Applied Medical Informatics
Original Research Vol. 28, No. 1/2011, pp: 16-22

Software Application for Data Collection and Analysis in Acute Myeloid Leukemia

Anca BACÂREA[1]*, Bogdan Adnan HAIFA[2], Marius MUJI[3], Alexandru ŞCHIOPU[1]

[1] University of Medicine and Pharmacy Târgu Mureş, Physiopathology Department, 38 Gh. Marinescu str., Tîrgu Mureş, Romania.
[2] University of Medicine and Pharmacy Târgu Mureş, Medical Informatics and Biostatistics Department, 38 Gh. Marinescu str., Tîrgu Mureş, Romania.
[3] University Petru Maior Târgu Mureş, Engineering Department, 1 Nicolae Iorga str., Tîrgu Mureş, Romania.
E-mail(s): bacareaa@yahoo.com, marius_muji@yahoo.com, bogdan_adnan@yahoo.com

* Author to whom correspondence should be addressed; Tel.: +4-0744325077.

Received: 12 December 2010 /Accepted: 21 February / Published online: 7 March 2011

Abstract

Aim: It is important in the context of the informatics development and also of medical research, that new software technology to be integrated in order to achieve easier research. The aim of this study was to develop a software application that uses few resources, and that enable data collection, their primary processing in statistical terms (e.g. mean, median, etc.), drawing of survival curves and survival Log Rank statistic testing according to the collected parameters. *Material and Method:* For this purpose, a database in SQLite3 was developed. Because the database engine is embedded in the Database Management System (DBMS) this program allows absolute portability. Graphical interface was made in wxWidgets. Statistical calculations were obtained using R software (the `addons` E1071 was used for descriptive statistics and the `Survival `for testing survival and Northest for Kaplan Meier survival curve). Patients were cases admitted and treated in the Hematology Department of County Emergency Hospital Tîrgu Mureş hospitalized and treated during 2007-2010. *Results*: We created a GUI in wxWidgets to collect the desired medical data: age, date of diagnosis, date of death, blood count values, and the CD leukocyte markers detected by flow cytometry. Entwining of medical data collection and processing statistics (for acute myeloid leukemia - survival, prognostic factors evaluation) is a further step in medical research. *Conclusion*: The tool presented is a useful for research. Application in acute myeloid leukemia derives from the author's interest in the subject; development of this tool in other directions is possible and desirable.

Keywords: Acute myeloid leukaemia; Database; Survival.

Introduction

It is important in the context of the informatics development and also of medical research, that new software technology to be integrated in order to achieve easier research. An important aspect of conducting a survey is to collect data. Data provided by the patient or taken from existing records, in most cases are handled directly by physically marking them on paper because of the still limited portability of computers issued by financial constrains. Adding patients (cases) in the

database of the study is thus a two phase process, which implies a lot of time allocated in data recording.

In the same context is the need for preliminary statistical evaluation of new data, from the evaluation of distributions in the sense of normality, to the advanced statistical evaluation of changes.

Acute myeloid leukemia is part of a group of diseases where the collected data have a distinctive character [1]. Normality tests regarding the blood analyses (e.g. leukocytes number) is an issue to be evaluated in the context of the disease. Survival evaluation after diagnosis is also crucial for disease management [2,3].

Bringing together these issues (software) in one form (application) in order to achieve efficiency in data collection and initial evaluation of data is the goal of this work.

The aim of this study is to develop a software application that uses few resources, and that enable data collection, their primary processing in statistical terms (e.g. mean, median, etc.), drawing of survival curves and survival Log Rank statistic testing according to the collected parameters.

Material and Method

We started from the premise that using a PDA or a smart phone can replace data collection using paper or laptops (with reduced independence anyway), in order to collect and primary process statistically the data and also to evaluate prognostic parameters in terms of survival.

For this purpose, we created a database in SQLite3 [4]. Because the database engine is embedded in the Database Management System (DBMS) this program allows absolute portability. There is no need to connect to servers; it requires no LAN, WAN connection. Other benefit refers to low memory consumption, representing an optimal variant for a future written application for PDA or Smartphone.

Graphical interface was made in wxWidgets [5]. This library was chosen for its cross platform type. Thus it allows the writing of a unique code for Windows operating systems, Linux or Macintosh. It is a library in a continuous process of development.

Statistical calculations were obtained using R program, using the `*addons*` E1071 for descriptive statistics, `*Nortest*` for testing normality (function lillie.test) and `*Survival*` for testing survival and for drawing Kaplan Meier and Log Rank survival curve [6].

The language used was C++, using the programming environment Code Blocks v. 10.5 on an Ubuntu10.10 operating system.

Patients were cases admitted and treated in the Hematology Department of County Emergency Hospital Targu Mures hospitalized and treated during 2007-2010. The authors decided that an exhaustive study reveals better the survival then a sample based study. Consecutively considering also the low incidence of acute myeloid leukemia all patients were included (75 adult patients, 29 females and 46 males).

Results

We created a GUI in wxWidgets to collect the desired medical data. We collected the following information regarding: age, date of diagnosis, date of death, blood count values, the CD leukocyte markers detected by flow cytometry (Figure 1).

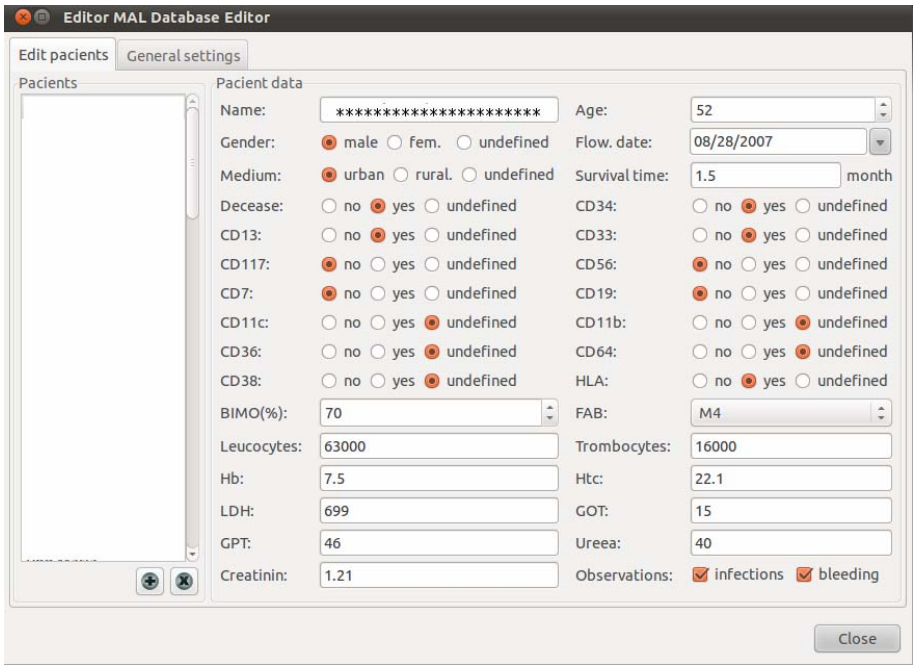

Figure 1. DBMS screenshot for patients with acute myeloid leukemia

In this window you can enter and edit in real-time database cases, changes made are saved in real time and statistical results are automatically updated. Standardizing the entry modes of data do not allow incomplete data entry or data to interfere with subsequent processing of data. For binary data, it will also enable the introduction of binary values and also a variant with lack of value.

In order to add new patients or to delete them from the database (counting unreliable or incomplete) the following buttons are used: .

For safety deletion (involuntary loss) program displays a confirmation window to do it.

To classify the type of leukemia in the FAB system it uses a list of options that is forcing the operator to properly mark the classification in a class.

In the "survival time" field there is automatically calculated the survival time since diagnosis to the entry in the database ("Data Flow" is the time of diagnosis) and the reference date is current date. If the patient is deceased recorded survival time value is inserted manually.

For assessment of and testing for survival according to blood values at diagnosis is needed binarization of quantitative results. This operation is specifically permitted in the program presented by changing the threshold value of binarization. It is known that there is no threshold value, for which the survival prognosis can be said to be good or bad. In order to clarify these challenges required the program provides the solution (Figure 2).

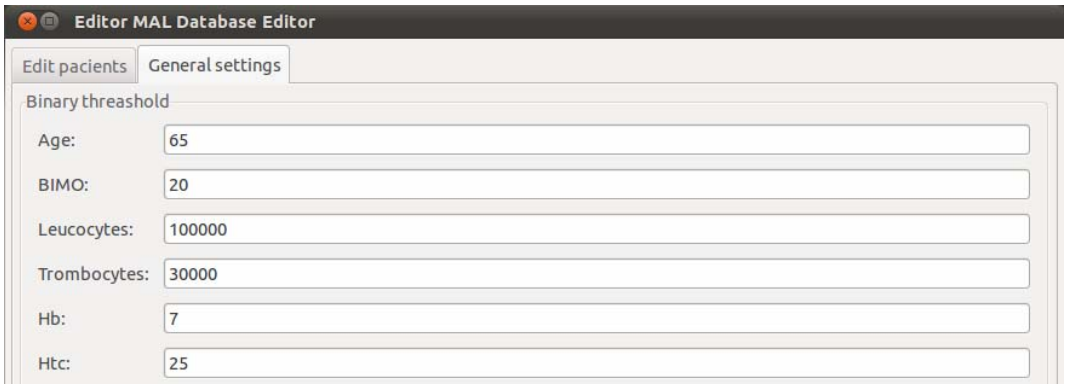

Figure 2. Menu "General Settings" for establishing threshold values for laboratory measurements

Entering a new patient in the database is done with all laboratory or demographic data. Missing values are designated as "-1" (Figure 3).

Given the fact that saving is done by default, there is no special button to save. Statistical engine automatically calculates descriptive statistics and Kaplan Mayer survival curve and test the difference in survival (LogRank test).

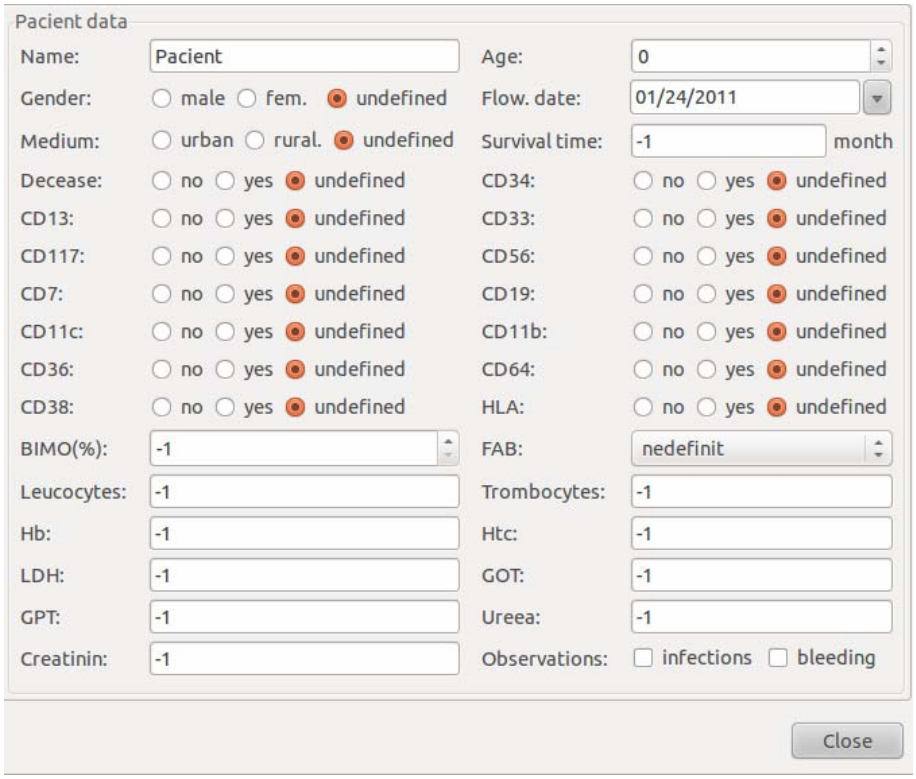

Figure 3. Entering a new patient in the database

Calculation of descriptive statistics parameters is needed to assess real-time values entered. It always puts the problem of outliers. This program allows the evaluation of real-time distribution (Kolmogorov Smirnof normality test). This option allows reanalysis of outliers (erroneous laboratory value, bad techniques for sample collection, contamination of the sample, the erroneous entry in the hospital's computer system, etc.). Thus, further work of "brushing" data can be avoided (Figure 4).

Figure 4. Statistics and normality testing parameters for selected variables

Real-time evaluation of the survival of patients with acute myeloid leukemia may be considered less relevant. However, in view of the low incidence of this and especially low survival (approximately 7 months) any entered case may affect the overall survival. There are many cases

with survival of only a few days or less. Assessment in terms of survival is discharge after an appropriate time and only after an initial assessment of it (Figure 5).

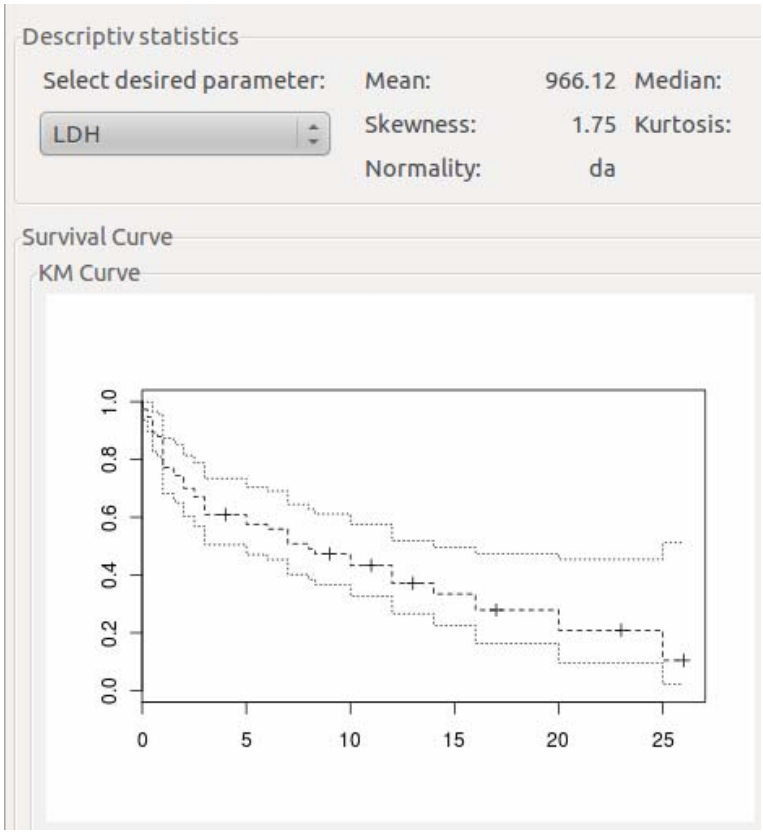

Figure 5. Automatic survival curve (Kaplan Mayer)

Variable "time" is automatically updated so that after updating database status "dead / alive", survival periods will increase. The only obligation of the researcher remains to update the database periodically regarding survival time if one of the patients died.

Testing picked parameters as prognostic factors in terms of survival is done by using the interface. For this purpose the desired parameters are selected. If the measured parameter is a continuous numerical variable, it will be entered like binary variable (as presented above). To evaluate the parameters which were entered as binary variables we directly select them from the options list (Figure 6).

In the example shown is tested in terms of survival the parameter 'age'. Being represented by a continuous numerical variable it was binarized according to the threshold value of 65. Survival curves are presented and the value returned by the statistical test is $p = 0.3368$. For comparison is shown in Figure 6 the curve obtained by the same test using the "R".

Discussion

There are several medical informatics systems for data collection. "Microsoft Office" offers "Access" [7]. This medical data collection tool allows the use of "form" sites that secures data in terms of their unitary character, like the program presented. Also the use of tables for data storage is done on the workstation. The benefits of our presented program are the "crossplatform" character- can be installed in multiple operating systems (Linux, McIntosh, Windows), and low consumption of resources (you can use your PDA or Smartphone). At the same time, it allows a

rapid assessment of the descriptive statistical parameters, but also of advanced statistical processing (survival test).

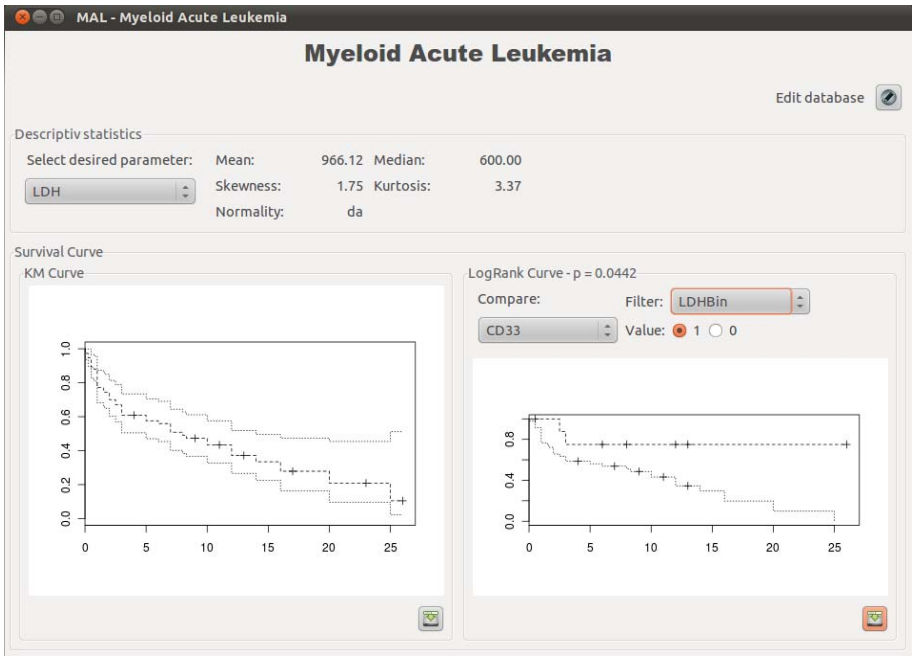

Figure 6. Survival curves and logrank test application

Another system of database management is "the application of medical data storage and management [8] which enables data security, placing them in a tree form, more records management for the same patient without deleting the original data collection, multicenter study. It allows only the use of it in network systems and WAN or Internet connection, it requires a server for data storage and a secure Internet connection for data replication.

The statistic in the "R" is powerful. Its use, however, requires a good knowledge of command line programming. Acquisition of routines / procedures called from the program allows the use of its statistical processing power with customized interface [9].

It can be further improved by adapting the graphical interface to run on a handheld PDA / Smartphone. Data processing code is compatible with such a system, the only change being a graphical interface that takes into account the small screens of these devices.

The shift in the field of biostatistics and medical informatics, to use dedicated software is a general goal in Romanian and international context [10]. Calling on the resources open source (Linux) allows, for the need of informatization and computerization of the health system, reducing costs. In research using such programs generate: lower cost, data classification (no Internet connection), faster data processing, etc.

Conclusions

The tool presented is a useful tool for research. Application in acute myeloid leukemia derives from the author's interest in the subject; development of this tool in other directions is possible and desirable.

Using devices with limited memory and / or low power processors (PDA / Smartphone) gives an absolute portability character.

It confers the privacy necessary for medical data activity (no need Internet connection, server data storage, etc.).

Entwining of medical data collection and processing statistics (for acute myeloid leukemia - survival, prognostic factors evaluation) is a further step in medical research.

Conflict of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

Anca Bacarea carried out the investigation of patients and participated in the design of this application. Bogdan Haifa and Marius Muji participated in the design and development of the applications. Alexandru Schiopu coordinated and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This paper is partially supported by the Sectoral Operational Programme Human Resources Development, financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/89/1.5/S/60782.

References

1. Röllig C, Thiede C, Gramatzki M, Aulitzky W, Bodenstein H, Bornhäuser M, et al. A novel prognostic model in elderly patients with acute myeloid leukemia: results of 909 patients entered into the prospective AML96 trial. Blood 2010;116(6):971-978.
2. Betz BL, Hess JL. Acute Myeloid Leukemia Diagnosis in the 21st Century. Arch Pathol Lab Med 2010;134:1427-1433.
3. Li X, Li J, Du W, Zhang J, Liu W, Chen X, Li H, Huang S, Li X. Relevance of immunophenotypes to prognostic subgroups of age, WBC, platelet count, and cytogenetics in *de novo* acute myeloid leukemia. APMIS 2011;119(1):76-84.
4. *** SQLite [online] [cited 2010 December]. Available from: URL: www.sqlite.org.
5. *** wxWidgets Cross platform GUI library. [online] [cited 2010 December]. Available from: URL: http://www.wxwidgets.org/
6. *** R project for statistical computing [online] [cited 2010 December]. Available from: URL: http://www.r-project.org/.
7. Muji M, Ciupa RV, Dobru D, Bica C, Olah P, Bacarea V, Marusteri M. Database Design Patterns for Healthcare Information Systems. International Conference on Advancements of Medicine and Health Care Through Technology 2009;26:63-6.
8. Bacârea V, Călinici T. Application of Complete Storing and Selection of Medical Data. Applied Medical Informatics 2004;14(1-2):42-7.
9. Bacârea V, Călinici T, Moldovan C, Sabău H, Măruşteri M. The Level of Computer Use and the Degree of the Acceptance of Informatics Applied in Medical Research. Applied Medical Informatics 2008;23(3-4):63-7.
10. Călinici T, Munteanu V, Bacârea V. DIASTAL - a Client Server Database Application for Laparoscopic Diagnosis and Staging in Abdominal Cancers. Applied Medical Informatics 2005;17(3-4):68-71.