than >=: 65

younger than <=: 75

Gender: female

Marital Status: -

Education Years: -

Handedness: right

Handedness (Value): -

Number of found Probands: 35

Studienstatus - Active study member

Studienstatus - Inactive study member

Studienstatus - Still no study member

Studienstatus - Unknown status (Error)

Pos.	Surname	Prename	Birthdate	Group	Studies	Details
44	B[redacted]	E[redacted]	[redacted] - (74)	DZNE	Biomarkerstudie (Proband unqualified)	Details
33	B[redacted]	E[redacted]	[redacted] - (70)	DZNE	Biomarkerstudie (Proband unqualified)	Details
113	B[redacted]	I[redacted]	[redacted] - (69)	DZNE	Biomarkerstudie (Prob. accepted (Active Proband)) MRT-Studie (Prob. accepted (Active Proband)) Hirndiagnostikstudie (Prob. asked & uncertain)	Details
36	B[redacted]	C[redacted]	[redacted] - (73)	DZNE	Biomarkerstudie (Proband unqualified) Hirndiagnostikstudie (Proband unqualified)	Details
51	B[redacted]	S[redacted]	[redacted] - (66)	DZNE	Biomarkerstudie (Proband unqualified) Hirndiagnostikstudie (Proband unqualified)	Details
37	B[redacted]	K[redacted]	[redacted] - (68)	DZNE	Biomarkerstudie (Proband unqualified)	Details
39	F[redacted]	U[redacted]	[redacted] - (69)	DZNE	Biomarkerstudie (Prob. accepted (Active Proband)) MRT-Studie (Prob. accepted (Active Proband)) Hirndiagnostikstudie (Prob. accepted (Active Proband))	Details
45	G[redacted]	M[redacted]	[redacted] - (67)	DZNE	Biomarkerstudie (Prob. declined)	Details
47	G[redacted]	W[redacted]	[redacted] - (71)	DZNE	Biomarkerstudie (Prob. declined)	Details

Figure 3: Interface for the Search and Retrieval Component

Search for medical trials: To find similar or interesting clinical trials three different types of information can be used:

1. The meta data available for each clinical trial describe objectives and boundary conditions. This information is available in track sheets and requests to the ethics committee. These documents can be stored in an information system. The information about the clinical trials can be extracted from the documents and made available for other co-workers. The information in these documents is semi-structured. We plan to use methods from text retrieval for making the information searchable and available.
2. A clinical trial can be described by the group of probands and the results. Furthermore, an integrated study information system delivers the information about how many probands already took part in a study, with which characteristics (e.g. gender, age, medication, previous or existing diseases) and what are the results of the medicals examinations. That means: a study is described by its participants, the examinations, and the results; all information can be found in the integrated system.
3. Results of clinical studies are published in scientific articles. To search for suitable and interesting publications describing clinical studies, information retrieval methods have to be applied.

4. Results & Discussion

We implemented an information system for managing studies and probands data. This system is already integrated within the research at DZNE. The scientists have now the opportunity to retrieve and analyze the data across several studies and compare their results with others.

At the moment, the system stores 1.200 probands within 11 clinical studies. A couple of further, multi-centric studies will be integrated soon. The study management system has been developed in the last 13 month and will be continuously redeveloped and completed.

The method described in this paper can be applied for the search of suitable probands for clinical trials. For this purpose the requirements of a clinical trial have to be determined in the first step. Secondly the information system can be queried with the described characteristics.

Probands that turn out to be appropriate for the new clinical trial can then be selectively recruited.

In a future step, we want to provide a possibility to define a distribution on the result set and the possibility for integrating and comparing study results. We want to motivate it with an example:

We are looking for 50 female left-handed probands with an age range between 50 and 90 years uniformly distributed?

1. building a query vector by selecting all relevant elements of the feature vector: *gender, handedness, age*
2. generating a ranked list of all probands with a significant similarity
3. group the proband list by the features age and similarity
4. iterate the groups of the result list and build a uniformly distributed ranked list of probands between 50 and 90 years

Defining a distribution on a characteristic of the feature vector, we have to define the set of probands to distribute and a suitable kind of distribution.

While the search for probands is already implemented, the realization of textual search methods for finding scientific articles about the studies is a future plan.

One reason is that the MySQL database solution offers only limited methods for text search. Therefore, another system has to be utilized for offering access to scientific articles. It was planned to develop an additional collaborative system similar to the technologies of social and business networks for supporting the collaboration and the scientific discussions and exchange.

5. Conclusions

The implemented study management system is working well so far. As the research in the field of neurodegenerative diseases goes on new insights will be gained, new questions will appear and thus, the search for probands and studies will have to take new characteristics into account. Accordingly the search and retrieval functions have to be modified. The presented methods are easily extendable and adaptable. They therefore provide a good basis for new development.

6. Clinical Relevance Statement

The system introduced in this paper can be used for the effective storage and retrieval of all results of clinical trials.

The system is independent from the type of studies, we developed it for clinical trials in the field of neurodegenerative diseases but it easily can be used for other medical trials.

The integration of several functions into one system (from the management of personal data of patients, results of examinations to timetables and management of appointments) increase the efficiency in the management of medical trials.

References

[1] DZNE.de [Internet]. Bonn: German Center for Neurodegenerative Diseases (DZNE e.V.) - Online Resources; c2009-11 [cited 2011 May 30]. Available from: http://www.dzne.de

[2] TMF-eV.de [Internet]. Berlin: Technology, Methods, and Infrastructure for Networked Medical Research - Online Resources; c2011 [cited 2011 May 30]. Available from: http://www.tmf-ev.de

[3] Reng CM, Debold P, Specker C, Pommerening K. Generic solutions to the data protection for the research nets of the medicine [german]. Berlin: Medical scientific publishing house company; 2006

[4] Teipel SJ, Meindl T, Grinberg L, Grothe M, Cantero JL, Reiser MF, Möller HJ, Heinsen H, Hampel H. The cholinergic system in mild cognitive impairment and Alzheimer's disease: An in vivo MRI and DTI study. Hum Brain mapp.; 2010.

[5] Hoffmeyer A, Teipel S, Kirste T. Evaluation of the cognitive state of dementia patients with the aid of mobile sensors [german]. GMDS Conference, Mannheim, Germany, 2010 (Bewertung des kognitiven Zustandes von Demenzpatienten mit Hilfe mobiler Sensoren)

[6] Frewer A, Schmidt U (Ed.). Standard of the research. Historical development and ethical bases of clinical trials [german]. Frankfurt am Main: Peter Lang publishing; 2007

[7] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems), 2006

Table I: Object data in the information system.

Origin	Name	Data Type	Description / Operations
	proband key	private Object ID	system-internal private key
	id	Object ID	
	control group	Boolean	yes or no
	death	Date	only filled with a date if proband already deceased

Table II: Anamnese data in the information system.

Origin	Name	Data Type	Description / Operations
personal data:	name	String	name of a person
	address	String	
	gender	Numeric	male or female
	date of birth	Date	used for age
	age	Numeric	Age selection (derived from date of birth)
	marital status	Enumeration	
	education	String	using terms of defined vocabularity
	education years	Numeric	
	handedness	Numeric -24 - +24	left-, right-handed, and both
	persons history	String	e.g., further addresses, attending physicians, involved nursing institutions, further contact persons), semistructured

Table III: Medical data: persons disease and medication history.

Origin	Name	Data Type	Description / Operations
	medication	Enumeration	elements from "Red List", classification of medication; used as selection attribute, exclusion criterion
	previous or existing diseases	Enumeration	elements of ICD-10; used as selection attribute, exclusion criterion

Table IV: Exploration data.

Origin	Name	Data Type	Description / Operations
	results of CERAD tests	Enumeration	using distance between CERAD-vectors
	diagnosis	Enumeration	elements of ICD-10; used as selection attribute, exclusion criterion
	results from MRT sequences	String	using terms of defined vocabularity
	Results from Liquor-Tests	String	using terms of defined vocabularity
	Results from Blood-Test	String	using terms of defined vocabularity