

Matchings and Decision Trees for Determining Optimal Therapy

Natalia Korepanova¹(✉), Sergei O. Kuznetsov¹,
and Alexander I. Karachunskiy²

¹ School of Applied Mathematics and Information Science,
National Research University Higher School of Economics, Moscow, Russia
korepanova.natalia@gmail.com, skuznetsov@hse.ru

² Research and Clinical Center of Pediatric Hematology,
Oncology and Immunology, Moscow, Russia
aikarat@mail.ru

Abstract. An approach to the study of different types of treatments in subgroups is proposed. This approach is based on matching algorithms and decision trees. An application to the data on children with acute lymphoblastic leukaemia is considered.

Keywords: Medical informatics · Decision trees · Optimal therapy · Machine learning for medicine

1 Introduction

Nowadays one of the most promising way of therapy optimization, especially in pediatric haematology, is conforming a therapy to various subgroups of patients which are described by patients' physiological features. Usually, the number of possible subgroup descriptions is large, and often physicians are able to chose subgroups for analysis relying just on their experience and observations. Therefore, in statistical terms any subgroup analysis which is aimed at showing the superiority in efficiency of one treatment strategy over one or several others seems doubtful as a rule. In the present paper the approach of finding subgroups with significantly different or equivalent response to two treatment strategies is proposed. Obtained hypotheses can underlie subgroup analysis with better choice of subgroups.

The analysis is carried out for the database on children with acute lymphoblastic leukemia (ALL) [1] who underwent a course of one of two types of induction therapy. The first step consists in finding the largest set of pairs of similar patients who took different drugs with the help of the Gale-Shapley algorithm [2–4] for computing optimal stable matching. This algorithm is based on the concept of physiological “similarity” between two patients; therefore, the definition of “distance” between two patients is introduced. After that the derived matching is examined for the existence of the classes in which treatment strategy strongly affects or does not affect treatment response. At the second step

we attempt to describe extracted classes by applying decision trees with various parameters [5, 6]. It appears that decision trees are not able to describe every class on the whole. However, some subgroups of patients for whom one drug is more appropriate than another one can be selected from the results of classification. Moreover, decision trees allow us to formulate hypotheses about comparison of therapy efficiencies in subgroups in form suitable for haematologists. And finally, results are approved or disapproved by classical medical statistical methods.

The rest of the paper is organized as follows. In Sect. 2 the dataset is described. In Sect. 3 the steps of proposed approach are presented in detail. In Sect. 4 the application to the initial dataset and its results are shown. Section 5 concludes the paper.

2 Dataset

The dataset consists of 1946 patients up to 19 years old of age with newly diagnosed acute lymphoblastic leukaemia (ALL). This dataset is stored as a database containing the following data fields: sex (male or female), age (in years), initial white blood count (per nl) (WBC), immuno-phaenotype (8 types), CNS status (3 types), palpable liver size (in cm), palpable spleen size (in cm), mediastinum status (3 types), date of allocation to treatment, last status report (alive, no information, death), date of the latest follow up visit, treatment strategy (2 types).

The analysis was based on the comparison of the efficiency of two treatment strategies: under DEXA $6 \text{ mg/m}^2/\text{d}$ and MePRED $60 \text{ mg/m}^2/\text{d}$ [1], which we call S1 and S2. To find relations between initial characteristics and survival rate all physiological features presented above were chosen: sex, age, initial WBC, immuno-phaenotype, CNS, palpable liver size, palpable spleen size, mediastinum status. To evaluate the therapy efficiency overall survival [7] was calculated with death as the event. Survival time was calculated from diagnosing until the date of last status report. If the value of one patient's characteristic was not determined, this patient was not included in analysis. Consequently, we obtained data on 1535 appropriate patients: 939 of them were assigned S1, and 596 of them were assigned S2.

3 Proposed Approach

Our procedure is intended to find subgroups where differences between two competing treatment strategies are noticeable or do not exist. The input data consist of two sets of patients corresponding to the strategies. All patients are described by several initial features each could be either numerical or categorical. The number of features can be reduced applying selection feature techniques, or relying on expert views. In the current research we were provided with information on 8 features which are the most influential in haematologists' sight, therefore feature selection techniques were not required. Nevertheless it can surprisingly appear that unprovided features affect intensively the survival time. So, as for medical

data, we believe that results of feature selection and expert views should be combined. Let us now move on to the steps of the proposed procedure.

3.1 Patient Distance

First of all, the procedure suggests defining distance between two patients (the inverse concept to “physiological similarity”). If all physiological features are numerical, it is possible to use one of the classical distance measures [8–10]. However, there are likely several categorical features in patient descriptions. Therefore it is required to modify classical definitions of the distance. It is considered that there is no sense to measure distance between two patients that cannot be compared. For this purpose it is necessary to give a definition of “comparability” of two patients.

Definition 1. *Two patients are **comparable** if the values of all their categorical physiological features coincide. If they differ in just one categorical feature, they are **incomparable**. So, if all initial features of patients are numerical, any two patients are comparable.*

Before computing distance between comparable patients, numerical features need to be normalized to make all feature impacts equivalent. Therefore all of them are centered by subtracting the mean value and then scaled by dividing by the range of values [11]. So, the distance between two comparable patients is computed based on the normalized values of their numerical features.

3.2 Pairs of Similar Patients

The distances between all pairs of patients where the first one is from the first set and the other one is from the second set are computed. To find pairs of similar patients *the deferred acceptance procedure* [2–4] is applied for two sets of patients who underwent different courses of treatment. This algorithm was developed to solve the marriage problem, i.e. the problem of finding stable matching. It is applicable to two sets of instances which are often referred to as men and women. Every man ranks women and every woman ranks men in accordance with their preferences. Thereupon each man proposes to his favourite woman, and each woman rejects all but her favourite, who becomes her marriage nominee. The rejected men propose to their next choices, and each woman chooses her favourite among the new proposers and the nominee rejecting all the rest, and so on. As soon as no men are rejected or they have no more choices each woman accepts her nominee. Eventually, we get the pairs consisting of one man and one woman. In other words, every pair includes two instances from different sets. The result of the algorithm application is stable, and optimal if preferences are complete. In case of one-to-one matching completeness also accounts for uniqueness of the provided matching. Before applying this algorithm to medical data definitions of preference and completeness should be given.

Definition 2. Patient p **prefers** patient q_1 to patient q_2 if the distance between p and q_1 is less than the distance between p and q_2 . Patient p is **indifferent** between patients q_1 and q_2 if the distances between p and q_1 and between p and q_2 are equal.

Definition 3. Preferences are **complete** if for every x, y from one set and for every z from the other one, z prefers x to y , or y to x , or it is indifferent between them.

Some patients may be incomparable w.r.t preference. However, it results from the definition of comparability that all patients can be partitioned into subsets where all patients are comparable with each other. Consequently, in these subsets patients preferences are complete, and the result of the *deferred acceptance procedure* application to every such subset is unique, stable and optimal [4], which means that the constructed matching is unique, stable and optimal in total.

3.3 Separation into Classes in Terms of Efficiency (Overall Survival)

Using the matching and patients' survival times we can determine classes of patients with quite clear or without any dissimilarities in survival time under treatment strategies. The simplest approach is to visualize the matching in any way and attempt to mark boundaries of such classes. Therefore, it is proposed to consider the coordinate plane where X-axis is survival time under the first curing strategy and Y-axis is survival time under the second one. Every pair in the matching is associated with a point on the plane. The first coordinate of the point is the survival time of the patient who has received the first kind of treatment, and the second one is the survival time of the patient who has received the other one.

All points (pairs of patients) are partitioned into several classes according to sensitivity to treatment strategies (e.g., survival time under treatment strategy 1 is superior to that under treatment strategy 2, survival time under treatment strategy 2 is superior to treatment strategy 1, short survival time under both treatment strategies, and long survival time under both of them). Further manipulations are carried out on data about individual patients.

3.4 Hypotheses Generation and Verification

It is insufficient to separate patients into classes in which survival times under different strategies differ or do not differ. It is more essential to obtain descriptions of these classes, so the classification problem arises. In the case of comparing treatment strategy efficiencies decision trees with various parameters [5, 6] seem appropriate because, in general, the accuracy of the other well-known methods is lower on the initial data. Moreover, the form of hypotheses generated by decision trees is comprehensible for physicians. So, in our computer experiments we used information gain, information gain ratio and Gini index as attribute selection criteria [12, 13]. Also, minimum number of instances in leaves, maximal allowable

tree depth and sufficient percent of majority class for nonsplitting were varied. The approach evaluation was conducted by means of 10-fold cross-validation.

Decision trees output data are descriptions of classes in terms of characteristics of the patients belonging to these classes. Those descriptions may be transformed into hypotheses about the existence or the absence of the difference in treatment strategy efficiencies. To show how it works, assume that any description of the class with the superiority of the first treatment strategy has been received. This assumption can be transformed into the following hypothesis: for the patients who fall under obtained description overall survival under the first strategy is higher than that under the second one. This sort of hypotheses are put forward on the basis of the most evident subgroups output by decision trees.

All formed hypotheses are tested by classical medical statistical tools. The first of them is Kaplan-Meier survival curves [14–16] which estimate sample survival rate functions for censored data. The second one is log-rank test [16, 17], a nonparametric hypothesis test to compare the survival distribution of two samples with no-difference null hypothesis and standard normally distributed statistics. In contrast to log-rank, the equivalence test [18–20] is applied to confirm that survival rates of two samples do not differ, and usually used if log-rank null hypothesis has not been rejected. Also for each hypothesis false negative error (type II error) is computed [21]. It is important to mention that hypotheses are tested on the set of patients that consists not only of those who have been included in classification, but also of those who have not been matched with anybody or have not been labeled with any class mark. If a hypothesis is confirmed by the tests and false negative error is not very large, then it can be analysed by physicians in further random trials. The necessity of new trials is specified by the worldwide statistical principles of clinical trials. According to the notes of European Medicines Agency [22], any clinical trials may have two aspects: confirmatory and exploratory ones. For the first of them the hypotheses are pre-defined, and are tested when the trial is complete, while the second aspect allows of the data dependent choice of hypotheses, and the ability of changes in response to accumulating results. Obviously, the proposed procedure is intended for the exploratory aspect of trials, therefore its results “cannot be the basis of the formal proof of efficacy” [22]. However, the exploratory investigations can serve “for suggesting further hypotheses for later research” [22]. The last statement clearly explains the main purpose of the proposed approach.

4 Analysis and Results

According to the proposed approach it is necessary to distinguish between numerical and categorical physiological features. The initial dataset contains four numerical features: age, initial WBC, palpable liver size, palpable spleen size; and four categorical features: sex, immuno-phaenotype, CNS, mediastinum status. The second ones determine the comparability of patients. However, immuno-phaenotype has two levels of categorization: B- or T-ALL (nominal values), and each of these types has four ordinal subcategories. Therefore, the condition of

Table 1. Numerical values of immuno-phaenotype.

B	-1
pre-B	-0.75
common-B	-0.5
pre-pre-B	-0.25
early-T	0.25
intermediate-T	0.5
mature-T	0.75
hybrid	1

comparability was weakened for this feature. Thus, two patients were comparable if both of them had B- or T-ALL and values of all other categorical features coincided.

To take into account the second level of immuno-phaenotype categorization, its values were transformed into numerical values according to Table 1.

So, for child-ALL data distance was computed using normalized values of numerical characteristics and numerical values of immuno-phaenotype. As it is mentioned before, any standard definition of distance can be chosen. Therefore, Manhattan [8,9], Euclidean [8,9], Minkowski [8,10] with factor 3, Minkowski with factor 100 and Chebyshev [8] distances were used in computer experiments. We found out that there was no meaningful difference between these measures, so for further analysis and method specification Euclidean distance was used.

After that the deferred acceptance procedure was applied, the scatter plot of the derived matching is shown in Fig. 1. There were about 5 most evident isolated classes of points (Fig. 2). As for classes 1 and 2 survival time under both strategies was not long. However, it was slightly longer under S1 in class 1 and slightly longer under S2 in class 2. We can also say that for class 3 the survival time under S1 was longer than under S2 and vice versa for class 4. The survival time for class 5 was long under both strategies. We attempted formalizing boundaries in the way shown in Fig. 3. It is important to mention that the 4-years boundary was not selected randomly. There are about 4 years between the latest diagnosing date and the latest last status report date. In other words, if survival time of any patient was shorter than 4 years this patient was certainly dead or escaped from the observation. For the sake of clear separation of the classes the points between dashed lines were excluded from the analysis.

By applying decision trees to all presented partitions we obtained several hypotheses, one of the most reliable hypothesis is presented below.

The hypothesis is “*MePRED is more efficient than DEXA for patients who are equal to or more than 6.6 years old, with palpable spleen size not smaller than 3.5 cm, and pre-pre- or pre-B immuno-phaenotype*”. There were 39 patients of that kind for classification and 47 such patients at all. This subgroup is not numerous, but, at first, Kaplan-Meier curves (Fig. 4) seem to confirm the hypothesis. The value of log-rang statistics is equal to 2.12 which allows one to reject

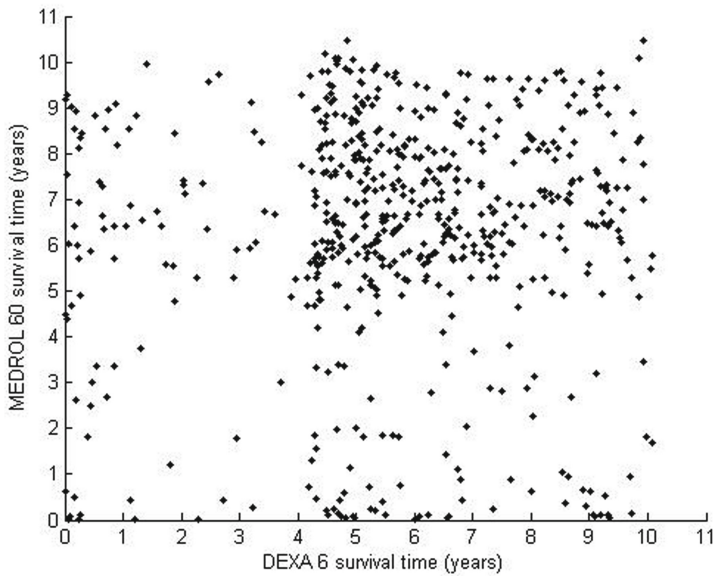


Fig. 1. Scatter plot of all matching pairs.

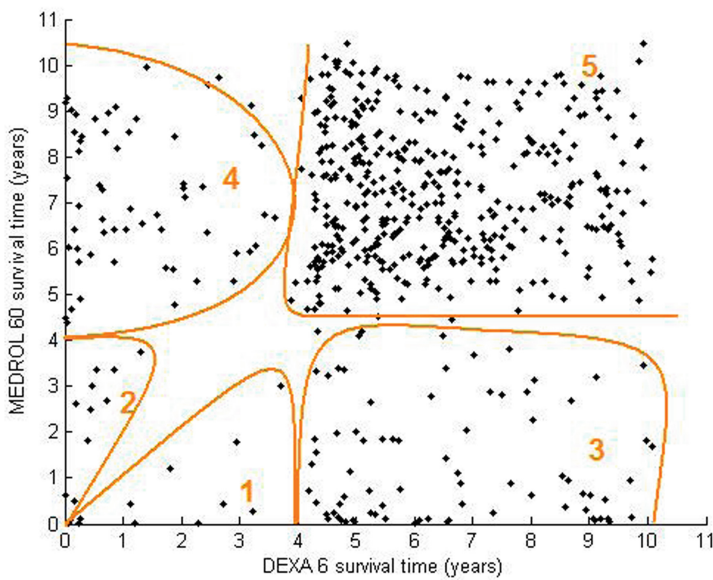


Fig. 2. Scatter plot of all matching pairs with outlined classes.

the hypothesis about no difference at confidence level of 0.95. The false negative error amounts to 0.31. This is quite good, so, we can propose this hypothesis to test in further clinical random confirmatory trials.

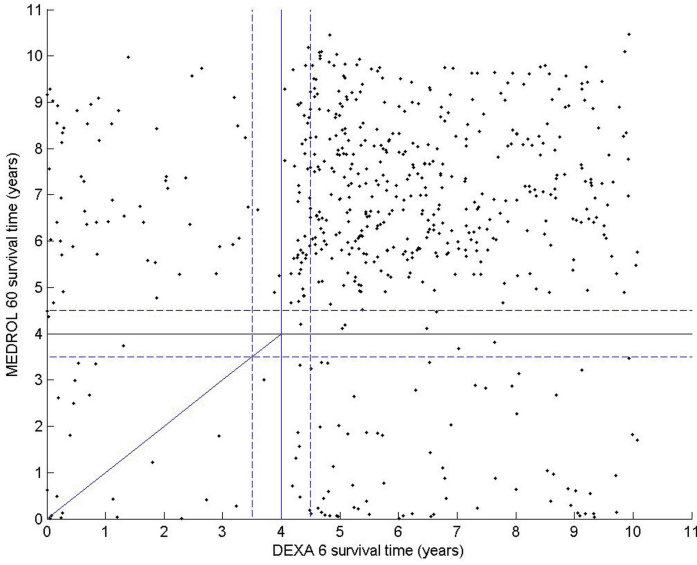


Fig. 3. Scatter plot of all pairs with partition into 5 classes.

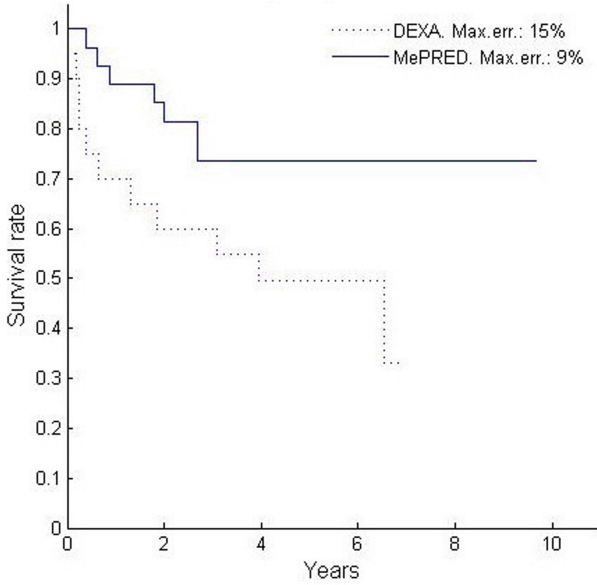


Fig. 4. Kaplan-Meier curves for the case of 6.6 years old and older patients with palpable spleen size not smaller than 3.5 cm and pre-pre- or pre-B immuno-phenotype.

5 Conclusion

In this paper we introduced a novel approach to solving the problem of determining relevant subgroups of patients for therapy optimization. Getting the dataset of patients described by their physiological characteristics, dates of diagnosing and last status report, the procedure constructs the optimal stable matching between patients who took different drugs and attempted to describe subgroups in which the efficiency of these drugs are different or approximately equal. In further studies other learning techniques will be used, in particular, those based on closed descriptions [23–25].

The proposed procedure can also be applied in other studies of subgroup analysis. Moreover, all parts of the procedure are flexible to changes and can be adapted to other practical problems of subgroup analysis. The main idea of this work consists in proposing the order in which data analysis techniques can be applied, and how they can influence any therapy optimization. Hopefully, the obtained hypotheses will be successfully used in Russian ALL-treatment studies.

References

1. Karachunskiy, A., Herold, R., von Stackelberg, A., et al.: Results of the first randomized multicenter trial on childhood acute lymphoblastic leukaemia in Russia. *Leukemia* **22**, 1144–1153 (2008)
2. Gale, D., Shapley, L.S.: College Admissions and the Stability of Marriage. *Am. Math. Mon.* **69**(1), 9–15 (1962)
3. Roth, A.E.: Differed acceptance algorithm: history, theory, practice, and open questions. *Int. J. Game Theory* **36**(3–4), 537–569 (2007)
4. Alkan, A., Gale, D.: Stable schedule matching under revealed preference. *J. Econ. Theory* **112**, 289–306 (2003)
5. Fürnkranz, J.: Decision tree. In: Sammut, C., Webb, G.I. (eds.) *Encyclopedia of Machine Learning*, pp. 263–267. Springer, New York (2010)
6. Rokach, L., Maimon, O.: Classification trees. In: Maimon, O., Rokach, L. (eds.) *Data Mining and Knowledge Discovery Handbook*, 2nd edn, pp. 149–174. Springer, New York (2010)
7. NCI Dictionary of Cancer Terms. <http://www.cancer.gov/dictionary?cdrid=655245>. Accessed 7 March 2014
8. Deza, M.M., Deza, E.: *Encyclopedia of Distances*, pp. 94, 323–324. Springer, Heidelberg (2009)
9. Shekhar, S., Xiong, H.: Distance measures. In: *Encyclopedia of GIS*, p. 245. Springer, New York (2008)
10. Fuhr, B.: Distance and similarity measures. In: *Encyclopedia of Multimedia*, pp. 188–189. Springer, New York (2008)
11. Mirkin, B.G.: *Core Concepts in Data Analysis: Summarization, Correlation, Visualization*. Springer, London (2011)
12. Raileanu, L.E., Stoffel, K.: Theoretical comparison between the Gini Index and Information Gain criteria. *Ann. Math. Artif. Intell.* **41**, 77–93 (2004)
13. Kotsiantis, S.B.: Decision trees: a recent overview. *Artif. Intell. Rev.* **39**, 261–283 (2013)

14. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**(282), 457–481 (1958)
15. May, W.L.: Kaplan-Meier survival analysis. In: Schwab, M. (ed.) *Encyclopedia of Cancer*, pp. 1590–1593. Springer, Heidelberg (2009)
16. Kleinbaum, D.G., Klein, M.: Kaplan-Meier survival curves and the log-rank test. In: Kleinbaum, D.G., Klein, M. (eds.) *Survival Analysis*, pp. 55–96. Springer, New York (2012)
17. Beyersmann, J., Schumacher, M., Allognol, A.: Nonparametric hypothesis testing. In: *Competing Risks and Multistate Models with R*, pp. 155–158. Springer, New York (2012)
18. Piaggio, G., Elbourne, D.R., Altman, D.G., Pocock, S.J., Evans, S.J.W., for the CONSORT Group: Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA* **308**(24), 2594–2604 (2012)
19. Machin, D., Gardner, M.J.: Calculating confidence intervals for survival time analyses. *Brit. Med. J.* **296**, 1369–1371 (1988)
20. Goberg-Maitland, M., Frison, L., Halperin, J.L.: Active-control clinical trials to establish equivalence or noninferiority: methodological and statistical concepts linked to quality. *Am. Heart J.* **146**(3), 398–403 (2003)
21. Glanz, S.A.: *Primer of Biostatistics*, 7th edn. McGraw-Hill Education, New York (2011)
22. ICH Topic E9: Statistical Principles for Clinical Trials. Step 5. (2.1 Trial Context.), pp. 6–7 http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf. Accessed 7 March 2014
23. Ganter, B., Kuznetsov, S.O.: Hypotheses and version spaces. In: Ganter, B., de Moor, A., Lex, W. (eds.) *ICCS 2003. LNCS (LNAI)*, vol. 2746. Springer, Heidelberg (2003)
24. Blinova, V.G., Dobrynin, D.A., Finn, V.K., Kuznetsov, S.O., Pankratova, E.S.: Toxicology analysis by means of the JSM-method. *Bioinformatics* **19**(10), 1201–1207 (2003)
25. Ganter, B., Grigoriev, P.A., Kuznetsov, S.O., Samokhin, M.V.: Concept-based data mining with scaled labeled graphs. In: Wolff, K.E., Pfeiffer, H.D., Delugach, H.S. (eds.) *ICCS 2004. LNCS (LNAI)*, vol. 3127, pp. 94–108. Springer, Heidelberg (2004)