Model of medicines sales forecasting taking into account factors of influence

A G Kravets¹, M A Al-Gunaid¹, V I Loshmanov¹, S S Rasulov², L B Lempert ²

¹ Volgograd State Technical University, 28, Lenin Av., Volgograd, 400005, Russia
² Volgograd State Medical University, 1, Pavshikh Bortsov Sq., Volgograd, 400131, Russia

E-mail: agk@gde.ru

Abstract. The article describes a method for forecasting sales of medicines in conditions of data sampling, which is insufficient for building a model based on historical data alone. The developed method is applicable mainly to new drugs that are already licensed and released for sale but do not yet have stable sales performance in the market. The purpose of this study is to prove the effectiveness of the suggested method forecasting drug sales, taking into account the selected factors of influence, revealed during the review of existing solutions and analysis of the specificity of the area under study. Three experiments were performed on samples of different volumes, which showed an improvement in the accuracy of forecasting sales in small samples.

1. Introduction
Accurate forecasting of sales is an important and inexpensive way to increase profits, which will also reduce the impact of risks on the main processes occurring within the company. Based on the results of the constructed forecast, it becomes possible to model the most favorable environment for the production and distribution of goods. The forecasting model is a functional and adequate representation of the process under investigation as well as the basis for calculating the specific values that represent the expected result - breakeven sales. The health of people in need of medical care who are the main consumers of drugs depends also on the timely distribution of medicines through the main channels.

In the constantly changing world, forecasting the demand for new products is a key factor in economic well-being. Such forecasting concerns serious problems caused by a lack of data and uncertainty about how consumers will be able to adopt a new nomenclature. Obviously, the use of previous sales data sets is not possible, since the product has not yet been purchased or purchased in small quantities. In this regard, it is not possible to build an adequate model for changing sales indicators [1].

One of the features of pharmaceutical products marketing is the division of medicines into two main groups, such as original brands and generics (analogs, whose trade name coincides with the international non-proprietary name or the name of the active substance). These groups of drugs differ from each other, due to the legislative framework with respect to patent rights, research, and marketing. Most large pharmaceutical companies are engaged in the production of original brand-named drugs. Generics are usually relatively cheaper than brand names. This leads to the fact that
generic medicines are in high demand in countries whose population has average or low salaries. The growth of new markets relates, first of all, to generic medicines. Most sales forecasts, including those in the generic pharmaceutical industry, are based on the implicit assumption that sales can be represented by a time series of historical data, a model that can be constructed without taking into account characteristics that do not have a sufficient impact on total sales [2].

One of the main directions in the system of marketing research of a pharmaceutical organization is the analysis and forecasting of the sales performance of products [3]. The use of effective forecasting techniques provides a competent and justified approach to the distribution of goods along the main distribution channels. To obtain a more accurate forecast, it is necessary to study the features and identify the main factors of influence in the study area [4].

2. The problem statement
Previously [5] the authors found that existing methods are ineffective for pharmaceutical companies, due to requiring a large sample of data. Drugs, in turn, are constantly replaced by analogs or are updated to enhance pharmacological effects or eliminate side effects. Therefore, it is necessary to build an accurate model for forecasting sales of pharmaceuticals using one of the methods of machine learning, taking into account the constant updating of medicines and the lack of sufficient data on past sales of preparations of each kind (figure 1).

![Figure 1. Stages of the prediction model construction](image)

This study compares the accuracy of models predicting drug sales volumes within Volgograd region, built using the random forest method and based only on historical data and taking into account the factors of influence. Accuracy is assessed using the model identification criteria.

3. Methods for constructing a forecast model
At the moment, there are many methods for constructing a predictive model. Fuzzy modeling is widely used in solving problems of forecasting [6].

Existing researches in the field of pharmacology identify several most effective and accurate methods: a linear regression (LR), random forest (RF) method, construction of a time series prediction using a neural network (NN) [7], the use of support vector regression (SVR) [8] and the Levenberg-Marquardt algorithm (LMA) [9].

The demand for most drugs depends on the periods of increased risk of spreading diseases. In this
case, in constructing the model it is necessary to take into account seasonal trends [10]. The Autoregressive Integrated Moving Average (ARIMA) model and methods of harmonic analysis [11] can reveal and take into account seasonal fluctuations of the sales volume. Further, the behavior of the model is evaluated on the independent data by cross-checking [12] or by calculating the mean-square error of the forecast.

4. Method of constructing forecast taking into account factors affecting sales of drug
In previous work [13], the factors were studied which impact the sales of prescription drugs. The purpose of the study was to evaluate the effect of price and advertising on sales of medicines. It was found that promotions may increase the frequency of prescribing and price growth may decrease it. There was also a correlation between drugs prescription with gender and age of the doctor. However, the distribution indicators in [13] were not considered. Based on the authors’ studies of the distribution of the drug, an original algorithm was developed for the method of constructing the forecast, taking into account the impact factors (figure 2).

Figure 2. An algorithm of the forecasting method taking into account the impact factors

4.1 Determination of the impact factors based on analysis of distribution indicators
As a result of the research, the following factors influencing sales of the drug were determined:
1) the percentage of points of sale, in which this drug is available;
2) the share of the product in the gross sales volume of the commodity group;
3) the number of visits to a doctor for the current month;
4) the period of conducting the marketing activities;
5) the share of the given drug in the full-scale sales volume of all drugs of one type in Russia;
6) the share of the given drug in the cost sales volume of all drugs of one type in Russia;
7) current season;
8) the price of the drug tested;
9) the percentage of changing prices of preparations of this type on the commercial market;
10) selling;
11) training sample.

4.2 Formation of training sample
Sales of the drug "Hydroxyzine", provided by drugstore chains in the city of Volgograd and Volgograd region from the beginning of 2014 through the end of 2017 were taken as historical data. Each record contains information on the sales profit, the number of units sold and the change in its price for the specified period. A training sample consisting of 200 entries was created, based on the historical sales data and the collected numerical indicators of the impact factors described above (an example of data is presented in Table 1).

Table 1. Example of training sample (fragment)

| Date         | Quantity | Price per packing | Amount | Visits to the doctor | Current season | Qualitative distribution | Quantitative distribution | Share in full-scale volume | Share in cost volume | Percent of price change | Advertising |
|--------------|----------|-------------------|--------|-----------------------|----------------|--------------------------|--------------------------|-------------------------|---------------------|------------------------|-------------|
| 28.01.2014   | 357      | 258.06            | 92127.42 | 232                  | Winter         | 0.0244                  | 0.0574                  | 11.28                   | 16.96               | 0.4                    | 0           |
| 17.02.2014   | 540      | 257.10            | 138835.62 | 260                  | Winter         | 0.0253                  | 0.0809                  | 11.89                   | 18.09               | 0.6                    | 0           |
| 23.03.2014   | 260.15   | 140481            | 93574.8  | 239                  | Spring         | 0.0249                  | 0.0962                  | 11.66                   | 17.5                | 0.2                    | 0           |
| 07.05.2014   | 360      | 259.93            | 93574.8  | 239                  | Spring         | 0.0245                  | 0.1115                  | 11.28                   | 16.86               | 1.2                    | 0           |
| 19.06.2014   | 360      | 259.57            | 93444.12 | 283                  | Summer         | 0.0241                  | 0.0565                  | 11.04                   | 17.62               | 0.8                    | 1           |
| 14.07.2014   | 360      | 256.95            | 92501.64 | 280                  | Summer         | 0.0237                  | 0.0654                  | 11.16                   | 19.3                | -0.7                   | 1           |
| 28.01.2014   | 357      | 258.06            | 92127.42 | 232                  | Winter         | 0.0244                  | 0.0574                  | 11.28                   | 16.96               | 0.4                    | 0           |

5. Program implementation of the forecasting method taking into account the influence factors
An application was developed in the course of the study, to forecast drug sales using the R programming language R of the RStudio environment and the ShinyApps framework (figure 3).

The developed application receives a file with input data in the *.csv format on the basis of which the learning sample is formed, and a file in the *.csv format containing the forecast terms and known influence factors at that time. It is necessary to indicate the degree of learning of models as an input parameter by which the accuracy of the methods used will be evaluated [14].

6. Results of computational experiments
Two experiments’ sets were conducted to evaluate the effectiveness of the forecasting model construction using the identified impact factors, in which models were trained at different volumes of data sampling based only on historical data and taking into account the factors described above.

For the experiment, the following actions were performed:
1. Two data sets, which contain 40 and 80, were compiled from the generated sample.
2. Each set was divided into two parts - the training and test, in the approximately equal ratio for each data set. The degree of the model training should be about 90%.
Figure 3. Application for forecasting sales of medicines

3. A random forest model was constructed for each set on the basis of the training part; two models were obtained in the issue:
   a. A model built only on the historical data of the drug sales.
   b. A model built taking into account the selected factors of influence.

4. The forecast for the test part was constructed using two models for each data set.

5. The estimated sales forecast was assessed using the model identification criteria and the average deviation of the forecast sales figures from the actual for the last three records of the sample being tested.

6. The results were tabulated, analysed and became the basis for conclusions.

   The first experiment. The results of the first experiment are shown in Table 2. The volume of the sample is 40 records. The degree of the model training is 88%.

| Table 2. Results of the first experiment |
|------------------------------------------|
| Month         | RMSE  | MSE     | MAPE    |
| Fact          |       |         |         |
| October 2015  | 138027.78 | 138263.5  | 138266.7 |
| November 2015 | 134709.59 | 135279.55 | 145375   |
| December 2015 | 135248.1 | 137650.1  | 143622.5 |

   Forecast based on historical data

| Month         | RMSE  | MSE     | MAPE    |
| Fact          |       |         |         |
| October 2015  | 138027.78 | 138263.5  | 138266.7 |
| November 2015 | 134709.59 | 135279.55 | 145375   |
| December 2015 | 135248.1 | 137650.1  | 143622.5 |

   Forecast made taking into account impact factors

| Month         | RMSE  | MSE     | MAPE    |
| Fact          |       |         |         |
| October 2015  | 138027.78 | 138263.5  | 138266.7 |
| November 2015 | 134709.59 | 135279.55 | 145375   |
| December 2015 | 135248.1 | 137650.1  | 143622.5 |

The second experiment. The results of the second experiment are shown in Table 3. The volume of the sample is 80 records. The degree of model proficiency is 91%.

| Table 3. Results of the second experiment |
|------------------------------------------|
| Month         | RMSE  | MSE     | MAPE    |
| Fact          |       |         |         |
| October 2016  | 135123.12 | 308070.2  | 175858.4 |
| November 2016 | 177100.59 | 180766.86 | 140206.04 |
| December 2016 | 133821.69 | 300349.77 | 161108.75 |

   Forecast based on historical data

| Month         | RMSE  | MSE     | MAPE    |
| Fact          |       |         |         |
| October 2016  | 135123.12 | 308070.2  | 175858.4 |
| November 2016 | 177100.59 | 180766.86 | 140206.04 |
| December 2016 | 133821.69 | 300349.77 | 161108.75 |

   Forecast made taking into account impact factors

| Month         | RMSE  | MSE     | MAPE    |
| Fact          |       |         |         |
| October 2016  | 135123.12 | 308070.2  | 175858.4 |
| November 2016 | 177100.59 | 180766.86 | 140206.04 |
| December 2016 | 133821.69 | 300349.77 | 161108.75 |
7. Conclusions
The following conclusions can be drawn based on the results of experiments:
1. The indices of error identifying criteria of the second forecast model are lower than the first one.
2. The average difference between the actual and forecasted sales values for the three-month first forecast model is about 4,500 for the first experiment and about 68,000 – for the second one.
3. The average difference between the actual and forecasted sales values for the three-month second forecast model is less than 3000 for the first experiment and less than 8000 for the second one.

The results of the experiments showed that the forecast, taking into account the impact factors, turned out to be more accurate compared with the same based only on historical data.

Based on this, it is possible to conclude that for time series with a small set of records, the most effective model for predicting drug sales will be a model constructed taking into account the factors of influence.

References
[1] Mas-Machuca M, Sainz M, Martinez-Costa C 2014 A review of forecasting models for new products Intangible Capital 10(1) 1-25.
[2] Salnikova N A, Lempert B A, and Lempert M B 2015 Integration of Methods to Quantify the Quality of Medical Care in the Automated Processing Systems of Medical and Economic Information Communications in Computer and Information Science 535 307-319
[3] Clark T D 2013 Probability-based forecasting for U.S. generic drug sales Journal of Generic Medicines 10(3–4) 184–192
[4] Dyachenko T, Ivanenko V, Lempert B, and Salnikova N 2017 Dynamics of Health Care Quality Indicators at Inpatient Hospitals of the Volgograd Region Estimated by an Automated Information System Communications in Computer and Information Science 754 847-857
[5] Koulouriotis D E, Mantas G 2012 Health products sales forecasting using computational intelligence and adaptive neuro fuzzy inference systems Operational Research 12 (1) 29-43.
[6] Tahmasebi N, Zadeh A K, Imani A, Golestani M 2013 Evaluation of Factors Affecting Sales of Prescription Medicines by Econometric Methods in Iran Pharmaceutical sciences 19(3) 101-107
[7] Gerget O M, Maruhina O V, Cherkashina Yu A, Krivonogova T S 2016 Visualization systems and multivariate data analysis medical and social research Scientific Visualization 8 (4) 80-90
[8] Al-Gunaid M A, Shcherbakov M V, Kamaev V A, Gerget O M, Tyukov A P 2016 Decision Trees based Fuzzy Rules Information Technologies in Science, Management, Social Sphere and Medicine’ (ITSMSSM 2016) 502-508
[9] Jiang X-F 2012 Research on the prediction of drug sales based on Levenberg-Marquardt algorithm Applied Mechanics and Materials 198-199 1452-1456
[10] Khalil Zadeh N, Sepehri M M, Farvaresh H 2014 Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach Mathematical Problems in Engineering art. no. 420310
[11] Dai W 2015 A Clustering-based Sales Forecasting Scheme Using Support Vector Regression for Computer Server Procedia Manufacturing 2 82 – 86
[12] Orudjev N Y, Lempert M B, Osaulenko I, Salnikova N A, Kuzmichev A A, Kravets A G 2016 Computer-Based Visual Analysis of Ecology Influence on Human Mental Health. 7th International Conference on Information, Intelligence, Systems and Applications (IISA2016) art.no. 7785416
[13] Wagner N, Michalewicz Z 2011 Intelligent techniques for forecasting multiple time series in real-world systems International Journal of Intelligent Computing and Cybernetics 4(3) 284-310
[14] Kravets A, Poplavskaya O, Lempert L, Salnikova N, and Medintseva I 2017 The Development of Medical Diagnostics Module for Psychotherapeutic Practice Communications in Computer
and Information Science \textbf{754} 872-883