Subsystem of prognostic risk factors analysis of childhood acute leukemias

Tom IE[1], Krasko OV[1], Zalessky BA[1], Novoselova NA[1], Suchkova AP[1], Sotikova EE[1], Skrygan NM[1], Potapnev MP[2], Uglova TA[2]

[1] United Institute of Informatics Problems of The National Academy of Sciences of Belarus, Minsk, Belarus
[2] Belarussian Center for Pediatric Oncology and Hematology, Minsk, Belarus

Abstract

Propose: **The purpose was to create the electronic register formed as a database where the results of medical examination for each patient are stored and to develop the software for statistical and intelligent data analysis applied to medical tasks of prediction outcome of induction therapy and patient risk group classification based on prognostic risk factor (PRF) combinations in childhood acute leukemias.**

Material and methods: **Statistical and intelligent methods for data analysis, modern programming technologies for database and software development. The medical data were obtained in Belarussian Research Center for Pediatric Oncology and Hematology (BRCPOH).**

Results: **The subsystem for multivariate statistical and intellectual analysis of data is realized as a special software "Professional Analysis of Prognostic Factors" (ProAPF). ProAPF includes a number of statistical and neural network analysis methods that were selected, upgraded and adopted according to the applied aim of the project.**

Conclusions: **The database for prognostic risk factors (PRFs), collected at diagnosis and during early treatment course in children with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), was created. The information technology of analysis of prognostic risk factors actualized as ProAPF software was developed to identify patient-specific risk group to apply risk group – oriented chemotherapy in patients with ALL or AML. The new PRFs and their combinations were explored, the combinations were selected to predict outcome of induction therapy.**

ADDRESS FOR CORRESPONDENCE:
Igor E. Tom
Laboratory of Bioinformatics
United Institute of Informatics Problems
The National Academy of Sciences of Belarus
6 Surganov Street, 220012, Minsk, Belarus
e-mail: tom@newman.bas-net.by

Key words: data analysis, prognostic risk factors, outcome of induction therapy.

Introduction

The modern information technologies of data collection, data storage and data analysis are effectively used for the different goals, in particular in the health care. Augmented clinical opportunities for specialized medical care for childhood leukemia currently require their selection based on particular sets of parameters, including diagnostic and prognostic factors. Selection the adequate strategy of therapy will result in shorten treatment course to reach recovery, better quality of patient's life, higher rate of survival and, finally, decreased cost of therapy. Sets of parameters to be taken into account are markedly increased and require their computing for medical decision making.

The purpose was to create the electronic register formed as a database where the results of medical examination for each patient are stored and to develop the software for statistical and intelligent data analysis applied to medical tasks of prediction outcome of induction therapy and patient risk group classification based on prognostic risk factors (PRFs) combinations in childhood acute leukemias.

Material and methods

Statistical and intelligent methods for data analysis, modern programming technologies for database and software development were used.

The medical data were obtained in Belarussian Research Center for Pediatric Oncology and Hematology (BRCPOH). Altogether, data from 189 patients with acute lymphoblastic leukemia (ALL) and 74 patients with acute myeloid leukemia (AML) were analyzed. To verify the results of application the analytical system testing sample was obtained in Medical High School Hannover (collaborator of the project Prof. K.Welte).

Results

The subsystem for multivariate statistical and intelligent data analysis is developed as a special software "Professional Analysis of Prognostic Factors" (ProAPF). This is one of the main parts of information analytical medical-oriented system. ProAPF software finds out prognostic factors and their combinations for medical classification and prediction tasks.

The main reason in developing ProAPF is to create modern information technology for discovering combination of various prognostic risk factors (PRFs) for induction therapy response in childhood acute leukemias.

The main goal of ProAPF application is to enhance the efficacy of therapy using the estimation of patient's state based on the prognostic risk factors for selection an adequate intensity of therapy.

ProAPF software actualizes the information technology of joint application multivariate statistical methods and intelligent neural network analysis of data with coherent input/output on the each stage of analysis.

Joint application of statistical and intelligent analysis is necessary for mutual compensation of their intrinsic limitations. The results of neural networks application are implicit, cannot be easy interpreted or modified. But they have low sensitivity to the form of analyzed data, hence they have high generality. The results of application of the statistical analysis are explicit and easy interpreted. The numbers of statistical and intelligent analysis methods are adapted to medical tasks of discovering prognostic risk factors in ProAPF software. Moreover, some special analytic methods were developed for ProAPF. This software can be used for solution the following tasks:

1. **Decrease of the dimension** of clinical and laboratory data **and** its **preprocessing**: digitalization and coding of continuous variables, normalization. This task is provided by univariate analysis (by t-tests, Chi-square tests), principal component analysis (statistical and neural network versions), multiple correspondence analysis, combined algorithm of entropy estimation and maximum-likelihood Chi-square test.

2. **Selection of the combinations** of the most significant PRFs for classification and prediction. This task is provided by stepwise algorithm for PRFs selection in linear regression, exhaustive search on full sets of PRFs, stepwise discriminant analysis based on Mahalanobis distance (stepwise with addition and exhaustive search).

3. **Verification of the selected combinations** by test cases classification and prediction. This task is provided by linear discriminant analysis; multilayer perceptron with different learning algorithms and different normalization type; self-organizing Kohonen networks for classification tasks.

4. **Prediction the outcome of induction therapy** and identification of patient's risk group by his/her individual profile of prognostic risk factors. This task is provided by multiple regression algorithms with exponential and logistic functions, multilayer perceptron for regression tasks, classification CART-trees (decision trees) with automatic and interactive construction type.

We have positive experience in processing of real oncohematological data that approves the efficiency of the developed information technology, especially ProAPF software for comprehensive processing and analysis of heterogeneous (laboratory, clinical, epidemiological) data based on joint using various statistical and intelligent neural network analysis.

The technique for PRFs combinations selection and patient's PRFs profile estimation to predict outcome of induction therapy are developed and applied in BRCPOH in cooperation with medical scientists from BRCPOH. According to the agreement the ProAPF software was successfully tested in Medical High School Hannover, collaborator of the project Prof. K.Welte.

For children with **acute myeloid leukemia** the detection at diagnosis more than 3.5% of apoptotic cells (PI+ cells) in blast population in patient's bone marrow (BM) is one of the most significant favorable PRFs of early response to induction therapy (on day 14 of therapy). The simultaneous identification of <3.5% of apoptotic cells and >70% of initial blasts in PB predict an unfavorable (poor) early response to therapy, estimated on day 14. Besides, the unfavorable PRFs of early response to therapy on day 14 are initial content of BM blasts of >80%, initial blasts in peripheral blood (PB) $>70{\cdot}10^9$/l as well as a high amount of CD34+ blasts (>70%) and HLA-DR+ blasts <70% in bone marrow. Significant PRFs for unfavorable treatment response to therapy on day 28 in children with AML are the combination of >10% blasts in BM (day 14-15) sample with the initial WBC of $>70\times10^9$/l and with a large number (>60%) of CD7+ blasts and of t (8;21). Unfavorable response on day 42 of therapy can only be associated with initially defined cytogenetic abnormalities in BM blasts – rearrangement 11q23 without t (9;11).

For children with **B-lineage acute lymphoblastic leukemia** the PRFs for unfavorable response to induction therapy (day 33 – 36 of protocol) are gender (boys), the presence of blasts in BM on day 15 of therapy >10%, initial count of CD2+ blast cells in BM>10%. The significant PRFs for early response to induction therapy (on day 14-15 of protocol) are gender, age at the moment of diagnosis, (more or less than 10 years old) in combination with initial WBC, WBC and blasts' count on day 8 of therapy, bone pains, initial count of CD 34+ blast cells in BM. Among PRFs the unfavorable early response to induction therapy for girls was related to: age >12 years, associated with the initial WBC of $>20\times10^9$/l, WBC count on day 8 of therapy $>2\times10^9$/l, initial count of blasts PB >70%, the persistence of blasts in PB on day 8 of treatment >10%, initial high level (>85%) of CD34+cells and the simultaneous identification >10% cyIgM+ and ≤10% sIgM+ blasts of in patient BM sample. Among boys the PRFs of unfavorable early response to therapy were: initial blasts in PB>70%, WBC count $>7\times10^9$/l on day 8 of therapy, the persistence of $>1\times10^9$/l of blasts or >1% of blasts in PB on day 8 of treatment, initial content of >10% CD2+ cells and low level CD20+ blasts in BM, hepatomegaly after day 8 of therapy.

The PRFs of unfavorable response to the induction therapy (on day 33 of protocol) in children with T-lineage acute lymphoblastic leukemia were: the persistence of >10% of blasts in BM on day 15 of therapy or combination the next factors – the presence of blast cells and especially >20% CD117-expressing and >30% CD34-expressing BM cells at diagnosis, as well as

the blast count of $>1\times10^9$/l on day 8 of therapy. The PRFs of unfavorable early response to induction therapy (day 15) are: blasts content $>1\times10^9$/l in PB on day 8 therapy, the absence of CD10-expressing BM cells. Myelocoexpression (CD13) on BM blast cells is also an unfavorable PRF for early treatment response (day 15) the children with T-lineage ALL.

Conclusion

1. The information technology of analysis of prognostic risk factors actualized as ProAPF software was developed to identify patient-specific risk group to apply risk group – oriented chemotherapy in patients with ALL or AML.

2. The technique of accounting the individual profile of patient-specific PRFs and their association with response to induction therapy at critical timepoints for patients with ALL or AML was elaborated using statistical and intelligent neural network analysis methods.

3. It was estimated, that for children with AML detection at diagnosis more than 3.5% of apoptotic cells (PI+ cells) in blast population in patient't bone marrow (BM) is one of the most significant favorable PRFs of early response to induction therapy (on day 14 of therapy). The simultaneous identification of <3.5% of apoptotic cells and >70% of initial blasts in PB predict an unfavorable (poor) early response to therapy, estimated on day 14. Unfavorable response on day 42 of therapy was associated with initially defined cytogenetic abnormalities in BM blasts – rearrangement 11q23 without t (9;11).

4. For children with B-lineage ALL the PRFs for unfavorable response to induction therapy (day 33-36 of protocol) were gender (boys), the presence of blasts in BM on day 15 of therapy >10%, initial count of CD2+ blast cells in BM>10%. The significant PRFs for early response to induction therapy (on day 14-15 of protocol) are gender, age at the moment of diagnosis, (more or less than 10 years old) in combination with initial WBC, WBC and blasts' count on day 8 of therapy, bone pains, initial count of CD 34+ blast cells in BM.

Acknowledgement

The financial support of European Community under the contract to The International Science and Technology Center (Project #B-522) is greatly acknowledged.