Prediction model of ischemic stroke based on machine learning

Zhijie Zhang
Lingnan Normal University
Zhihong Zou (zouhong1118@126.com)
Shenzhen Traditional Chinese Medicine Hospital
Junjie Yang
Lingnan Normal University
Xin Guan
Lingnan Normal University
Xialan Chen
Lingnan Normal University

Research Article

Keywords: Extreme learning machine, stroke, predict, risk factors

Posted Date: April 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1452411/v1

Prediction model of ischemic stroke based on machine learning

Zhijie Zhang[1], Zhihong Zou[2,*], Junjie Yang[1], Xin Guan[1], Xialan Chen[3]

1 School of Computer Science and Intelligence Education, Lingnan Normal University, 524048 Zhanjiang, China

2 Shenzhen Traditional Chinese Medicine Hospital, 518033 Shenzhen, China

3 School of Life Science and Technology, Lingnan Normal University, 524048 Zhanjiang, China

* Corresponding author: Zhihong Zou (e-mail: zouhong1118@126.com)

Abstract

Background Existing research focuses on the identification of key risk factors for stroke, improves the accuracy of stroke risk prediction, and provides more evidence for the scientific diagnosis, treatment and intervention of stroke.

Methods We included 4785 cases, stratified by sex (men: n=3156; women: n=1629) and age (18-40 years; 41-54 years; 55-69 years; 70+ years), the survey data of stroke patients conducted by Cooperative Hospital from 2019 to 2020. After data preprocessing, an extreme learning machine was used to construct a stroke risk prediction model.

Results The stroke risk prediction model was identified that total cholesterol and high-density lipoprotein were the 10 most important risk factors affecting the onset of stroke. The prediction accuracy rate of the risk prediction model is 97%.

Conclusion The method in this paper can quickly and effectively dig out the key risk factors affecting the onset of stroke from the data, and predict the risk of the onset, which has high application value.

Key Words: Extreme learning machine—stroke—predict—risk factors

Highlights

- This study constructed an accurate and efficient stroke prediction model based on extreme learning machine.
- The ten important features were discovered, which provide support for the rapid construction of stroke risk prediction models.
- Our research will provide support that can be applied to clinical practice, that is, it will help doctors quickly identify people at high risk of stroke and make decisions.
- Data preprocessing are known to reduce the impact on the model prediction results and we will account for this in the analysis.
- We will address the potential limitation of important feature detection by performing feature selection methods.

Background

Stroke is caused by the development of cerebrovascular disease to a certain extent, with high morbidity, disability and mortality, and has become a major disease that seriously endangers human health and life safety in the world today.

At present, stroke is the chronic disease with the highest rate of death and disability in my country. In recent years, the incidence rate has continued to increase. It is estimated that the incidence of cerebrovascular disease events in my country in 2030 will increase by about 50%

compared with 2010.

As patients with stroke are irreversible, difficult to heal, and high nursing costs, the medical burden is showing an increasing trend. Therefore, early prevention of stroke is particularly important.

Stroke prevention research generally uses statistical methods to model the relationship between risk factors and disease incidence, and then quantify the risk of disease incidence, so as to carry out stroke and high-risk population health management [1-3].

At present, for ischemic stroke, stable scoring methods and predictive models have been constructed [4-9]. Experts generally recommend choosing modified Framingham stroke scale, pooled cohort equation, stroke risk calculator and other tools for ischemic stroke risk assessment.

With the development of big data technology, the latest research focuses on the tracking and processing of medical electronic medical record data of stroke patients, using machine learning, data mining and other methods to build a high-precision stroke risk prediction model [10-12].

In stroke prediction methods, more methods are used: artificial neural network method, decision tree method, logistic regression method, etc.

Methods

Study settings and design

In this study, the decision tree method is a fixed analysis that puts all variables at the same level, and the algorithm of the decision tree structure has "sequential bias". It is recommended to combine decision trees with other models [13-14].

Compared with logistic regression, in the classification of multi-dimensional and non-linear medical data, the classification performance, generalization ability, modeling calculation ability of logistic regression does not have obvious advantages [15-16].

The extreme learning machine algorithm can incorporate more risk factors, realize complex nonlinear mapping, automatically extract appropriate solving rules, and has good generalization, generalization and learning capabilities.

This study uses an extreme learning machine to complete the modeling. The model can reasonably classify data in function, and has better nonlinearity, generalization and fault tolerance in performance, which is better than ordinary ANN models. For the identification of risk factors, this article refers to the relevant literature on stroke risk factors and the standards of related research on the impact of cardiovascular and cerebrovascular diseases [17], through extreme learning machine learning modeling, in order to obtain a more complete risk prediction model.

By scientifically assessing the risk status of stroke, identifying high-risk stroke patients, moving the prevention and treatment center of stroke disease forward, and turning passive disease treatment into active health management.

Statistical analysis

The samples of this study are derived from the survey data of stroke patients conducted by China National Stoke Registry from 2019 to 2020. The scope of the research object is: residents who are 16 years old and above, have lived in the local area for more than 5 years, have a clear consciousness, and are sick or not sick for the first time. The length of the data set collection is 10 months, and a total of 4785 valid sample data were collected. The detailed relevant information of the data set is shown in Table 1.

Table1 Dataset description.

Attribute	Type	Unit	Value
Age	Numerical	Years	18-40; 41-54; 55-69; ≥70
Gender	Nominal	None	Male=1; Female=0
Body mass index(BMI)	Numerical	Kg/m^2	[12.7, 68.4]
Hypertension	Nominal	None	Yes=1; No=0
Diabetes	Nominal	None	Yes=1; No=0
History of stroke	Nominal	None	Yes=1; No=0
Heart disease	Nominal	None	Yes=1; No=0
History of smoking	Nominal	None	Yes=1; No=0
History of drinking	Nominal	None	Yes=1; No=0
Hyper homocysteinemia	Nominal	None	Yes=1; No=0
Hyperlipemia	Nominal	None	Yes=1; No=0
High uric acid	Nominal	None	Yes=1; No=0
Arteriosclerosis/stenosis/occlusion	Nominal	None	Yes=1; No=0
Systolic blood pressure	Numerical	mmHg	[48, 249]
Diastolic blood pressure	Numerical	mmHg	[35, 180]
White blood cell count	Numerical	$10^9/L$	[0.23, 27.8]
Total cholesterol	Numerical	mmol/L	[0.71, 9.4]
Triglycerides	Numerical	mmol/L	[0.05, 19.7]
High density lipoprotein	Numerical	mmol/L	[0.27, 3.15]
Low density lipoprotein	Numerical	mmol/L	[0.11, 8.79]
Albumin	Numerical	g/L	[28.3, 65.8]
Homocysteine	Numerical	umol/L	[0.09, 28.74]
Sodium	Numerical	mmol/L	[89.2, 153.8]
Potassium	Numerical	mmol/L	[1.74, 6.92]
Stroke	Nominal	None	Yes=1; No=0

Among them, there were 3156(65.96%) males and 1629 (34.04%) females, with an average age of (63.30 ± 16.60) years. The data set mainly contains the following content: (1) Basic personal information (gender, age, heart disease, hypertension, diabetes, smoking history, drinking history, Hyper homocysteinemia, Hyperlipemia, High uric acid, BMI, etc.); (2) Physical examination data (systolic blood pressure, diastolic blood pressure, etc.); (3) Laboratory examination data (triglycerides, total cholesterol, high density lipoprotein, low density lipoprotein, homocysteine, white blood cell count, Albumin, Sodium, Potassium, etc). The survey data is based on the information obtained during the last survey.

Model construction

Based on the high-dimensional and nonlinear characteristics of the effective data set obtained after preprocessing, this study uses an extreme learning machine to construct a stroke risk prediction model.

Extreme learning machine (ELM) was proposed for training single hidden layer feedforward neural networks (SLFNs); it can act as an efficient learning solution for regression problem [18].

The essence of ELM is that: unlike the common understanding of learning, the hidden layer of SLFNs should not be tuned. Considering N training data $\left\{\left(x_i,t_i\right)\mid x_i\in R^n,t_i\in R^m\right\}_{i=1}^{N}$, if an SLFN with L hidden nodes can approximate the mentioned N samples with zero error, it implies the existence of β,w and b; thus, it yields

$$f\left(x_i\right)=h\left(x_i\right)^T\beta=\sum_{j=1}^{L}\beta_jG\left(w_j,b_j,x_i\right),\quad i=1,...,N \tag{1}$$

Where $\beta_j=\left[\beta_{j1},...,\beta_{jm}\right]^T$ denotes the vector of the output weights between the hidden layer and the output layer, $w_j=\left[w_{j1},...,w_{jn}\right]^T$ is the input weights connecting input nodes with the j th hidden node, b_j represents the threshold of the j hidden node, and $G\left(w_j,b_j,x_i\right)$ is the activation function (e.g., $G\left(w_j,b_j,x_i\right)=1\Big/\left(1+\exp\left(-\left(w_j^T\cdot x_i+b_j\right)\right)\right)$) satisfying ELM universal approximation capability theorems.

To enhance the generalization ability of the traditional SFLNs based on ELM, Huang et al. [19] proposed the equality constrained optimization-based ELM. In their approach, structural risk considered as the regularization term is introduced. The so-called RELM is capable of regulating the proportion of structural risk and empirical risk using the parameter C. The proposed constrained optimization can be formulated as

$$\begin{aligned}&\min\ L_{RELM}=\frac{1}{2}\|\beta\|^2+C\frac{1}{2}\sum_{i=1}^{N}\|\xi_i\|^2\\&s.t.\ h\left(x_i\right)^T\beta=t_i-\xi_i\quad i=1,...,N\end{aligned} \tag{2}$$

Where ε_i denotes the slack variable of the training sample x_i and C controls the tradeoff between the output weights and the errors. Eq. (2) is similar to the classical optimization problem of Support vector machine (SVM), despite the simpler constraints, and it is valid for regression, binary, and multiclass cases [20]. Thus, (2) achieves a solution in the closed form

$$\beta=H^T\left(HH^T+\frac{I}{C}\right)^{-1}T \tag{3}$$

Where $H=\left[h\left(x_1\right),...,h\left(x_N\right)\right]_{N\times L}^T$ denotes the hidden layer output matrix, I indicates the identity matrix and $T=\left[t_1,...,t_N\right]_{N\times m}^T$. The RELM output function can be further derived as

$$\begin{aligned}&f\left(x_i\right)=h\left(x_i\right)^T\beta=h\left(x_i\right)^TH^T\left(HH^T+\frac{I}{C}\right)^{-1}T\\&i=1,...,N\end{aligned} \tag{4}$$

The extreme learning machine is a feed-forward neural network, and the parameter selection has an important influence on its prediction effect. Because the multi-hidden-layer model is complex and difficult to solve, this study chooses one hidden layer. Based on the Loppmann method, the range of the optimal number of neurons in the hidden layer is determined to be [21-22], and the number of neurons is adjusted in turn to train the model, and the optimal number of neurons is 9. The dependent variable in this study is a binary variable, and the activation functions

are $G(w,b,x)=1/\left(1+\exp\left(-\left(w\cdot x+b\right)\right)\right)$, and the prediction accuracy of this model are better. The input weight and hidden layer threshold of the network are randomly generated, the value range is [-1,1], the ratio of training sample and test sample data is set to 7:3, and the learning rate is determined to be 0.9 based on the experience value.

Results

Aiming at the problems of unstructured data such as electronic medical records and inspection results, such as standard irregularities, data missing, data noise, and system deviations, this study adopts data preprocessing operations such as data cleaning, integration, and dimensionality reduction. By discarding and forcibly replacing the data with many missing values and obviously unreasonable, and using the average value to fill in the missing data of continuous variables, and using the maximum and minimum method for normalization, 3962 valid data were finally obtained.

In view of the extreme learning machine is a supervised learning algorithm, the An assessment matrix is selected as the evaluation standard, and the loop debugging method is used to determine the optimal parameter value.

An assessment matrix [23] is illustrated in Table 2 as an effective measurement method to assess the experimental results. In this matrix, the true positive (TP) reveals that the stroke is correctly classified, the false negative (FN) means that the stroke is misclassified into non-stroke, the false positive (FP) denotes that the non-stroke is misclassified into stroke, and the true negative (TN) reveals that the non-stroke are classified accurately.

TABLE 2. The assessment matrix.

	Assessment	
	Stroke	Non- stroke
Actual stroke	TP	FN
Actual non- stroke	FP	TN

The accuracy is indicated by the percentage of correctly identified examples in the total number of examined examples, as expressed by Eq. (35).

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

The TPR is the ratio of correctly classified stroke to the total number of actual stroke, as defined by Eq. (36).

$$TPR=\frac{TP}{TP+FN} \tag{6}$$

The precision is indicated by the proportion of correctly classified stroke to the total number of data that are classified as stroke, as expressed in Eq. (37).

$$Precision=\frac{TP}{TP+FP} \tag{7}$$

The F-measure is the assessment accuracy combining both the precision and TPR, as calculated by Eq. (38).

$$F-measure=\frac{2*TPR*Precision}{TPR+Precision} \tag{8}$$

Use Python language to build a stroke risk prediction model, and conduct model training and

testing. The results of an assessment matrix are shown in Table 3. The accuracy value of ELM is 0.97, the TPR value is 0.99, the precision value is 0.94, and the F-measure value is 0.96, as shown in Figure 1. The prediction accuracy value of the support vector machines (SVM) and decision tree (DT) is moderately, and the prediction value of the logistic regression (LR) is low. The experimental results show that the stroke risk prediction model based on ELM has a high accuracy rate and can accurately predict stroke population and non-stroke population.

Table 3 The results of model prediction based on ELM.

Predicted value	Stroke	Non-stroke	All
Stroke	508	5	513
Non-stroke	30	645	675
All	538	650	1188

Through the analysis of variance and the chi-square test, the features with relatively low importance are removed [24], and 10 important features are obtained, as shown in Figure 2. The patient's hypertension, stroke history, age, total cholesterol, triglycerides, high density lipoprotein, Homocysteine, heart disease, low density lipoprotein, Hyperlipemia are the top 10 significant features that cause stroke. According to the different number of features in the dataset, LR, DT, SVM and ELM algorithm are combined to predict the risk of stroke respectively, as shown in Table 4.

TABLE 4. The classification results from different kinds of features in dataset.

Algorithm	Accuracy (%)				
	All features	Top 10	Laboratory features	Physical features	Basic features
LR	71.47	61.85	58.29	50.93	47.24
SVM	77.63	68.14	64.17	55.08	54.73
DT	82.75	73.91	67.84	58.49	57.63
ELM	**97.10**	**87.53**	**75.29**	**66.52**	**64.91**

Discussion

The study provides a one-year clinical data as a basis. The collected data set contains a total of 24 index parameters including Basic personal information, Physical examination data, and Laboratory examination data, which can accurately describe the influencing features of stroke populations. When the ELM algorithm is used for prediction, if all the attribute features in the data set are used, the prediction accuracy is as high as 97%. When the 10 important features are used for prediction, the prediction accuracy is 88%. When using Basic personal information, Physical examination data and Laboratory examination data as attributes to predict, the prediction accuracy is 75%, 67% and 65% respectively. When LR, DT and SVM algorithms are used for prediction, the prediction accuracy will also vary with the number of selected features, as shown in Table 4. Our results show that the number of features in the dataset is closely related to the prediction accuracy of the model, and a more comprehensive number of features is conducive to obtaining accurate prediction results. Among the four machine learning algorithms, when all features are used, the prediction accuracy is the highest; when the 10 important features are used, the prediction accuracy is also excellent.

The study uses ELM to predict the risk of stroke. Compared with traditional feedforward neural networks, the training speed is slow, the local minimum is easily trapped, and the learning rate is

sensitive. The ELM algorithm randomly generates the input layer and the hidden layer. The connection weight and the threshold value of hidden layer neurons, and there is no need to adjust during the training process, only need to set the number of hidden layer neurons, and the only optimal solution can be obtained. At the same time, compared with other traditional machine learning algorithms, ELM has the advantages of high learning efficiency and strong generalization ability. As shown in Figure 1, in accuracy, TPR, precision, F-measure and other parameters, the performance of ELM is better than LR, DT and SVM. Our research results show that ELM has excellent performance in the construction of stroke risk prediction model.

The study showed that hypertension, stroke history, age, total cholesterol, triglycerides, high density lipoprotein, Homocysteine, heart disease, low density lipoprotein, and Hyperlipemia are the 10 important features of the stroke risk, which have important clinical significance. Doctors and patients can formulate and optimize diagnosis and treatment plans in a timely manner by observing the characteristic values of important features, which is conducive to controlling the risk of disease and improving the probability of survival.

Conclusions

In conclusion, a multi-layer extreme learning machine model for predicting the risk of ischemic stroke was constructed, and 10 important features affecting the onset of stroke were screened out. This model can quickly and effectively dig out the key features affecting the onset of stroke from a large amount of data, and predict the risk of the onset of stroke, which has good practical application value.

The contributions of this research are as follows:

(1) The constructed risk prediction model has a high prediction accuracy rate and will provide more bases for the scientific diagnosis, treatment and intervention of stroke.

(2) Improve the identification of stroke risk factors, and identify low-density lipoprotein, total cholesterol, blood creatinine, triglycerides, stroke history, etc., as the 10 most important features affecting the incidence of stroke.

(3) Assist doctors in decision-making, pay attention to high-risk groups, detect and prevent diseases in time.

Abbreviations

ML: Machine learning; ELM: Extreme learning machine; SLFNs: Single hidden layer feedforward neural networks; SVM: Support vector machine; TP: True positive; FN: False negative; FP: False positive; TN: True negative; DT: Decision tree; LR: Logistic regression.

Declarations

Ethics approval and consent to participate

This article is a research project supported by the Lingnan Normal University and Shenzhen Traditional Chinese Medicine Hospital. All methods of the present study were performed in accordance with the relevant guidelines and regulations of the ethical committee of Lingnan Normal University, Shenzhen Traditional Chinese Medicine Hospital, and China National Stroke Registry, and also this study was approved by the ethical committee of Lingnan Normal University.

Participation was voluntary, the informed consent was verbal and written, for participants aged 16

and below from the parents, but all participants responded via email or text message to approve their participation. Participants had the right to withdraw from the study at any time without prejudice.

Consent for publication

Not applicable.

Availability of data and materials

All data generated and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study is supported by 2020 Zhanjiang City Science and Technology Development Special Fund Competitive Allocation Project (grants no. 2020A01043); 2020 Zhanjiang City Non-funded Science and Technology Research Project (grants no. 2020B01001).

Authors' contributions

Z.J.Z. was responsible for the study design, data analysis, and initial drafting of the manuscript. Z.H.Z. developed the model system. J.J.Y. conceived the study, guided the study design. X.G. helped to analyze the data. X.L.C. helped to write the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the Lingnan Normal University for financially supporting this project. We also would like to thank all experts who participated in this study.

Author information

School of Computer Science and Intelligence Education, Lingnan Normal University, 524048 Zhanjiang, China

Zhijie Zhang, Junjie Yang, Xin Guan

Shenzhen Traditional Chinese Medicine Hospital, 518033 Shenzhen, China

Zhihong Zou

School of Life Science and Technology, Lingnan Normal University, 524048 Zhanjiang, China

Xialan Chen

Corresponding author

Correspondence to Zhihong Zou.

References

[1] Hijazi Z, Lindb?Ck J , Alexander J H , et al. The ABC (age, biomarkers, clinical history) stroke risk score: a biomarker-based risk score for predicting stroke in atrial fibrillation.

European Heart Journal, 2016(20):1582-1590.

[2] Wang Y, Wang J, Cheng J, et al. Is the Population Detected by Screening in China Truly at High Risk of Stroke?. Journal of Stroke and Cerebrovascular Diseases, 2018, 27(8): 2118-2123.

[3] Shao Zeguo, Chen Chen, Chen Wei. Analysis of Risk Factors of Daily Life Habits in Stroke Based on Optimal Decision Tree. Modern Preventive Medicine, 2018, 45(15): 2689-2693.

[4] Thomas, Lumley, and, et al. A stroke prediction score in the elderly: validation and Web-based application. Journal of Clinical Epidemiology, 2002.

[5] The Stroke RiskometerTM App: Validation of a data collection tool and stroke risk predictor. International Journal of Stroke, 2015, 10(2): 231-244.

[6] Meinshausen M , Rieckert A , Renom-Guiteras A , et al. Effectiveness and patient safety of platelet aggregation inhibitors in the prevention of cardiovascular disease and ischemic stroke in older adults – a systematic review. Bmc Geriatrics, 2017, 17(S1).

[7] Erdur H , Scheitz J F , Ebinger M , et al. In-hospital stroke recurrence and stroke after transient ischemic attack: frequency and risk factors. Stroke; a journal of cerebral circulation, 2015, 46(4):1031-7.

[8] Wang W Y , Sang W W , Yan S M , et al. Cox regression analysis of risk factors for recurrent acute ischemic stroke in 1-year follow-up[J]. Chinese Journal of Geriatric Heart Brain and Vessel Diseases, 2016.

[9] Garcia-Carretero R , Barquero-Perez O , Mora-Jimenez I , et al. Identification of clinically relevant features in hypertensive patients using penalized regression: a case study of cardiovascular events. Medical & Biological Engineering & Computing, 2019, 57(5).

[10] Ji X , Chang W , Zhang Y , et al. Prediction Model of Hypertension Complications Based on GBDT and LightGBM. Journal of Physics: Conference Series, 2021, 1813(1):012008 (8pp).

[11] Chauhan S , Vig L , Grazia M , et al. A Comparison of Shallow and Deep Learning Methods for Predicting Cognitive Performance of Stroke Patients From MRI Lesion Images. Frontiers in Neuroinformatics, 2019, 13.

[12] Almadani O, Alshammari R.Prediction of Stroke Using Data Mining Classification Techniques. International Journal of Advanced Computer Science and Applications, 2018, 9(1): 457-460.

[13] Qian S X , Zhuang J H , Yue W Q , et al. Prediction of post-stroke depression based on classification and regression tree model. Journal of Clinical Neurology, 2018.

[14] Costa H , Fernandes A , Oliveira D , et al. Intergame Analysis of Upper Limb Biomechanics of Stroke Patients in Real and Virtual Environment. 2020.

[15] Lin B S , Lee I J , Hsiao P C , et al. An Assessment System for Post-Stroke Manual Dexterity Using Principal Component Analysis and Logistic Regression. IEEE transactions on neural systems and rehabilitation engineering, 2019, 27(8):1626-1634.

[16] Min S N , Lee K S , Park S J , et al. Development of Stroke Diagnosis Algorithm Through Logistic Regression Analysis with National Health Insurance Database. Springer, Cham, 2017.

[17] Mackay J, Mensah G A, Greenlund K. The Atlas of Heart Disease and Stroke[M]. World Health Organization, 2004.

[18] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and applications.

Neurocomputing, 2006, 70(1/3):489-501.

[19] Hu B . Extreme Learning Machine for Regression and Multiclass Classification. 2012, 42(2): 513-529.

[20] Evgeniou T, ontil M, Poggio T. Regularization Networks and Support Vector Machines. Advances in Computational Mathematics, 2000, 13(1):1-50.

[21] Dewang R K, Singh A K. State-of-art approaches for review spammer detection: a survey. Journal of Intelligent Information Systems, 2018, 50(2):231-264.

[22] He H, Tiwari A, Mehnen J, et al. Incremental information gain analysis of input attribute impact on RBF-kernel SVM spam detection. Evolutionary Computation. IEEE, 2016: 1022-1029.

[23] Varadarajan S, Miller P, Zhou H. Region-based Mixture of Gaussians modelling for foreground detection in dynamic scenes. Pattern Recognition, 2015, 48(11):3488-3503.

[24] Zhao J, X Lei, X Yang, et al. A New Method for Identification of Essential Proteins by Information Entropy of Protein Complex and Subcellular Localization. Springer, Cham, 2019.

Figures

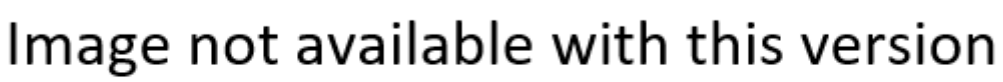

Figure 1

This image is not available with this version.

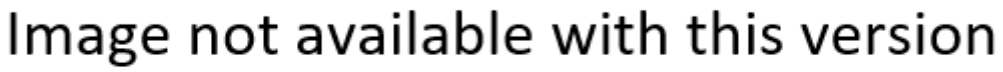

Figure 2

This image is not available with this version.