Arh. farm. 2022; 72: 701 – 715 *Original scientific paper/Originalni naučni rad*

Data mining techniques applied in the analysis of historical data

Jovana Kovačević[1*], Aleksandar Kovačević[2], Tijana Miletić[1], Jelena Djuriš[3], Svetlana Ibrić[3]

[1] Hemofarm AD, Product development, Beogradski put BB, 26300 Vršac, Serbia

[2] University of Novi Sad - Faculty of Technical Sciences, Department of Computing and Control Engineering, Trg Dositeja Obradovića, Novi Sad, Serbia

[3] University of Belgrade - Faculty of Pharmacy, Department of Pharmaceutical Technology and Cosmetology, Vojvode Stepe 450, 11221 Belgrade, Serbia

[*]Corresponding author: Jovana Kovačević, E-mail: jovana.kovacevic@hemofarm.com

Abstract

Understanding the effect of the characteristics of formulation and process parameters on the physicochemical properties of a pharmaceutical product is very significant for the development of solid dosage forms, as the knowledge gained on small scale batches in the early phase of development is used in the later phases of product lifecycle or in the development of other products. One of the approaches for gaining a better understanding of the effects of the formulation and production process on the quality of the finished product is to apply a systematical approach which concerns performing experiments according to a predefined factorial or fractional factorial experimental plan. However, often it is the case that there are available data gathered in a non-systematic way, because experiments were not performed according to a predetermined experimental plan. In such a case, data mining techniques could be used to extract useful data from the historical data set. In this research, the possibility of using several data mining techniques to build models that describe the effect of formulation characteristics on acid resistance and dissolution profile of a model drug from gastro-resistant pellets. The model drug used in the research is duloxetine hydrochloride from the group of antidepressants. It belongs to the BCS 2 class of active pharmaceutical ingredients, and it is therefore necessary for the release profile of duloxetine to be characterized by a higher number of tested time points. The developed models can be used for planning future laboratory trials, or in the development of other products.

Key words: drug manufacturing, gastro-resistant pellets, modelling, release profile, acid-resistance

https://doi.org/10.5937/arhfarm72-41368

Introduction

Data mining is a branch of computer science dealing with untrivial extraction of implicit, previously unknown and potentially useful information from data bases. It combines several machine learning, artificial intelligence, pattern recognition and statistic techniques, and is also known as exploratory data analysis and data driven discovery. Data mining is an integral and essential part of knowledge discovery in databases (KDD) which designates the whole process of conversion of data into useful information (1). The process of knowledge discovery is iterative and consists of 9 steps: the definition of target of KDD, selection and creation of data set on which KDD will be used, preprocessing and cleaning of data, data transformation, choice of data mining type to be used (descriptive or predictive), choice of algorithm to be used, application of algorithm, evaluation and application of discovered knowledge.

Data mining techniques are used in the optimization of formulation of pharmaceuticals (2, 3), optimization of production process (4, 5), analysis of historic data to improve the understanding of the process (6, 7), correlating *in vitro* results with *in vivo* data (8), and to test the stability of drug products (9).

Two main types of data mining are the one oriented towards verification, when the system confirms a user's hypothesis, and the one oriented towards discovery, when the system autonomously finds new patterns and rules. The methods oriented towards verification are some of the usual statistical methods, such as analysis of variance (ANOVA), hypothesis testing (for example t-test of means), or testing of behavior of distribution. These methods are not often related to data mining, because most of the problems with which data mining deals are based on discovery and not on hypothesis testing. The part of data mining oriented towards discovery comprises descriptive and predictive methods. Descriptive methods discover patterns among data and enable a better understanding of the way in which the examined data are related. Predictive methods automatically build models based on known results (1).

Data mining concerns dealing with following tasks (10, 11):

- Anomaly detection – a descriptive task which concerns finding unusual data that need to be examined in greater detail.
- Association rule learning – a descriptive task which concerns discovering relations between variables.
- Clustering – a descriptive task which concerns discovering similar groups and structures within the evaluated data.
- Classification – a predictive task concerning the discovery of the category (subpopulation) to which a new case belongs, based on the training data set which contains cases of known categories.
- Regression – a predictive task which concerns discovering functions that model data with minimal error, i.e. that find relations between the output variable and one or more input variables. The difference between regression and classification lies in

the nature of the output variable. Regression is applied in solving problems which have continuous values as output variables, while classification is used when output variables are categorical values.

- Summarization – a descriptive task that concerns devising a description of a group (or subgroup) of data.

The type of the most suitable method used for solving regression problems depends on the available data set: the number of predictors (input variables), number of responses (output variables) and the effect that predictors have on responses. The effect of predictors on responses is a priori unknown, so different techniques need to be applied on a certain data set to determine which technique is optimal for the given set. The modelling technique and its parameters need to be chosen so as to correspond to the data. Otherwise, the model could underfit (it would perform poorly on the training data set and therefore it would not be able to make predictions with new data) or overfit (it would perform too well with the training data set and would therefore be unable to make accurate predictions with the new data).

The objective of our study is to evaluate the possibility of using different data mining techniques to model the impact of formulation of enteric coated pellets on acid resistance and dissolution profile of a model drug. The model drug in this study was duloxetine hydrochloride, which degrades to a significant extent under acidic conditions. To prevent drug degradation and subtherapeutic levels in plasma caused by loss of drug in the stomach, duloxetine has to be formulated like a gastro-resistant dosage form. Furthermore, to ensure adequate stability of the finished product, contact between acid labile drug and acidic enteric polymer has to be prevented by an isolating layer of adequate thickness (12). Therefore, if duloxetine is formulated as pellets, these pellets need to contain at least three layers: drug layer, isolating layer and gastro-resistant layer. A historical data set obtained in trial and error laboratory trials was used to build models. The task of predicting acid resistance and drug release from enteric coated pellets was formalized as a regression problem, and the suitability of several regression techniques for solving this problem was tested.

Materials

Duloxetine hydrochloride of particle size distribution d (90) < 10 μm (JiuZhou, China), non-pareil seeds 20 – 25 mesh and 25 – 30 mesh (JRS Pharma, Germany), povidone K-30 (ISP, Switzerland), hypromellose 3 cp (Taian Ruitai Cellulose Co. Ltd., China), hypromellose 6 cp (The Dow Chemical Company, The UK), triacetin (Eastman Chemical B.V., Switzerland), talc (Imerys Talc, Italy), hypromellose acetate succinate MF (Shin-Etsu, Japan), hypromellose acetate succinate LF (Shin-Etsu, Japan), sucrose (Sunoko, Serbia), hydroxypropylcellulose EF and LF (Ashland, The USA), ethylcellulose 10 cp (Ashland, The USA), isopropanol (Brenntag-CEE, Germany) and ethanol concentrated (Swan Lake, Serbia) were used in the performed experiments. All the other reagents were of analytical grade.

Methods

Preparation of coated pellets

All coating trials were carried out in a fluid-bed dryer granulator Glatt GPCG2 (Glatt GmbH, Germany) in a configuration with Wurster partition and two fluid nozzle of 1.2 mm opening placed on the bottom of the device. The production process parameters were set bearing in mind the basic principles of pellet coating by using this technique so as to avoid agglomeration of pellets and prevent the effect of process parameters on the responses. For example, when the coating liquid contained water, trials were carried out at lower spray rates than when the coating liquid was completely solvent based: trials where the drug was dissolved in a mixture of purified water and isopropanol were carried out at the spray rate of 4 g/minute, and trials where the drug was suspended in isopropanol were carried out at spray rate of 6 g/minute. Inlet air flow rate for all the trials was between 50 m3/h and 100 m3/h. Atomizing air pressure was kept at 1.8 bar ± 0.5 bar for all the trials. All coating liquids were prepared by dissolving soluble materials in solvent, and insoluble materials (if there were some in the formulation) were suspended and homogenized separately prior to the addition of soluble materials to the solution.

Assay of duloxetine

An Agilent HPLC system equipped with a 1100 series high pressure binary gradient pump along with a pulse damper, photo diode array detector with auto liquid sampler handling system was used for the analysis of the samples. The samples were tested using a validated in-house analytical procedure for assay determination. The HPLC system was equipped with a 4.6mm x 15 cm analytical column Zorbax XDB C-18, 150 x 4.6 mm with 3.5µm particle size (Agilent Technologies, USA). The column eluent was monitored at the detection wavelength of 225nm. The mobile phase consisted of a filtered and degassed mixture of buffer, HPLC grade acetonitrile and HPLC grade tetrahydrofuran in 65:20:15 ratio. The buffer was prepared by dissolving 7.4 g of sodium perchlorate monohydrate in 1000 ml of purified water, adding 3 ml of triethylamine and adjusting pH with perchloric acid to 2.0. Chromatography was performed maintaining the column at room temperature with the flow rate of 1.5 ml/min. The data was recorded using the OpenLab CDS Ezchrom edition software (Agilent Technologies, USA).

Dissolution profile of gastro-resistant pellets

The dissolution study was conducted on a DT800 Dissolution Tester (Erweka, Germany). The dissolution profile of gastro-resistant pellets was conducted according to the PhEur 2.9.3, method B with apparatus I (basket apparatus). The test was carried out at 37 ± 0.5 ºC in 1000 ml of 0.1M HCl (pH 1.2) for 120 minutes (acid stage) and in 1000 ml of phosphate buffer pH 6.8 (buffer stage) for another 90 minutes. Basket rotation was adjusted to 100 rpm. At predetermined time points 4 ml of test medium were withdrawn, filtered through a 1.0 µm glass fibre filter and the first 2 ml of filtrate were discarded. The volume of the test medium (4 ml) withdrawn was not replaced. The concentration of the drug was determined by the previously described validated HPLC method. Drug release

was determined in minimum three samples. Acid resistance is the amount of dissolved drug after the acid stage of the dissolution test. The percent of dissolved drug after 5, 10, 15, 20, 30, 45, 60 and 90 minutes is the percent of drug dissolved in the buffer stage of dissolution (after 120 minutes of testing in the acidic medium).

Data set

An overview of input and output parameters that were included in the development of models is given in Table I.

Table I Input and output variables included in formulation modelling

Tabela I Ulazni i izlazni parametri uključeni u modelovanje formulacije

Qualitative input variables		
Diameter of inactive pellets	20 – 25 mesh, 25 – 30 mesh	
Polymer in API layer	HPMC 6 cp, PVP K-30, HPC EF, HPC LF	
Solvent in API layer	water, isopropanol, mixture of water and isopropanol	
Polymer in isolation layer	HPMC 6 cp, HPMC 6 cp + HPC EF, HPMC 6 cp + HPC LF, HPMC 6 cp + EC	
Polymer in GR layer	HPMCAS LF, HPMCAS MF	
Solvent in GR layer	water, mixture of water and ethanol (20:80)	
Quantitative input variables	*Minimum (%)*	*Maximum (%)*
Content of inactive pellets	20	36
Content of polymer in API layer	20	52
Weight gain of isolating layer	18	47
Polymer content in isolating layer	10	41
Content of talc in isolating layer	10	67
Filler content in isolating layer	0	80
Weight gain of gastro-resistant film	20	66
Polymer content in gastro-resistant film	37	100
Lubricant content in gastro-resistant film	0	56
Plasticizer level in gastro-resistant film	0	20
Output variables		
Acid resistance	0.01	18.10
% of dissolved API after 5 minutes	0.2	4.3
% of dissolved API after 10 minutes	2.5	26.5
% of dissolved API after 15 minutes	13.9	51.8
% of dissolved API after 20 minutes	31.9	72.4
% of dissolved API after 30 minutes	46.5	82.3
% of dissolved API after 45 minutes	57.5	95.5
% of dissolved API after 60 minutes	64.9	101.4
% of dissolved API after 90 minutes	73.7	102.1

A historical data set obtained in 19 experiments not carried out according to a factorial or fractional factorial experimental plan was used for the modelling of formulation. The data set comprised all the available information regarding qualitative

and quantitative composition of gastro-resistant pellets. The preparation of data for further processing included transforming qualitative data into coded numerical values. Whenever it was possible, coded numerical values were assigned in a logical way to qualitative input parameters. Therefore, pellets of 20 – 25 mesh size were assigned designation 0 because their surface area in the finished product is smaller than the surface area of pellets of 25 – 30 mesh size, which were assigned designation 1. Furthermore, polymer hypromellose acetate succinate grade MF that dissolves at a lower pH value of solution was given designation 0, and grade LF was given designation 1 because it dissolves at a higher pH value of solution. Other qualitative input parameters were given designations randomly. For all the models, except for regression trees and the ensemble of regression trees obtained by the "boosting" technique, categorical attributes were transformed so that they could have values 0 or 1. For example, if the categorical attribute “Polymer content in API layer” had values 0, 1, 2 and 3, the attribute was transformed into four categorical attributes: polymer 0, polymer 1, polymer 2 and polymer 3, all of which could have values of 0 or 1. Therefore, after transformation, there were 27 input variables. The transformation of categorical attributes was done because regression would have treated numbers 0, 1, 2 and 3 as numbers which differ from each other by exactly 1, which is not the case. On the other hand, regression trees are models that can deal with categorical attributes with nonnumerical values, and for this modelling technique it was not necessary to transform categorical attributes so that they have values 0 or 1. The values of quantitative input variables for all the models, except regression trees and the ensemble of regression trees obtained by the "boosting" technique, were standardized/ normalized by using z-transformation.

Output variables were acid resistance and the percent of dissolved drug after 5, 10, 15, 20, 30, 45, 60 and 90 minutes of dissolution testing. Qualitative and quantitative composition of evaluated formulations is given in Table II, and acid resistance and dissolution profiles of these formulations are given in Table III.

Table II Input parameters evaluated by different data mining techniques

Tabela II Ulazne promenljive ispitane različitim tehnikama istraživanja i analize podataka

	Inactive pellets		Drug layer			Isolating layer					Gastro-resistant layer					
Exp.	Diameter[1]	Content (%)	Polymer[2]	Polymer content (%)	Solvent[3]	Weight gain (%)	Polymer[4]	Polymer content (%)	Lubricant content (%)	Filler[7] content (%)	Weight gain (%)	Polymer[5]	Solvent[6]	Polymer content (%)	Lubricant[8] content (%)	Plasticizer[9] content (%)
1	0	29	0	40	0	20.3	0	41	39	20	34.0	0	0	71	21	7
2	0	30	0	40	0	20.3	0	41	39	20	31.6	1	1	63	19	18
3	1	24	0	44	0	25.0	0	33	67	0	66.0	1	1	53	32	15
4	1	23	0	44	0	35.5	0	40	20	40	42.4	0	1	53	32	15
5	0	27	1	35	1	21.7	0	33	67	0	59.3	1	0	72	22	6
6	0	30	1	35	1	18.8	1	33	67	0	44.3	1	0	71	21	7
7	0	22	0	52	2	23.4	0	33	67	0	31.6	1	0	71	21	7
8	0	20	0	52	2	27.0	1	33	67	0	39.0	1	0	71	21	7
9	0	26	1	35	1	21.5	1	33	67	0	30.2	1	0	40	40	20
10	0	26	1	35	1	31.1	2	33	67	0	20.6	1	0	40	40	20
11	0	24	1	35	1	46.6	2	33	67	0	19.8	1	0	37	56	7
12	0	25	1	35	1	46.6	2	33	67	0	15.4	1	0	37	56	7
13	0	30	1	35	1	17.4	3	33	67	0	21.8	1	1	100	0	0
14	0	28	1	35	1	21.3	3	33	67	0	23.6	1	1	100	0	0
15	0	28	1	35	1	21.3	3	33	67	0	24.2	1	1	50	50	0
16	0	36	2	20	1	17.6	0	40	20	40	23.8	1	1	100	0	0
17	0	31	2	20	1	38.6	0	40	20	40	23.2	1	1	100	0	0
18	0	34	3	20	1	22.7	0	10	10	80	26.8	1	1	100	0	0
19	0	32	3	20	1	34.3	0	10	10	80	20.4	1	1	100	0	0

[1] 0 – inactive pellets 20 – 25 mesh; 1- inactive pellets 25-30 mesh

[2] 0 – Hypromellose 6 cp; 1 – Povidone K-30; 2-Hydroxypropylcellulose LF; 3-Hydroxypropylcellulose EF

[3] 0 – water; 1-isopropanol; 2- water + isopropanol

[4] 0 – Hypromellose 6 cp; 1- Hypromellose 6 cp + Hydroxypropylcellulose EF; 2- Hypromellose 6 cp + Hydroxypropylcellulose LF; 3- Hypromellose 6 cp + Ethyl cellulose

[5] 0 – Hypromellose acetate succinate MF; 1 – Hypromellose acetate succinate LF

[6] 0 – water + ethanol (20:80), 1 – water

[7] Sucrose is used as a filler in the isolating layer

[8] Talc is used as lubricant in the gastro-resistant layer

[9] Triethylcitrate is used as plasticizer

Table III Acid resistance and dissolution data of experiments that were modelled by different data mining techniques

Tabela III Gastrorezistencija i brzina rastvaranja u eksperimentima korišćenim za modelovanje različitim tehnikama istraživanja i analize podataka

Experiment	Acid resistance (%)	5. minute (%)	10. minute (%)	15. minute (%)	20. minute (%)	30. minute (%)	45. minute (%)	60. minute (%)	90. minute (%)
1	0.2	0.2	6.5	28.1	50.3	77.1	93.8	99.0	100.4
2	0.1	0.4	5.4	23.1	42.8	66.0	87.0	96.9	100.2
3	0.9	2.4	9.4	34.1	55.9	77.5	92.8	98.6	100.6
4	0.1	1.6	12.4	37.9	60.0	82.3	95.5	99.4	100.0
5	0.2	2.3	5.5	35.8	60.4	77.1	87.0	92.7	97.5
6	0.1	1.3	22	50.7	65.4	76.7	85.2	90.4	95.0
7	0.2	0.2	4.1	16.7	33.2	64.3	91.1	98.0	98.8
8	0.2	0.9	2.5	13.9	31.9	66.1	95.4	101.4	102.1
9	0.3	1.9	26.5	55.0	68.6	78.8	86.7	91.5	96.3
10	0.4	3.8	23.2	51.8	69.5	82.2	90.8	94.9	98.0
11	0.3	2.1	16.7	45.7	65.4	80.0	89.9	94.6	97.6
12	3.0	3.2	25.7	57.1	72.4	83.5	90.8	94.2	96.2
13	0.3	2.2	23.2	43.5	56.1	69.7	82.5	91.2	97.3
14	0.1	1.7	24.0	44.6	56.7	70.7	83.0	92.3	99.4
15	18.1	1.7	12.2	24.6	34.5	46.5	57.5	64.9	73.7
16	0.4	3.9	9.0	19.1	32.2	52.2	70.4	82.1	93.6
17	0.9	4.3	9.9	20.0	32.1	49.6	66.2	77.9	89.8
18	0.0	2.8	10.4	27.3	44.5	64.3	80.0	89.0	96.2
19	0.1	2.1	9.3	27.1	44.4	63.8	80.3	90.6	98.0

Methods for modelling acid resistance and release of duloxetine from gastro-resistant pellets

Several approaches to modelling acid resistance and release of duloxetine from gastro-resistant pellets were evaluated: multiple linear regression, stepwise regression, lasso regression, ridge regression, elastic net, regression trees, ensemble of regression trees obtained by the "boosting" technique and artificial neural networks.

All modeling techniques except artificial neural networks were tested by using software Matlab® 2013b (Mathworks, USA). Default settings were applied in the Matlab software. The data set was randomly split into a training set and a test set so that 17 examples were used for training and 2 examples (10% of all the data) were used for testing developed models.

To develop models by using an artificial neural network Statistica® 8.0 (StatSoft, USA) was used. The data was split into three groups: training set, test set and validation set. Out of 19 examples, 15 were used for training, 2 for the test and 2 for validation. The test set was the same as the one used in the Matlab software. Network architecture was

varied so as to have between 5 and 27 hidden layers. The network was trained until the smallest error on the validation data set was obtained. The error is calculated as a sum of squares of differences between predicted and real values:

$$E = \sum_{i=1}^{N}(y_i - t_i)^2 \qquad (1)$$

where N is the number of examples used in the given phase, y_i is the value predicted by the network, and t_i is the accurate value.

The validation data set was randomly chosen from the remaining examples in the training set. To build models, artificial neural networks of the multilayer perceptron (MLP) type were used.

Model evaluation

The developed models were evaluated by comparing root mean square errors (RMSE) for the test data set applied to test models. RMSE is calculated by using the following formula:

$$RMSE = [\sum (y_i p - y_i m)^2 / n]^{1/2} \qquad (2)$$

where yp is the accurate value, i.e. the value obtained experimentally, ym is the value predicted by the model, and n is number of examples.

The experimentally obtained results of dissolution profiles in phosphate buffer and values of dissolution profiles predicted by model were compared by calculating the similarity factor f2 according to the following formula (FDA, 1997):

$$f_2 = 50 \cdot \log\{[1+(1/n)\Sigma_{t=1}^{n} (R_t - T_t)^2]^{-0.5} \cdot 100\} \qquad (3)$$

where Rt and Tt are dissolution values of the examined batches at time t, and n is the number of time points included in the calculation. In our calculations, we considered all time points before dissolution of 85%, and only one time point after dissolving more than 85% of both batches. f_2 values greater than 50 ensure the equivalence of tested dissolution curves.

Results and discussion

A graphical representation for all calculated RMSE for all tested points of dissolution profile by applying all evaluated techniques is given in Figure 1. Looking at single techniques, elastic net performs the best, with the calculated RMSE range from 0.72 to 6.40. There were no literature data found on formulation modelling by using the elastic net technique. Ibrić et al. modelled the release of aspirin from prolonged release tablets and obtained RMSE for test data in the range from 1.85 to 5.19. Experiments were carried out according to a structured experimental plan. Models developed by using the

elastic net enabled predicting values based on a historical data set in such a way that RMSE was of a similar order of magnitude compared to the ones calculated for the data set obtained in a structured experimental plan by applying the artificial neural network technique (3).

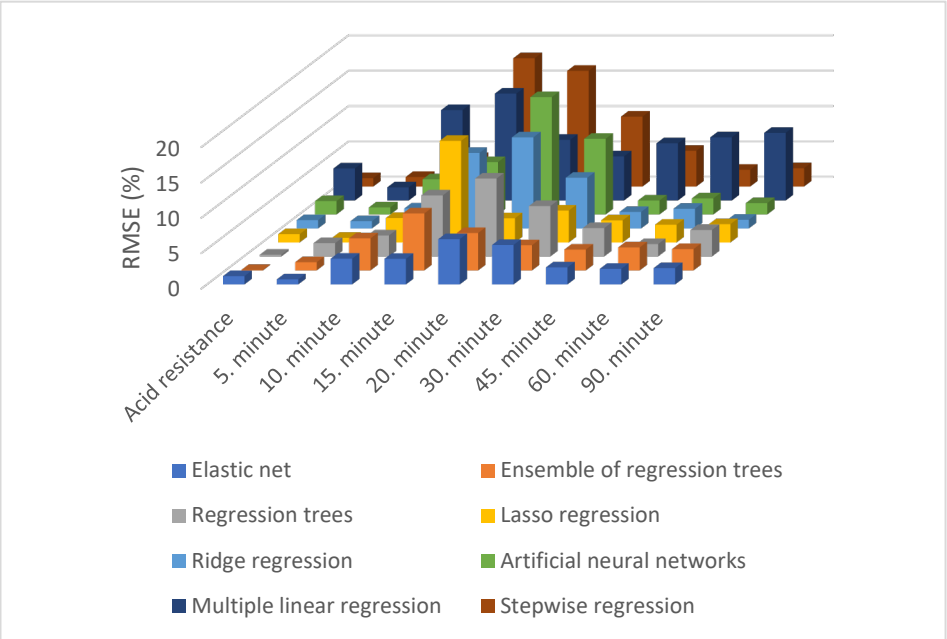

Figure 1. Comparative overview of RMSE for all evaluated points of release profiles and for all evaluated modelling techniques.

Slika 1. Uporedni pregled korena kvadrata srednje greške za sve ispitane tačke profila oslobađanja i za sve ispitane tehnike modelovanja

Looking at the calculated f2 values for release profiles in phosphate buffer obtained experimentally and the ones that were predicted by the model, it can also be concluded that elastic net performs better than any other tested technique, as the calculated f2 are the highest for this technique. An overview of the calculated f2 values for all evaluated techniques is presented in Table IV.

Table IV Calculated values of similarity factor f2 for test examples for all evaluated modelling techniques

Tabela IV Izračunate vrednosti faktora sličnosti f2 za sve ispitane tehnike modelovanja

	Multiple linear regression	Stepwise regression	Lasso regression	Ridge regression	Elastic net	Regression tress	Ensemble regression trees	ANN
Test 1	48.8	43.9	56.9	54.4	68.3	68.0	58.9	47.6
Test 2	54.6	57.1	60.2	59.9	70.3	66.0	59.9	60.3

ANN – Artificial neural networks

Petrović et al. (14) applied different types of artificial neural networks to model the release of diclofenac and caffeine from matrix tablets. They gathered the data in a systematic manner and obtained calculated f2 values between experimentally obtained and predicted release profiles in the range from 49.3 to 86.9. Chansanroj et al. (2) modelled drug release from matrix tablets by applying artificial neural networks and self-organizing maps. A comparison of the experimentally obtained and predicted dissolution profiles showed that the calculated similarity factors, f2, were in the range from 42.3 to 95.8.

However, models developed by the elastic net technique did not have the lowest calculated RMSE for all points of the release profile. Therefore, the highest value for the similarity factor, f2, between the experimentally obtained release profile and release profile predicted by a model is obtained when several techniques are combined. The techniques that had the lowest RMSE for a certain time point of release profiles are presented in Table V.

Table V Optimal techniques for predicting the release of model substance from gastro-resistant pellets

Tabela V Optimalne tehnike za predviđanje oslobađanja model supstance iz gastro-rezistentnih peleta

Response	*Techniques*	*RMSE for test data set*
Acid resistance	Ensemble of regression trees obtained by "boosting" technique	0.08
5. minute (%)	Lasso regression	0.66
10. minute (%)	Ridge regression	2.83
15. minute (%)	Elastic net	3.62
20. minute (%)	Lasso regression	3.42
30. minute (%)	Ensemble of regression trees obtained by "boosting" technique	3.56
45. minute (%)	Artificial neural networks	1.98
60. minute (%)	Regression trees	1.75
90. minute (%)	Ridge regression	1.23

When different modelling techniques are combined, all calculated RMSE values for the test data set are lower than 4. By combining different modelling techniques, given in Table V, to model different parts of the release profile, better similarity between the experimentally obtained values and those predicted by models is obtained, as it is presented in Figure 2.

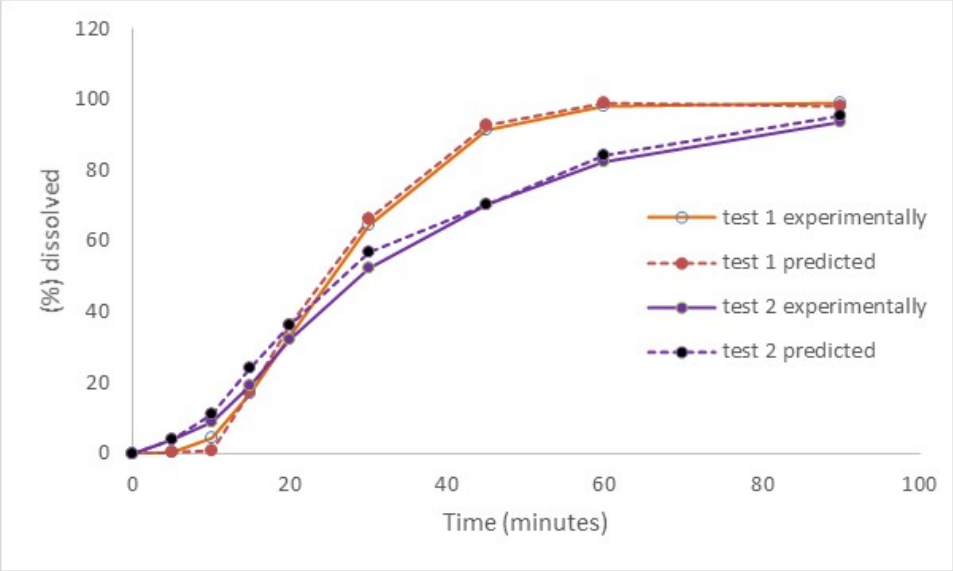

Figure 2. **Comparative release profiles of test formulations in phosphate buffer – experimental results vs. results predicted by models developed by different data mining techniques.**

Slika 2. **Uporedni profili oslobađanja test formulacija u fosfatnom puferu – eksperimentalno dobijeni rezultati vs. rezultati predviđeni modelima razvijenim različitim tehnikama istraživanja i analize podataka**

The calculated similarity factors, f2, for experimentally obtained and values predicted by different models combined are 82.3 for formulation test 1 and 74.6 for formulation test 2, which is higher than the obtained similarity factors for any single technique. The reason for this is the fact that the release of duloxetine from gastro-resistant pellets is a complex dynamic process. The behavior of the system changes over time, and it is therefore difficult to obtain equally good predictions for all tested time points by using only one modelling technique. For each tested time point, different conditions were optimal for modelling, which is caused by the complexity of the formulation: the presence of different layers in the formulation, their interaction in the product and their interaction in the dissolution medium.

Conclusion

The suitability of different data mining regression techniques for modelling acid resistance and dissolution profile of gastro-resistant pellets based on formulation characteristics was evaluated. Historical data were used for developing models, and the following regression techniques were evaluated: multiple linear regression, stepwise regression, lasso regression, ridge regression, elastic net, regression trees, ensemble of regression trees obtained by the "boosting" technique and artificial neural networks. Six qualitative and ten quantitative input variables were used to describe the formulation

characteristics of compositions used for training and testing the model. Elastic net was the modelling technique that had the overall lowest RMSE values and the highest value of the calculated similarity factor, f_2. However, this modelling technique did not have the lowest RMSE values for all the points of release profile. Therefore, it was concluded that the highest similarity of experimentally obtained release profiles and predicted release profiles would be achieved if techniques with lowest values for different points of release profile would be combined. By doing this, the calculated similarity factor between experimentally obtained and predicted values of release profiles increased to 82.3 for formulation test 1 and 74.6 for formulation test 2.

Therefore, it can be concluded that, by applying an adequate regression technique, historical data can be used for developing models suitable for predicting acid resistance and dissolution profile of gastro-resistant pellets based on the formulation characteristics. These models could be used for formulation optimization in the same way as models developed by using data gathered in experiments with structured experimental design.

References

1. Maimon O, Rokach L. Introduction to knowledge discovery and data mining. In: Maimon O, Rokach L, editors. Knowledge discovery and data mining handbook. 2nd edition. New York, USA: Springer; 2010; p. 1-13.
2. Chansanroj K, Petrovic J, Ibric S, Betz G. Drug release control and system understanding of sucrose esters matrix tablets by artificial neural networks. Eur J Pharm Sci. 2011;44:321–331.
3. Ibrić S, Jovanović M, Djurić Z, Parojčić J, Solomun L. The application of generalized regression neural network in the modeling and optimization of aspirin extended release tablets with Eudragit RS PO as matrix substance. J Control Release. 2002;82:213–222.
4. Mihajlović T, Ibrić S, Mladenović A. Application of Design of Experiments and Multilayer Perceptron Neural Network in Optimization of the Spray-Drying Process. Dry Technol. 2011;29(14),1638-1647.
5. Mansa RF, Bridson RH, Greenwood RW, Barker H, Seville JPK. Using intelligent software to predict the effects of formulation and processing parameters on roller compaction. Powder Technol. 2008;181:217-225.
6. Ronowicz J, Thommes M, Kleinebudde P, Krysinski J. A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm. Eur J Pharm Sci. 2015;73:44–48.
7. Mendyk A, Kleinebudde P, Thommes M, Yoo A, Szleka J, Jachowicz, R. Analysis of pellets with use of artificial neural networks. Eur J Pharm Sci. 2010;41:421–429.
8. Parojčić J, Ibrić S, Đurić Z, Jovanović M, Corrigan OI. An investigation into the usefulness of generalized regression neural network analysis in the development of level A in vitro–in vivo correlation. Eur J Pharm Sci. 2007;30:264-272.
9. Ibrić S, Jovanović M, Đurić Z, Parojčić J, Solomun Lj, Lučić B. Generalized regression neural networks in prediction of drug stability. J Pharm Pharmacol. 2007;59:745-750.

10. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37-54.
11. Sondwale PP. Overview of Predictive and Descriptive Data Mining Techniques. Int J Adv Res Comput Sci Softw Eng. 2015;5:263-265.
12. Jansen PJ, Oren PL, Kemp CA, Maple SR, Baertschi SW. Characterization of impurities formed by interaction of Duloxetine HCl with enteric polymers hydroxypropyl methylcellulose acetate succinate and hydroxypropyl methylcellulose phthalate. J Pharm Sci. 1998;87(1):81-85.
13. FDA Guidance for industry "Dissolution testing of immediate release solid oral dosage forms". US Department of Health and Human Services, CDER. 1997.
14. Petrović J, Ibrić S, Betz G, Đurić Z. Optimization of matrix tablets controlled drug release using Elman dynamic neural networks and decision trees. Int J Pharm. 2012;428:57– 67.

Tehnika istraživanja i analize podataka primenjena u analizi istorijskih podataka

Jovana Kovačević[1*], Aleksandar Kovačević[2], Tijana Miletić[1], Jelena Đuriš[3], Svetlana Ibrić[3]

[1] Hemofarm AD, Sektor razvoja, Beogradski put BB, 26300 Vršac, Srbija

[2] Univerzitet u Novom Sadu – Fakultet tehničkih nauka, Departman za računarstvo i automatiku, Trg Dositeja Obradovića, 21000 Novi Sad, Srbija

[3] Univerzitet u Beogradu – Farmaceutski fakultet, Katedra za farmaceutsku tehnologiju i kozmetologiju, Vojvode Stepe 450, 11221 Beograd, Srbija

[*]Autor za korespondenciju: Jovana Kovačević, E-mail: jovana.kovacevic@hemofarm.com

Kratak sadržaj

Razumevanje uticaja karakteristika formulacije i procesnih parametara na fizičko-hemijske osobine farmaceutskog proizvoda je vrlo značajno u razvoju čvrstih doziranih oblika jer se znanje stečeno u fazi razvoja koristi i u svim sledećim fazama životnog ciklusa proizvoda, a može da se primeni i u razvoju drugih proizvoda. Jedan pristup ka postizanju boljeg poznavanja proizvoda i procesa je primena sistematičnog pristupa koji podrazumeva izvođenje eksperimenata u skladu sa predefinisanim faktorijalnim ili frakcionim faktorijalnim eksperiementalnim planom. Međutim, čest je slučaj da dostupni podaci nisu prikupljeni na sistematičan način zato što eksperimenti nisu izvođeni po predefinisanom planu. Tada se mogu primeniti tehnike istraživanja i analize podataka da bi se iz seta istorijskih podataka izdvojili korisni podaci. U ovom istraživanju smo ispitali mogućnost korišćenja različitih tehnika istraživanja i analize podataka za razvoj modela koji opisuju efekte formulacije na gastrorezistenciju i profil oslobađanja model supstance iz gastrorezistentnih peleta. Model supstanca je duloksetin hidrohlorid iz grupe antidepresiva. Pripada BCS 2 klasi lekovitih supstanci, te je poželjno da profil brzine rastvaranja duloksetina iz peleta bude okarakterisan većim brojem vremenskih tačaka. Razvijeni modeli se mogu koristiti za planiranje narednih proba ili u razvoju drugih proizvoda.

Ključne reči: proizvodnja lekova, gastrorezistentne pelete, modelovanje, profil oslobađanja, gastrorezistencija