Original Article

Healthc Inform Res. 2021 October;27(4):279-286.
https://doi.org/10.4258/hir.2021.27.4.279
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Estimating the Optimal Dexketoprofen Pharmaceutical Formulation with Machine Learning Methods and Statistical Approaches

Atakan Başkor[1], Yağmur Pirinçci Tok[2], Burcu Mesut[2], Yıldız Özsoy[2], Tamer Uçar[3]
[1]Department of Big Data Analytics and Management, Institute of Science and Technology, Bahcesehir University, Istanbul, Turkey
[2]Department of Pharmaceutical Technology, Faculty of Pharmacy, Istanbul University, Istanbul, Turkey
[3]Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, Istanbul, Turkey

Objectives: Orally disintegrating tablets (ODTs) can be utilized without any drinking water; this feature makes ODTs easy to use and suitable for specific groups of patients. Oral administration of drugs is the most commonly used route, and tablets constitute the most preferable pharmaceutical dosage form. However, the preparation of ODTs is costly and requires long trials, which creates obstacles for dosage trials. The aim of this study was to identify the most appropriate formulation using machine learning (ML) models of ODT dexketoprofen formulations, with the goal of providing a cost-effective and time-reducing solution. **Methods:** This research utilized nonlinear regression models, including the k-nearest neighborhood (k-NN), support vector regression (SVR), classification and regression tree (CART), bootstrap aggregating (bagging), random forest (RF), gradient boosting machine (GBM), and extreme gradient boosting (XGBoost) methods, as well as the *t*-test, to predict the quantity of various components in the dexketoprofen formulation within fixed criteria. **Results:** All the models were developed with Python libraries. The performance of the ML models was evaluated with R^2 values and the root mean square error. Hardness values of 0.99 and 2.88, friability values of 0.92 and 0.02, and disintegration time values of 0.97 and 10.09 using the GBM algorithm gave the best results. **Conclusions:** In this study, we developed a computational approach to estimate the optimal pharmaceutical formulation of dexketoprofen. The results were evaluated by an expert, and it was found that they complied with Food and Drug Administration criteria.

Keywords: Machine Learning, Statistics, Data Analysis, Dexketoprofen Trometamol, Pharmaceutical Preparations

Submitted: January 18, 2021
Revised: 1st, May 14, 2021; 2nd, July 4, 2021; 3rd, August 17, 2021
Accepted: August 23, 2021

Corresponding Author
Tamer Uçar
Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, Istanbul, Turkey. Tel: +90-2123810575, E-mail: tamer.ucar@eng.bau.edu.tr (https://orcid.org/0000-0002-9397-6656)

I. Introduction

Dexketoprofen is a powerful pain reliever that dissolves well in water and has relatively few side effects. It is used for the symptomatic treatment of postoperative pain, musculoskeletal system pain, menstrual pain, and toothache in adults [1]. Dexketoprofen is a cyclooxygenase-1 (COX-1) and COX-2 inhibitor and has been proven to be an effective pain reliever in human clinical studies and animal studies. Clinical studies have demonstrated that the effectiveness of this pain reliever begins within 30 minutes and lasts for up to 6–8 hours. However, the active ingredient, which acts quickly and has high activity, has a very bitter taste [2].

Drug formulations, which are developed through research and development studies, are analyzed to obtain market approval for drugs. The data obtained after the analysis must be within the limits requested by the authorities. Pharmaceutical companies are therefore required to conduct numerous trials to reach the desired limits. However, the limited availability of raw materials before sale limits the number of trials. The expectation in pharmaceutical companies is to achieve results in a short time with relatively few trials and to reduce costs. Artificial intelligence approaches have been able to make positive contributions towards these goals [3].

Previous studies have investigated this issue. For instance, Dere and Ayvaz [4] evaluated the effectiveness of traditional approaches used to model drug-drug interactions (DDIs). The aim of their study was to provide a cost-effective and scalable solution to evaluate the effectiveness of similarity-based *in-silico* computational DDI estimation approaches and to estimate potential DDIs. Commonly known similarity-based computational DDI estimation methods were used to discover new potential DDIs. The drug interaction profile was found to be a better predictor of DDIs than drug side effects and protein similarities between DDI pairs [4].

Machine learning (ML) models are also used in many other areas in the health sector. For instance, an ML algorithm model was developed that could correctly diagnose coronavirus disease 2019 (COVID-19). Tree-based algorithms were used extensively in that study. The extreme gradient boosting (XGBoost) algorithm was found to predict the spread of COVID-19 with high accuracy; therefore, XGBoost can assist in the early detection of COVID-19 [5].

Another study used ML algorithms to evaluate creatinine, which plays a significant role in the detection of end-stage renal disease. Regression methods were used in the dataset. When the results were evaluated, the most important variable was determined to be blood urea nitrogen. Mixed-effects least-squares support vector regression was determined to be the best method to predict serum creatinine levels [6].

In this study, dexketoprofen data were first evaluated with statistical approaches during pre-processing. The *t*-test was applied using the SPSS program to analyze all output results. Subsequently, mostly tree-based ML algorithms are applied using the Python programming language to predict the dexketoprofen outputs. Seven ML methods were compared to find the best method for estimating the optimal dexketoprofen pharmaceutical dosage formulation. The predicted system output was tested by specialists in the laboratory and the results were evaluated.

This study has potential to save considerable time and monetary resources by making the process of manual formulation iterations obsolete.

II. Methods

1. Dataset Preparation

Each formulation (input) of the dexketoprofen dataset was prepared as a tablet. Granulation was done with two different Eudragit coating levels: low coating (group 1), 15.16%; high coating (group 2), 17.34%. The dexketoprofen inputs had Prosolv ODT (orally disintegrating tablet) (mg) as a filling material, Emdex (mg) as a flavoring, and MagnaSweet (%) to hide the bitter taste, and a tablet compression force (psi) was applied to finalize the tablets. The dataset contained 27 low-coating formulations and 27 high-coating formulations, resulting in a total of 54 formulations. The dataset had 10 different outputs: friability (%), hardness (N), weight variation (mg), and the dissolution rate (%). Details on the input variables in the tablet formulation preparation steps are shown in Table 1.

2. Statistical Analysis and ML Models used in the Dexketoprofen Dataset

Dexketoprofen pharmaceutical dosage formulation data were analyzed to establish the normality of the data distribution for the relationship between Eudragit coating values and each output. This analysis is important for the correct evaluation of the data. The appropriate test was then performed

Table 1. Formulation parameters in the preparation stage

Variable	Box Behnken	Eudragit 15.16%	Eudragit 17.34%
Prosolv ODT (mg)	-1	150	150
	0	200	200
	+1	250	250
Emdex (mg)	-1	100	100
	0	150	150
	+1	200	200
MagnaSweet (%)	-1	0.02	0.02
	0	0.13	0.13
	+1	0.24	0.24
Tablet compression force (psi)	-1	250	250
	0	500	500
	+1	750	750

ODT: orally disintegrating tablet.

according to whether the distribution was normal or non-normal. The Student *t*-test was used when the data had two independent groups with normal distributions [7,8]. Otherwise, the Mann-Whitney U test was used [9]. Moreover, if two Eudragit (15.16% and 17.34%) coating amounts are established to have differences from each other, that finding will provide support to make the right decisions in the next steps.

After the statistical analysis, seven different ML models were applied to the dataset: k-nearest neighbors (k-NN) [10], support vector regression (SVR) [11], classification and regression tree (CART) [12], bootstrap aggregating (bagging) [13], random forest (RF) [14], gradient boosting machine (GBM) [15], and XGBoost [16]. These models have some characteristic hyperparameters [17] to improve their prediction values. Each model was compared step by step for every output and the best model for each output was defined to predict the best formulation inputs.

1) Pre-processing for the dexketoprofen dataset

First, the Eudragit coating amounts were statistically compared in terms of hardness, friability, and disintegration time in SPSS. The normality assumption was checked for each output in group 1 and group 2 using histograms, the Q-Q plot, and the Shapiro-Wilk test ($n < 50$) in order to determine whether the group had a normal or non-normal data distribution [18].

Based on the normality tests, the Student *t*-test and Levene test [19] were implemented for hardness values. The Mann-Whitney U test was implemented for friability and disintegration time values.

The coating proportions are important for stomach absorption and for hiding the bitter taste in the mouth. Knowing whether there is a difference between the coatings will support subsequent production steps.

2) ML models for the dexketoprofen dataset

The dataset was imported into the Python program after splitting it into training (85%) and testing (15%) sets. The dataset contained 54 different formulations, 45 of which were used for training and nine of which were used for testing.

Every model had different characteristic parameters to improve its own performance measures. The parameters were defined for seven ML algorithms, as shown in Table 2.

All models were applied individually, using the appropriate hyperparameters. The ML models were trained with data from a limited number of trials and the best ones were chosen. In order to find the optimal values or best formulation values, intermediate input values must be known. Therefore, intermediate values between the values of each input were produced. For example, Prosolv ODT (mg) was produced with values between 150 and 250. Thus, a total of 2,500,000 intermediate input values were produced. Estimates were made across the intermediate values produced by the models with the best output.

Table 2. Hyperparameter selection in machine learning models

Model	Parameter	Value
k-NN	N Neighbors	Range (1, 20)
SVR	Kernel	Radial basis function
	C	Range (1, 300)
CART	Min samples split	Range (2, 100)
	Max leaf nodes	Range (2, 20)
Bagging	N Estimators	Range (2, 100)
RF	Max depth	Range (1, 10)
	Max features	Range (1, 6)
	N Estimators	200, 500, 700, 1000
GBM	Learning rate	0.001, 0.01, 0.1, 0.2
	Max depth	3, 5, 8, 50, 100
	N Estimators	200, 500, 1000, 2000
	Subsample	1, 0.5, 0.75
XGBoost	Colsample bytree	0.7, 0.3, 0.1, 0.8, 0.9
	Max depth	30, 25, 20, 5, 8, 10, 15
	N Estimators	100, 200, 500, 1000
	Learning rate	0.0005, 0.005, 0.0001, 0.001, 0.1, 0.01, 0.5

k-NN: k-nearest neighbors, SVR: support vector regression, CART: classification and regression tree, Bagging: bootstrap aggregating, RF: random forest, GBM: gradient boosting machine, XGBoost: extreme gradient boosting.

3) Evaluation criteria

The performance of models was evaluated by using R^2 (coefficient of determination) and root mean square error (RMSE). The model with the best R^2 and RMSE for each output was selected and saved for final predictions. The obtained results were filtered according to the criteria determined by the Food and Drug Administration (FDA) [20] and the International Council for Harmonisation (ICH) Q6 series [21]. These criteria were friability <1%, disintegration <30 seconds, and a dissolution rate of 100%. The inputs determined after the filtering process were transferred to experts for testing.

III. Results

1. Statistical Analysis of the Dexketoprofen Dataset

1) Hardness values between groups

The tablet coating amount of Eudragit (low vs. high levels) was found to be normally distributed for the hardness output; therefore, the *t*-test was implemented with the Levene test, which yielded a *p*-value of 0.435. Since this value was greater than 0.05, the variance was equal between groups. The *p*-value obtained using the *t*-test (0.504) was substantially greater than 0.05. Therefore, hardness showed no statistically significant difference between the tablet coating groups at a 5% significance level.

2) Friability values between groups

The amount of Eudragit tablet coating (low vs. high levels) was not found to be normally distributed for friability; therefore, the Mann-Whitney U test was implemented, yielding a *p*-value (0.640) that substantially exceeded the threshold of 0.05. Thus, friability had no statistically significant difference between tablet coating groups at a 5% significance level.

3) Disintegration time values between groups

The Eudragit coating (low vs. high levels) did not show a normal distribution for disintegration time; therefore, the Mann-Whitney U test was implemented, resulting in a *p*-value (0.993) that was significantly higher than 0.05. Therefore, disintegration time showed no statistically significant difference between the tablet coating groups at a 5% significance level.

2. Machine Learning Models Based on the Dexketoprofen Dataset

Regarding the results in Table 2, the output variable of hardness had an explanatory power (R^2) of 99% and an RMSE of 2.88. For friability, the model's explanatory power was 92% (R^2) and the RMSE was 0.02. For disintegration time, the model's explanatory power (R^2) was 97% and the RMSE was 10.09. The explanatory power for dissolution varied based on the time range; as shown in Table 2, the RMSE values were distributed between 1.89 and 5.92 and the R^2 values ranged from 0.65 to 0.94. All ML model results of the outputs are shown in Table 3 in detail.

In most cases, the results also had logical interpretations in addition to the numerical values. Although the ultimate target is to determine output predictions, the overall model evaluation is also important. Therefore, the importance of each feature for label prediction is also part of the total analysis. The feature importance of inputs for the dependent variables was calculated using the "feature_importance" property of the *sci-learn* library by choosing the model with the best predictive success for the dependent variable. The important point here is to find the model with the lowest RMSE and then determine input importance using the related properties. Graphical interpretations in terms of feature importance for hardness, friability, and disintegration time outputs are given in Figures 1–3.

The dexketoprofen dataset has limited observation capabilities and time constraints, resulting in reduced iteration opportunities. Given this context, optimal formulation prediction is a challenge. In order to cope with this challenge, global optimal values are targeted. However, a limitation of this study is that it is very difficult to find targeted global optimal values with limited data. The best model was selected and new iterations were repeatedly executed to identify new formulations according to each output. The features given in Table 1 are predicted with the best algorithms for each output using intermediate values to create actual values. For instance, an attempt was made to predict Prosolv ODT values between 150 and 250 mg using trained models that are generated via default ones. This analysis provided a total instance number of 2,500,000 for the best prediction results. The algorithm outputs for friability, hardness, disintegration time, and dissolution rate within the fixed constraints were generated for new active pharmaceutical ingredient formulations [22].

The predicted formulation according to the best algorithms is given in Table 4, as well as friability, hardness, and disintegration values in Table 5. The similarities in dissolution time are given in Table 6 and values are visualized in Figure 4. The actual values and predicted values for dissolution time were compared using the *t*-test to determine whether there were statistically significant differences between them. The *t*-test *p*-value was 0.548, which exceeded 0.05, meaning that there was no significant difference between the actual values and predicted values.

Figure 4 clearly shows the closeness between the actual value and the predicted values. In addition, the study was evaluated by an expert. The tablets also met the FDA criteria by dissolving by more than 85% in 15 minutes [23].

IV. Discussion

ML algorithms have been successfully used to determine optimal values for new medicine formulations in the medical industry, indicating that it is also possible to use algorithms

Table 3. Coefficients of determination and RMSE performance of all models for the outputs

Outputs	k-NN		SVR		CART		Bagging		RF		GBM		XGBoost	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Hardness	0.94	7.47	0.84	12.75	0.99	3.59	0.98	4.91	0.98	3.80	**0.99**	**2.88**	0.98	4.60
Friability	0.90	0.03	0.75	0.05	0.84	0.05	0.83	0.04	0.81	0.04	**0.92**	**0.02**	0.82	0.05
Disintegration time	0.96	12.34	0.59	25.52	0.85	19.96	0.85	21.86	0.69	35.07	**0.97**	**10.09**	0.96	10.05
Dissolution time														
1 min	0.89	4.43	0.42	5.22	0.83	4.40	0.72	6.17	0.89	5.35	0.94	2.61	**0.92**	**1.89**
3 min	0.90	4.72	0.71	9.09	0.90	5.42	0.77	6.51	0.81	9.15	0.84	8.31	**0.94**	**4.00**
5 min	0.78	3.85	0.31	9.50	0.79	8.25	0.37	14.07	0.61	7.03	0.71	9.86	**0.83**	**4.69**
10 min	0.64	4.44	0.56	5.06	0.56	6.37	0.41	6.98	**0.67**	**4.29**	0.61	6.00	0.56	4.67
15 min	0.57	5.41	0.46	7.02	0.52	8.53	0.50	3.85	0.46	6.69	**0.66**	**5.92**	0.47	7.07
20 min	0.75	4.74	0.12	8.65	0.51	7.42	0.45	6.84	0.57	6.20	**0.65**	**5.58**	0.57	5.68
30 min	0.74	5.88	0.40	6.72	0.58	5.57	0.47	6.76	0.45	5.01	**0.85**	**3.57**	0.80	3.95

Bold text indicates the best performing models.

RMSE: root mean square error, k-NN: k-nearest neighbors, SVR: support vector regression, CART: classification and regression tree, Bagging: bootstrap aggregating, RF: random forest, GBM: gradient boosting machine, XGBoost: extreme gradient boosting.

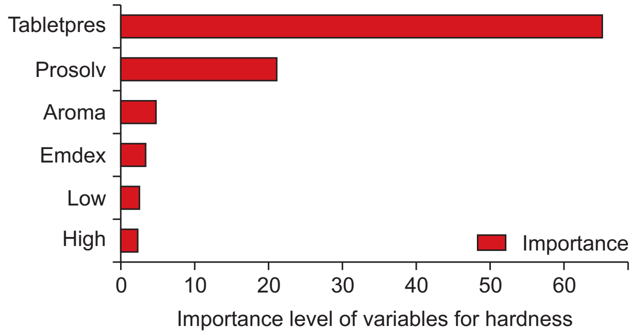

Figure 1. Importance of variables for hardness.

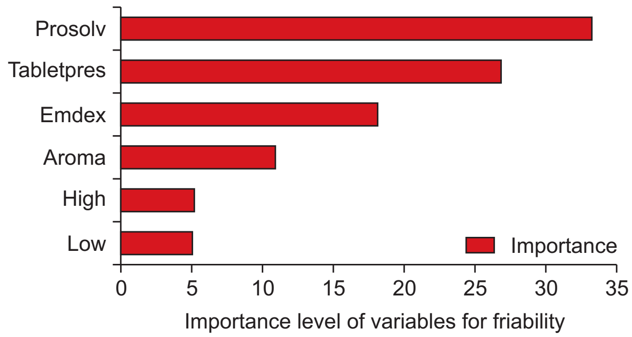

Figure 2. Importance of variables for friability.

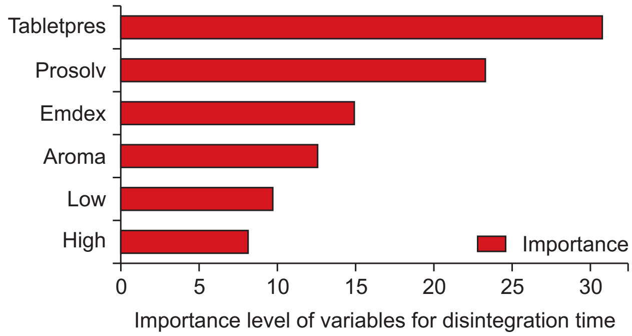

Figure 3. Importance of variables for disintegration time.

Table 4. Recommended formulation according to the algorithm

Coating	Eudragit 17.34%
Prosolv ODT (mg)	150
Emdex (mg)	176
MagnaSweet (%)	0.02
Tablet compression force (psi)	250

ODT: orally disintegrating tablet.

Table 5. Friability, hardness, and disintegration results

	Value
Friabilit (%)	0.43
Hardness (n)	68
Disintegration (s)	86

Table 6. Algorithm-predicted and actual results for dissolution

Time (min)	Algorithm prediction for dissolution rate (%)	Actual dissolution rate (%)
1	42.51434300	61.0439230
3	84.59233000	87.1458178
5	88.46628600	87.9163231
10	80.44402607	87.4055398
15	86.24988264	88.0383315
20	85.60514994	86.1494829
30	83.68538536	84.8540741

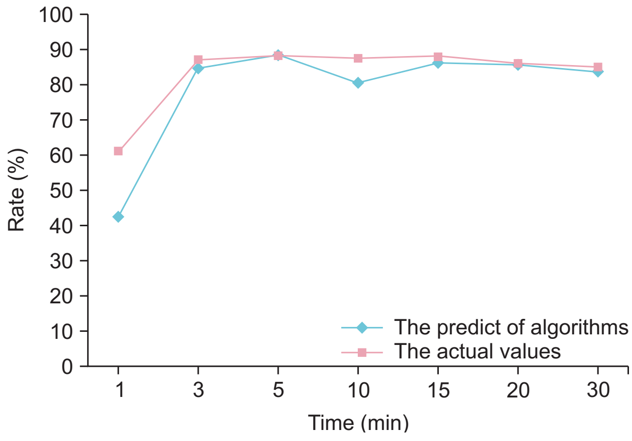

Figure 4. Line chart of the algorithmic prediction of the dissolution rate and actual dissolution rate in ratios over time.

in this sector in parallel to other working fields. Tree-based models have better predictive results in comparison to other models. GBM and XGBoost yielded better predictive results than other tree-based models. Moreover, statistical normalization and the *t*-test were successfully implemented beforehand in the pre-evaluation period. The proposed approach in this study has eliminated the necessity for many trials, and prevented the use of a limited amount of active ingredients, which would have significant impacts in terms of cost and time. However, a limitation of the study is the difficulty of finding targeted global optimal values with limited data. The other principal difficulty of the study is that the new dataset produced for forecasting was too large. The creation of a new method in this area can provide faster results. This research program can be improved via the development of new models and statistical analysis to use new medicine formulations as per specific requirements for the relevant analysis.

https://doi.org/10.4258/hir.2021.27.4.279

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This study was a part of the first author's master's thesis on big data analytics and management.

ORCID

Atakan Başkor (https://orcid.org/0000-0002-4739-4700)
Yağmur Pirinçci Tok (https://orcid.org/0000-0001-6915-0283)
Burcu Mesut (https://orcid.org/0000-0003-2838-1688)
Yıldız Özsoy (https://orcid.org/0000-0002-9110-3704)
Tamer Uçar (https://orcid.org/0000-0002-9397-6656)

References

1. Ezcurdia M, Cortejoso FJ, Lanzon R, Ugalde FJ, Herruzo A, Artigas R, et al. Comparison of the efficacy and tolerability of dexketoprofen and ketoprofen in the treatment of primary dysmenorrhea. J Clin Pharmacol 1998; 38(S1):65S-73S.
2. Kara I, Tuncer S, Erol A, Reisli R. [The effects of preemptive dexketoprofen use on postoperative pain relief and tramadol consumption]. Agri 2011;23(1):18-21.
3. Mesut B, Aksu N, Ozsoy Y. Design of sustained release tablet formulations of alfuzosin HCl by means of neurofuzzy logic. Lat Am J Pharmacy 2013;32(9):1288-97.
4. Dere S, Ayvaz S. Prediction of drug-drug interactions by using profile fingerprint vectors and protein similarities. Healthc Inform Res 2020;26(1):42-9.
5. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, Uddin S, Lio P, Xu H, et al. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl 2020;160:113661.
6. Amiri MM, Tapak L, Faradmal J, Hosseini J, Roshanaei G. Prediction of serum creatinine in hemodialysis patients using a kernel approach for longitudinal data. Healthc Inform Res 2020;26(2):112-8.
7. Vetter TR. Fundamentals of research data and variables: the devil is in the details. Anesth Analg 2017;125(4): 1375-80.
8. Kim TK. T test as a parametric statistic. Korean J Anesthesiol 2015;68(6):540-6.
9. McKnight PE, Najab J. Mann-Whitney U test. In: Weiner IB, Edward Craighead W, editors. The Corsini encyclopedia of psychology. Hoboken (NJ): John Wiley & Sons; 2010.
10. Adebowale A, Idowu SA, Amarachi A. Comparative study of selected data mining algorithms used for intrusion detection. Int J Soft Comput Eng 2013;3(3):237-41.
11. Gunn SR. Support vector machines for classification and regression. Southampton, UK: University of Southampton; 1998.
12. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14(1):1-37.
13. Breiman L. Bagging predictors. Mach Learn 1996;24(2): 123-40.
14. Breiman L. Random forests. Mach Learn 2001;45(1):5-32.
15. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189-232.
16. Chen T, Guestrin C. XgBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. p. 785-94.
17. Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 2019;20(1):1934-65.
18. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 2012;10(2):486-9.
19. O'Neill ME, Mathews KL. Levene tests of homogeneity of variance for general block and treatment designs. Biometrics 2002;58(1):216-24.
20. US Food and Drug Administration. Q6A specifications: test procedures and acceptance criteria for new drug substances and new drug products: chemical substances [Internet]. Silver Spring (MD): Food and Drug Administration; 2000 [cited at 2021 Oct 10]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q6a-specifications-test-procedures-and-acceptance-criteria-new-drug-substances-and-new-drug-products.
21. European Medicines Agency. ICH topic Q6a specifications: Test procedures and acceptance criteria for new drug substances and new drug products: Chemical substances [Internet]. London, UK: European Medicines Agency; 2000 [cited at 2021 Oct 10]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-q-6-test-procedures-acceptance-criteria-new-drug-substances-new-drug-products-chemical_

en.pdf.

22. Mesut B, Baskor A, Pirincci Tok Y, Alkan B, Erginer Y. Statistical investigation of the effect of excipients particle size on orally disintegrating tablets: mannitol grades. Informatica 2020;31(7):69-91.
23. Center for Drug Evaluation and Research. Scale-up and postapproval changes: chemistry, manufacturing, and controls: in vitro dissolution testing, and in vivo bioequivalence documentation [Internet]. Silver Spring (MD): Food and Drug Administration; 1995 [cited at 2021 Oct 10]. Available from: https://www.fda.gov/media/70949/download.