Classification of Visit-to-Visit Blood Pressure Variability: A Machine Learning Approach for Data Clustering on Systolic Blood Pressure Intervention Trial (SPRINT)

Kelvin KF Tsoi[1,2], Max WY Lam[2], Felix CH Chan[2], Hoyee Hirai[2], Baker KK Bat[2], Samuel YS Wong[1], Helen ML Meng[2,3]
1. Jockey Club School of Public Health and Primary Care, 2. Stanley Ho Big Data Decision Analytics Research Centre,
3. Department of Systems Engineering and Engineering Management. The Chinese University of Hong Kong

ABSTRACT

Background: Blood pressure variability (BPV) is associated with the cardiovascular disease. However, there is no standard risk stratification method to evaluate BPV. Our study aims to cluster BPV into three levels, namely, low, medium and high levels, by a machine learning approach. **Methods:** The Systolic Blood Pressure Intervention Trial (SPRINT) dataset, which includes patients with hypertension or at risk of cardiovascular diseases, was obtained from a clinical data sharing platform. In the clinical trial, participants with systolic blood pressure (SBP) of at least 130 mmHg and an increased cardiovascular risk were randomized to receive intensive treatment (targeting SBP below 120 mmHg) or standard treatment (targeting SBP below 140 mmHg), and blood pressure (BP) were measured and recorded during the follow-up periods. Visit-to-visit BPV was measured by the deviation between the observed records and the personalized BP trends, and two-dimensional clustering on SBP and diastolic BP were applied. Different curve fitting techniques (linear regression and cubic regression) and clustering methods (K-means and Agglomerative Clustering) were attempted and compared with each other. **Results:** With 8,092 participants and a median follow-up of 3.26 years, linear regression was a simple and reliable method to capture the BP trend. K-means model showed stable data clustering results. Intensive treatment showed to be effective for participants with a high level of BPV. **Conclusion:** Machine learning can be used for data clustering on BPV.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

1 BACKGROUND

Hypertension is the major risk factor for the cardiovascular disease [1,2]. Mean blood pressure (BP) is commonly used as the risk indicator for the cardiovascular disease, but BP readings show oscillations over the time [3]. Visit-to-visit blood pressure variability (BPV) is associated with the risk of cardiovascular outcomes [4]. However, there is no standard method for the categorization of BPV. The aim of this study is to apply a machine learning approach to cluster the visit-to-visit BPV.

DH '17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07.
http://dx.doi.org/10.1145/3079452.3079454

2 METHODS

A randomized controlled trial Systolic Blood Pressure Intervention Trial (SPRINT) found that subjects in the intensive treatment group (targeting systolic blood pressure (SBP) lower than 120 mmHg) reduced the risk of a cardiovascular event than subjects in the standard treatment group (targeting SBP lower than 140 mmHg); corresponding result was published in the New England Journal of Medicine in 2015 [5]. The SPRINT dataset was obtained through the National Institute of Health, a clinical data sharing platform in the United States.

It remains a great challenge to determine the cut-off of subject categorization. This problem has been addressed by employing machine learning clustering algorithms. Clustering is the process of grouping individuals in a population according to some similarity measures. It creates grouping such that individuals are similar to one another in the same cluster and are dissimilar to individuals in the other clusters. Variables denoting BPV need to be defined before clustering. There are studies using different sets of definitions [5]. One straightforward measure is to use the standard deviation of visit-to-visit SBP. It was noticed that in the SPRINT dataset subjects in the intensive treatment group were prone to have larger standard deviation than that in the standard treatment group. Also, subjects who initially had high baseline SBP were more sensitive to the treatment, yielding a rapid drop in the first few months. Evidently, it would be unfair if we use the standard deviation of SBP readings for all subjects to address for BPV.

To tackle this problem, curve fitting were employed to remove the apparent trends in the subjects' BP records such that only the fluctuations along the trends would be counted. In practice, both linear regression and cubic regression were attempted. While the fitted curve was obtained, the average absolute value of residual was used as the measure for visit-to-visit BPV. Note that our approach was similar to the work in [6], though they used the standard deviation of the residuals. In practice, by plotting the subjects' BP records, we concluded that the average absolute value of residual was a more sensible choice for our concerned dataset. Besides using the readings of SBP along the treatment, we computed the average absolute value of residual from the diastolic BP with the same procedure as additional information.

K-means is one of the popular machine learning clustering methods. In our work, Lloyd's algorithm [7] was used to perform K-means clustering. Subjects were separated into k clusters represented by k centroid, in a way that individuals within a cluster were closer to their centroid than the centroids of any

other clusters. To increase the robustness, clustering has been re-initialized 100 times using Arthur's approach [7].

Hierarchical clustering is another popular clustering method that produces a structure of multi-scale hierarchical clusters [8]. Agglomerative Clustering [9] is a kind of hierarchical clustering that construct clusters, from small to large, in a bottom-up fashion. Each individual is assigned to a cluster that only contains itself at first. Then, two closest clusters successively merge together according to the distance matrix. The accumulation continues until they form one cluster.

The final model was used to associate with the primary outcomes which were composite of myocardial infarction, acute coronary syndromes, stroke or death from cardiovascular causes. Serious adverse events, such as fatal or life-threatening cases, were considered. Cox-proportional regression models were performed independently for each BPV levels. The benefits from the reduced primary outcomes and the risk of serious adverse events were compared. Subgroup analyses were performed according to age groups, gender, smoking status, and baseline SBP.

3 RESULTS

A total of 8,092 participants with BP records in the first 18 months of follow-up were included in the machine learning models, with a median follow-up of 3.26 years. The mean SBP were 123.7 mmHg in the intensive treatment group and 135.5 mmHg in the standard treatment group. For the machine learning approaches, K-means showed to be more stable than Agglomerative Clustering (Figure 1) and fitting with linear trend showed to be comparable to the cubic curve. Among the participants, 3,596 (44.4%), 3,378 (41.7%), and 1,118 (13.8%) were clustered as with low, medium, and high levels of BPV, respectively.

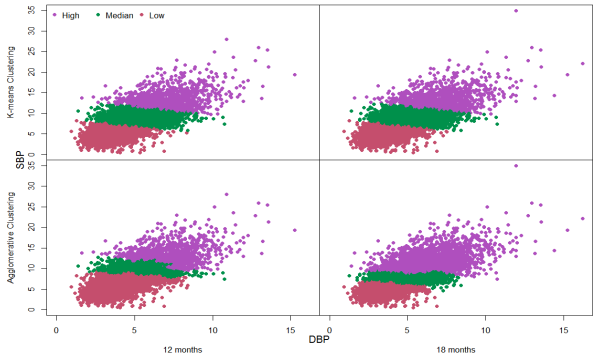

Figure 1: Clustering diagram based on BPV.

Visit-to-visit BPV was associated with the risks of primary outcome after controlling for the intensive treatment (P < 0.001). For participants with a low level of BPV, the intensive treatment group showed a significantly lower rate of primary outcome than the standard treatment group (1.4% per person-year vs 2.3% per person-year; hazard ratio (HR, 95%CI = 0.66, 0.45 to 0.96; P=0.003), but the intensive treatment did not show benefit among those with medium or high levels of BPV (Figure 2). For serious adverse events, participants with low or medium levels of BPV showed comparable risks between the intensive and standard treatments, but participants with high levels of BPV showed significant more adverse events (HR, 95%CI = 1.26 (1.05 to 1.51). Subgroup analyses also demonstrated that the intensive treatment had benefited those who were (i) aged 80 years or above (HR, 95%CI = 0.54, 0.32 to 0.91); (ii) male (HR, 95%CI = 0.69, 0.53 to 0.88); and (iii) non-smokers (HR, 95%CI = 0.67, 0.53 to 0.85).

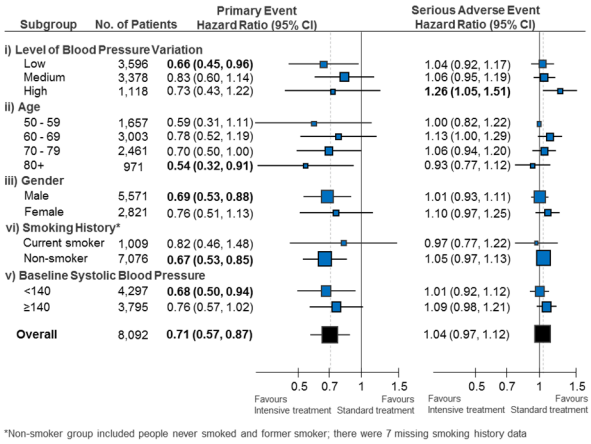

Figure 2: Subgroup Analyses for the Hazard Ratios of Intensive Treatment with reference to Standard Treatment from Cox-proportional Regression Models with adjustment for history of clinical CVD, 10-year CVD risk, and other subgroup variables.

4 CONCLUSIONS

The machine learning approach successfully clustered BPV for risk stratification. The average absolute value of residual with the linear regression fitted curve was used as the measure for visit-to-visit BPV. K-means model showed stable data clustering results. The intensive treatment, as a proactive approach to control an SBP of less than 120 mmHg, showed to be effective for participants who were (i) with a low level of blood pressure variation; (ii) aged 80 years or above; (iii) male; and (iv) non-smokers. Due to the intensive treatment incurred considerable serious adverse events among people with high BP variability, further investigations on populations with different risk levels are needed.

REFERENCES

[1] Lim, S. S., et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. The lancet 2013, 380 (9859). 2224-2260.

[2] Law, M. R., Morris, J. K. and Wald, N. J. Use of BP lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. BMJ 2009, 338. b1665.

[3] Grassi, G., et al. Total cardiovascular risk, BP variability and adrenergic overdrive in hypertension: evidence, mechanisms and clinical implications. Current Hypertension Reports 2012, 14 (4). 333-338.

[4] Stevens, S. L., et al. BP variability and cardiovascular disease: systematic review and meta-analysis. BMJ 2009, 354. i4098.

[5] SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med, 2015 (373). 2103-2116.

[6] Shimbo, D., et al. Association Between Annual Visit-to-Visit BP Variability and Stroke in Postmenopausal Women Novelty and Significance. Hypertension 2012, 60 (3). 625-630.

[7] Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. in 18th annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

[8] Johnson, S. C. Hierarchical clustering schemes. Psychometrika 1967; 32 (3). 241-254.

[9] Gowda, K. C. and Krishna, G. Agglomerative clustering using the concept of mutual nearest neighbourhood. Pattern Recognition 1978; 10 (2). 105-112.