Initial Data-Driven Model for Estimating Impact of Antihypertensive Drug Amount on Blood Pressure Lowering*

Anna Semakova[1][0000−0002−5858−5959] and Nadezhda Zvartau[1,2][0000−0001−6533−5950]

[1] ITMO University, St. Petersburg 197101, Russia
a.a.semakova@gmail.com
[2] Almazov National Medical Research Centre, St. Petersburg 197341, Russia
zvartau@almazovcentre.ru

Abstract. Due to the increasing popularity of clinical decision support systems, the problem of personalized drug dose identification becomes more relevant and substantial. In this paper, the authors introduce a data-driven model designed to operate in this case. Current work comprises general problem formulation, description of data and its preprocessing steps, model design overview, its first stage model tuning and training, evaluation metrics used to estimate the quality, achieved values.

Keywords: Digital healthcare · Personalized dose identification · Classification algorithms · Electronic health records · Decision support systems.

1 Introduction

Currently, decision support systems that consider patient characteristics are gaining more popularity and impact on the process of treatment [1]. Individual treatment rule (ITR) that assigns an appropriate treatment to the specific patient based on his/her characteristics is one of decision support systems important elements [2]. An individual dosage rule (IDR) can be considered as a part of ITR. It maximizes the expected treatment outcome for each patient by defining individual drug dosages.

ITR was studied for various diseases, such as oncology [3] or genome-guided therapy [4]. In this research, the ITR is constructed for patients with arterial hypertension. A similar problem was considered in the case of antihypertensive monotherapy, where the patients get treatment with a single drug class [5].

* The reported study was funded by RFBR according to the research project #18-37-00441.

This task of obtaining ITR was being solved in specific areas with such approaches as the Q-learning [6] and O-learning (Outcome Weighted Learning) [7] algorithms, and statistical random-effects linear models [8].

In this paper, the authors consider a supervised learning approach to obtain ITR for personalized combined drug therapy.

2 Problem statement

Given the vector $X_j(t_i)$ of patient profile j in the moment of time t_i before treatment as:

$$X_j(t_i) = \left\{ x_j^{(h)}(t_i) \right\} \tag{1}$$

Obtain the set of antihypertensive therapy:

$$Y_j(X_j(t_i)) = \left\{ (y_j^{(k,1)}, y_j^{(k,2)}) \right\}, \tag{2}$$

where $j = \overline{1,m}$; $i = \overline{1,n}$; $h = \overline{1,p}$; $k = \overline{1,q}$ and each drug y_j^k is a vector containing a drug International Nonproprietary Name (INN) and optimal daily dosage. The model, which authors propose in current paper, is designed to consist of three data-driven submodels: the first model receives a vector of patient features as an input and predicts an optimal drugs count (n_{opt}), the second model extends the result specifying drug INNs $\left\{ y_j^{(k,1)} \right\}_{k=1}^{n_{opt}}$ and the third model defines the desired daily dosages of each drug INN $\left\{ y_j^{(k,2)} \right\}_{k=1}^{n_{opt}}$ resulting with $\left\{ (y_j^{(k,1)}, y_j^{(k,2)}) \right\}_{k=1}^{n_{opt}}$.

3 Data description

The data used in this study were collected from 2010 to 2015, depersonalized and provided by Almazov National Medical Research Centre.

In the work, 16 features are grouped into a vector describing the patient profile. They are the following: age, sex, body mass index (BMI), systolic and diastolic blood pressure before treatment, smoking status, impaired glucose tolerance (IGT), left ventricular hypertrophy (LVH), chronic heart failure (CHF), ischemic heart disease (IHD), dyslipidemia, diabetes, microalbuminuria, cardiovascular diseases (CVD) among relatives, chronic kidney disease stage (CKD), and combination of concomitant drugs.

In the authors' previous study, eight patient clusters were been identification based on the feature tuples. The patient groups accept the clinical interpretation. As time passes, a patient profile will be changing due to disease development and ageing patient. It means that the same patient can belong to various groups depending on the disease dynamics at different points in time. To probabilistic model the arterial hypertension development in a patient, we use a *Markov chain* of transition from one cluster to another cluster. Therefore, the dynamic model of the hypertensive patient transitions process from cluster i ($i = \overline{1,8}$) to cluster j ($j = \overline{1,8}$) is presented as a graph, where the nodes are

clusters and the edges are transition probabilities P_{ij} (Fig. 1). It's noted, the probabilities P_{ij} are determined in a way that the transition probabilities sum is equal to one. Also, the patient condition will be able to remain the same then the patient will transition to the same cluster. These transitions are presented as a loop on between groups graph in Fig. 1. However, the certain clusters are incompatible for relative transitions, in particular, due to gender characteristics and/or the chronic concomitant diseases.

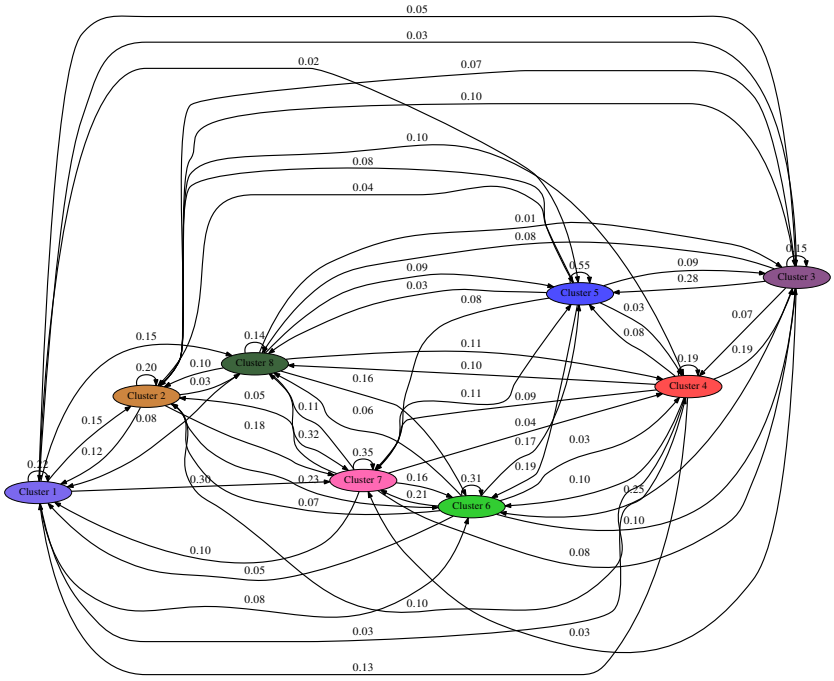

Fig. 1. Hypertensive patient state space.

Additionally, training dataset contains an outcome field as a result of the one-month treatment process. This value accepts various ways to be set. In this particular research, it is based on clinical guidelines and points out if systolic and diastolic blood pressure levels have reached the target values less than 140/90 mm Hg [5].

Since the task is to predict the optimal therapy for each patient based on his/her features vector, such fields as drug names and daily dosages have to be in the training data. So that, data was merged (on patient ID and outpatient visit date) with records contained medical prescriptions for these patients. As long as medical texts are written in natural language, they require additional processing to distinguish desired information and fill these fields.

4 Data preprocessing

Due to the lack of training dataset, the processing task was resolved in this research with a sequence of regular expressions and extracting rules implementing so-called rule-based natural language processing (NLP). Below is the overall pipeline that was applied to each outpatient visit:

1. Split the medical prescriptions field into substrings with '\n' (newline) delimiter. In the provided data most of such substrings contain, if they do, only one prescription of the drug.
2. Find all entries of drugs in each substring using the dictionary prepared by authors. This dictionary includes drug brand-names, their different writing options that may show up in a natural language text, INNs, and pharmacological classes.
3. The substring may have no medical prescriptions because it has general guidance, referral to laboratory testing, etc. Such substring is not involved in further processing.
4. If substring contains several drug brand-names, then distinguish them as an alternative or as a combination. Patterns, which are used in this step, include checking: their INNs – the same indicates the alternative, their location in string boundaries, and the presence of 'and' symbols, commas between them, conjunctions.
5. Check the presence of words that mean cancellation, dosages and frequency indicators. Substrings that don't have dosages and frequency can be involved in the dataset with filling missing values using appropriate machine learning algorithms in further preprocessing.
6. Extract dosages using regular expressions with measurement units, frequencies – using regular expressions with parts of the day patterns.
7. Aggregate INNs of all extracted drugs in the INN field, calculate their daily dosages and write them in the Dosage field using specified delimiter (in this research '|').

5 Classifier implementation

This paper is aimed to describe the first submodel in detail, which is proposed to be an extension of a treatment outcome classifier. In the cycle, it concatenates the vector of the patient features with every possible drug count $g = \overline{1, r}$ and utilizes the classifier to predict the probabilities of treatment ineffectiveness (negative class, 0) and effectiveness (positive class, 1) denoted as $\{(p(0), p(1))_g\}$.

The combination with the maximum outcome probability is assumed to contain an optimal number of drugs.

$$n_{opt} = \arg\max_{p(1)} \left\{(p(0), p(1))_g\right\} \tag{3}$$

In the current Python implementation, several scikit-learn classifiers (library version 0.21.3) were trained, tuned and evaluated, including C-Support Vector

Classifier (SVC), random forest classifier (RF), Multi-layer Perceptron (MLP), classifier as well as LightGBM (library version 2.3.0).

Hyper-parameters estimation using cross-validation led to the following values (parameters not mentioned below are expected to have the default values for the specified library version):

- SVC: radial basis function (RBF) kernel, balanced class weights, enabled probability estimation, kernel coefficient $\gamma = 0.04$, penalty parameter of the error term $C = 8.0$.
- Random forest classifier: entropy criterion, balanced class weights, with 150 trees in the forest of maximum depth 2 and 10 minimum samples to split.
- MLP: stochastic gradient descent (sgd) solver, invscaling learning rate, maximum number of iterations is 20, one hidden layer with 76 neurons, regularization term is 0.2, using Nesterov's momentum, shuffle samples in each iteration.
- LightGBM: random forest boosting type, balanced class weights, bagging frequency is 1, bagging fraction is 0.9, learning rate is 0.01, number of estimators is 110.

6 Assessment

After the preprocessing step the dataset containing 4521 records were divided into three datasets: training, validation, and test in a ratio of 0.63:0.27:0.1 respectively. It was decided to consider both Sensitivity (Recall) and Specificity while estimating the treatment outcome classifiers parameters. This decision is based on the requirement to efficiently identify both ineffective and effective treatment and use it in revealing the optimal drug amount.

The first two datasets were used in 5 splits with 7 repeats cross-validation, results of which are presented in Table 1. Table 2 gives the results of classifiers quality evaluation on the third (test) dataset. Although Sensitivity and Specificity were considered as target metrics, the tables additionally include values of such metrics as Accuracy, Precision, F1 score and ROC AUC (Receiver Operating Characteristic Area Under ROC Curve).

Table 1. Cross validation scores.

Classifier	Accuracy	Precision	Recall	Specificity	F1 score	ROC AUC
SVC	0.56133	0.44986	0.73122	0.45855	0.55668	0.62650
RF	0.57787	0.45993	0.68487	0.51284	0.54999	0.62827
MLP	0.51973	0.39233	0.49659	0.53435	0.43776	0.53079
LGPM	0.57695	0.45488	0.61014	0.55715	0.52049	0.61416

As can be seen from the tables that all classifiers avoided overfitting. MLP classifier performed worse than the others, which showed comparatively close

Table 2. Test scores.

Classifier	Accuracy	Precision	Recall	Specificity	F1 score	ROC AUC
SVC	0.56076	0.45691	0.70763	0.46783	0.55528	0.58773
RF	0.56979	0.46409	0.71186	0.47989	0.56187	0.59588
MLP	0.50082	0.38925	0.50636	0.49732	0.44015	0.50184
LGPM	0.56897	0.45293	0.54025	0.58713	0.49275	0.56369

results. The random forest classifier is assumed to perform the most optimal way.

The most important patient features and their importances with more than 1% impact returned by random forest classifier are presented in Fig. 2. They are calculated as the impurity decrease from each feature: the reduce in node impurity weighted by the probability of reaching the node. It can be seen that the drug count provides around 4.0% of the total decision.

However, authors assume that using a bigger training dataset will improve the results and the impact of the drug count feature, which is now considered to be not sufficient enough to reliably separate the effective therapy from the ineffective.

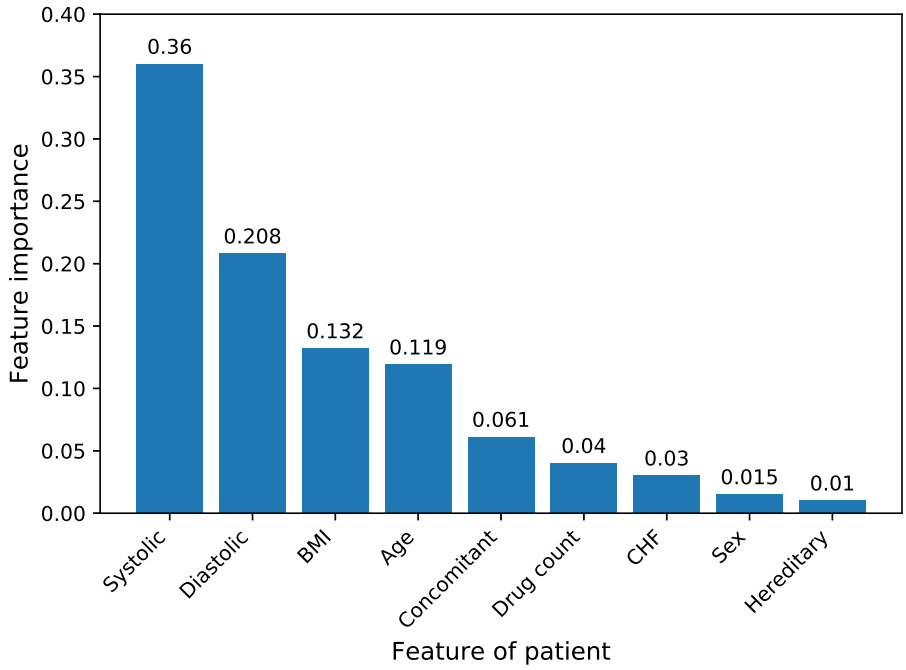

Fig. 2. Feature importances.

7 Conclusion and Future works

As a result of this work, the model that predicts the optimal antihypertensive drug count was presented, implemented, trained, and evaluated. This model is a part of the proposed general model predicting the optimal antihypertensive drug dosages based on the patient features.

Future works of this research include:

- preparation and preprocessing of new data collected from 2016 to 2019 that will be provided by Almazov National Medical Research Centre;
- further training and parameters tuning of additional classifiers predicting the optimal amount of prescription drugs for a patient with arterial hypertension;
- development of the data-driven model predicting the most effective individual antihypertensive therapy including drug INNs and daily dosages.

Acknowledgements

The reported study was funded by RFBR according to the research project #18-37-00441.

References

1. Somogyi, R., McMichael, J.P., Baranzini, S.E., Mousavi, P., Greller, L.D.: 10 Advanced data mining and predictive modelling at the core of personalised medicine. Studies in Multidisciplinarity **3**, 165–192 (2005)
2. Darwich, A.S., Ogungbenro, K., Vinks, A.A., Powell, J.R., Reny, J.L., Marsousi, N., Daali, Y., Fairman, D., Cook, J., Lesko, L.J., McCune, J.S., Knibbe, C.A.J., de Wildt, S.N., Leeder, J.S., Neely, M., Zuppa, A.F., Vicini, P., Aarons, L., Johnson, T.N., Boiani, J., Rostami-Hodjegan, A.: Why has model-informed precision dosing not yet become common clinical reality? lessons from the past and a roadmap for the future. Clin. Pharmacol. Ther. **101**(5), 646–656 (2017)
3. Barbolosi, D., Ciccolini, J., Lacarelle, B., Barlési, F., André, N.: Computational oncology-mathematical modelling of drug regimens for precision medicine. Nat Rev Clin Oncol **13**(4), 242–254 (2016)
4. Bielinski, S., Olson, J., Pathak, J.: Preemptive genotyping for personalized medicine: Design of the right drug, right dose, right timedusing genomic data to individualize treatment protocol. Mayo Clinic Proceedings **89**(1), 25–33 (2014)
5. Semakova, A., Zvartau, N., Bochenina, K., Konradi, A.: Towards Identifying of Effective Personalized Antihypertensive Treatment Rules from Electronic Health Records Data Using Classification Methods: Initial Model. In: Procedia Computer Science, pp. 852–858. Elsevier B.V. (2017)
6. Moodie, E.E., Chakraborty, B., Kramer, M.S.: Q-learning for estimating optimal dynamic treatment rules from observational data. Can J Stat **40**(4), 629–645 (2012)
7. Chen, G., Zeng, D., Kosorok, M.R.: Personalized Dose Finding Using Outcome Weighted Learning. J Am Stat Assoc **111**(516), 1509–1521 (2016)
8. Diaz, F.J., Yeh, H.W., de Leon, J.: Role of Statistical Random-Effects Linear Models in Personalized Medicine. Curr Pharmacogenomics Person Med **10**(1), 22–32 (2012)