Association of Morphological Features with Hematocrit Levels in Korean Adults: Classification Approach using Machine Learning

Bum Ju Lee, Jong Yeol Kim

Medical Research Division, Korea Institute of Oriental Medicine, Deajeon 305-811, Republic of Korea
jupiter-lee@hanmail.net, bjlee@kiom.re.kr

Abstract: Iron deficiency, which is used to diagnose anemia, is often identified by the measurement of hematocrit levels, and hematocrit levels have been associated with specific morphological features. However, no studies have evaluated the best indicator for identifying hematocrit levels derived from morphological features using machine learning. The objectives of the present study were to identify the best indicator of hematocrit levels among several morphological features and to predict hematocrit levels using a combination of morphological features based on data mining techniques. A total of 1,838 subjects participated in this study. We used two machine learning algorithms, logistic regression (LR) and naive Bayes (NB) algorithms, to identify the best indicator among several morphological features. To overcome the class imbalanced problem and select important features, the synthetic minority over-sampling technique and wrapper-based variable selection were applied to the data set in prediction experiments using combined features. Among all individual features, the best indicator for predicting high and low hematocrit levels was age ($p = <0.0001$; OR = 0.352; AUC = 0.756 by naive Bayes and 0.759 by logistic regression), and among all morphological features, the strongest predictor was body weight ($p = <0.0001$; OR = 1.724; AUC = 0.639 by naive Bayes and 0.641 by logistic regression). For the combined features examined, the area under the receiver operating characteristic curve (AUC) of the four models ranged from 0.745 to 0.789. The method using NB algorithm with wrapper-based variable selection showed the best predictive accuracy (AUC = 0.789; MCC = 0.445) and proved suitable for the prediction of low and high hematocrit levels; this method decreased the model complexity, resulting in the best prediction accuracy and providing the most cost-effective approach. The findings of the present study provide medical knowledge for primary screening and support the use of tools to predict hematocrit levels in both on-site and remote-site healthcare services.
[Bum Ju Lee, Jong Yeol Kim. **Association of Morphological Features with Hematocrit Levels in Korean Adults: Classification Approach using Machine Learning**. *Life Sci J* 2014;11(7):597-601] (ISSN:1097-8135). http://www.lifesciencesite.com. 82

Keywords: Hematocrit level; machine learning; morphological measurement; data mining; prediction

1. Introduction

Iron deficiency is a frequent health problem worldwide (Alleyne et al., 2008; Mwanri et al., 2000) and is diagnosed using hematocrit or hemoglobin levels to identify anemia (Khusun et al., 1999; Scholl, 2005). Anemia is a disease in which the number of red blood cells is not suitable to meet normal physiologic requirements due to iron deficiency (Cusick et al., 2008). The hematocrit denotes the percentage of red blood cells in the blood, and the hematocrit concentration is commonly used to identify iron deficiency because the procedure is cheap, simple, and readily available (Trost et al., 2006).

In recent decades, studies on associations between the hematocrit level and health problems such as obesity have been performed in several countries. For instance, several studies indicated that the hematocrit level serves as an independent risk factor of cardiovascular disease, type 2 diabetes mellitus, and stroke (Tong et al., 2006; Wannamethee et al., 1996; Tamariz et al., 2008; Wannamethee et al., 1994). Other studies reported that the hematocrit level is related to obesity after analyzing body mass index (BMI) or weight compared to the hematocrit concentration, indicating that obese subjects may have greater iron stores than normal subjects (Fricker et al., 1990; Paknahad et al., 2008). However, no studies have reported the best morphological indicator for identifying hematocrit levels derived from morphological features based on machine learning.

The objectives of the present study were to identify the strongest indicator of hematocrit levels among several morphological features and to predict hematocrit levels using a combination of morphological features based on data mining techniques. In our previous study (Lee and Kim, 2014), we described a prediction method for hematocrit levels using combined morphological features, although the best indicator of hematocrit levels was not revealed in this study; the current study represents an extended version of this previous study (Lee and Kim, 2014). The findings of this study will provide valuable information for the development of initial screening tools to assess blood hematocrit levels.

Table 1. Baseline characteristics assessed in the present study. The data are expressed as the mean (standard deviation).

	High level	Low level	Description
Subject	1712	126	Number of subjects
Hematocrit (%)	44.7 (2.765)	36.32 (2.715)	Hematocrit level
AGE	45.3 (13.96)	58.9 (13.82)	Age
HEIGHT	170.7 (6.08)	167.5 (6.645)	Height
WEIGHT	70.94 (10.26)	65.98 (11.33)	Weight
BMI	24.31 (3.044)	23.48 (3.493)	Body mass index
BODY1	57.74 (1.962)	57.06 (1.971)	Forehead circumference (the levels of the glabella and occiput)
BODY2	38.51 (2.554)	38.21 (3.074)	Neck circumference (the levels of the thyroid cartilage and cricoid cartilage)
BODY3	97.21 (6.378)	95.13 (7.009)	Axillary circumference (the levels of the left and right axilla)
BODY4	94.75 (6.975)	93.53 (7.349)	Chest circumference (the levels of the left and right nipples)
BODY5	88.06 (7.447)	88.14 (7.746)	Rib circumference (the levels of the left and right 7^{th} and 8^{th} prominences of the costochondral junction)
BODY6	87.62 (8.222)	87.15 (10.08)	Waist circumference (the level of the umbilicus)
BODY8	95.47 (6.241)	93.81 (7.023)	Hip circumference (the level of the upper edge of the pubis)

2. Material and Methods

All the data in this study were obtained from the Korean Health and Genome Epidemiology study database (KHGES). Written informed consent was obtained from all participants, and the Korea Institute of Oriental Medicine (KIOM) Institutional Review Board approved the study.

A total of 1,838 men aged 20–80 years participated in this study. For the diagnosis of low and high hematocrit values, we considered the recommendations of the World Health Organization (WHO). Therefore, for adult men, low hematocrit status was defined as less than 39%, and high hematocrit status was defined as more than 39%.

For the morphological measurements of body shape, body weight and height were estimated to the nearest 0.1 kg and 0.1 cm, respectively, using appropriate equipment (LG-150; G Tech International Co., Ltd., Uijeongbu). Body shapes were measured to the nearest 0.5 cm using non-elastic tape on the forehead, neck, axilla, chest, rib, waist, and hip positions, according to the methods reported in previous studies (Jang et al., 2012, Lee et al., 2013; Lee and Kim, 2014). The baseline characteristics evaluated in this study with brief descriptions are presented in Table 1.

All statistical analyses and prediction experiments were carried out using SPSS 19 for Windows (SPSS Inc., Chicago, IL, USA) and the Waikato Environment for Knowledge Analysis data mining tool (WEKA) (Hall et al., 2009). To perform the statistical analyses of individual indices, logistic regression (LR) was used to evaluate statistically significant differences between high and low hematocrit levels after a standardization transformation was applied to the data. To predict hematocrit levels using combined morphological features, we used two machine learning algorithms, logistic regression (LR) and the naive Bayes (NB) algorithm, to obtain better prediction performance and more reliable results. For variable subset selection, we used the wrapper-based variable subset selection technique with a greedy stepwise backward search. To overcome the class imbalanced problem, the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) was applied to our data set. In our experiments, we used four prediction methods using two classification algorithms and two data sets. The four methods are as follows: naive Bayes with full variable set (NB-full), logistic regression with full variable set (LR-full), naive Bayes with wrapper technique (NB-wrapper), and logistic regression with wrapper technique (LR-wrapper).

For the primary criterion of prediction performance of high and low hematocrit levels, we used the area under the receiver operating characteristic curve (AUC). In addition, we obtained the Matthews correlation coefficient (MCC) for performance in each method and determined the sensitivity, 1-specificity, precision, and F-measure to support the specific results of the prediction models. All prediction experiments were performed using 10-fold cross validation.

3. Results

3.1 Analysis of individual morphological features

Table 2. Statistical analysis and predictive power of individual morphological features (OR: odd ratio, AUC: the area under the receiver operating characteristic curve, NB: naive Bayes, LR: logistic regression).

Variable	p	OR	AUC by NB	AUC by LR
AGE	<0.0001	0.352 (0.284 - 0.437)	0.756	0.759
HEIGHT	<0.0001	1.655 (1.379 - 1.985)	0.633	0.635
WEIGHT	<0.0001	1.724 (1.402 - 2.121)	0.639	0.641
BMI	0.0034	1.33 (1.099 - 1.609)	0.571	0.572
BODY1	0.0003	1.368 (1.154 - 1.62)	0.601	0.601
BODY2	0.2227	1.121 (0.933 - 1.347)	0.525	0.526
BODY3	0.0005	1.39 (1.155 - 1.674)	0.584	0.595
BODY4	0.0587	1.196 (0.993 - 1.44)	0.536	0.554
BODY5	0.9071	0.989 (0.826 - 1.185)	0.44	0.426
BODY6	0.5468	1.057 (0.882 - 1.268)	0.532	0.486
BODY8	0.0043	1.31 (1.088 - 1.576)	0.582	0.58

Table 2 lists the statistical analysis and predictive power of individual morphological factors for the comparison of high and low hematocrit levels. AGE was the best indicator for predicting high and low hematocrit levels among all features evaluated (p = <0.0001; OR = 0.352; AUC = 0.756 by NB and 0.759 by LR). Among the morphological features examined, the strongest predictor was WEIGHT (p = <0.0001; OR = 1.724; AUC = 0.639 by NB and 0.641 by LR). Additionally, HEIGHT (p = <0.0001; OR = 1.655; AUC = 0.633 by NB and 0.635 by LR), BODY1 (p = 0.0003; OR = 1.368; AUC = 0.601 by NB and 0.601 by LR), and BODY3 (p = 0.0005; OR = 1.39; AUC = 0.584 by NB and 0.595 by LR) showed useful predictive powers for identifying hematocrit levels.

3.2 Prediction of combined morphological features

Figure 1 presents the results of the prediction experiments using the four methods. Among the four prediction methods, the NB-wrapper method showed the best prediction accuracy (AUC = 0.789; MCC = 0.445). When comparing the models using the full variable set and the models using variable subsets with the wrapper technique, the prediction accuracies of the variable subsets were slightly better than those of the full variable sets for both NB and LR. The NB-wrapper method obtained a better accuracy compared to the NB-full method (accuracy was improved by 0.044 in the AUC value), while the LR-wrapper method showed very little improvement (0.002). The best prediction method in our experiments slightly differed according to the criteria of prediction performance, although we considered the AUC as the primary criterion.

Table 2 lists the specific results for the prediction experiments using the four methods, including the sensitivity, 1-specificity, precision, and F-measure results. For example, the model constructed using the NB-full method demonstrated the lowest prediction performance (0.745 for the AUC); the sensitivity results for high and low hematocrit levels were 0.659 and 0.714, respectively, and the corresponding F-measure results were 0.678 and 0.695, respectively. The prediction model constructed using the NB-wrapper method included AGE, BMI, and BODY6, and the model built using the LR-wrapper method included AGE, HEIGHT, WEIGHT, BODY4, BODY5, and BODY8.

In our prediction experiments, the model using the NB-wrapper method proved suitable for the prediction of low and high hematocrit levels because this method decreased the model complexity, resulting in the best prediction accuracy as well as the most cost-effective model.

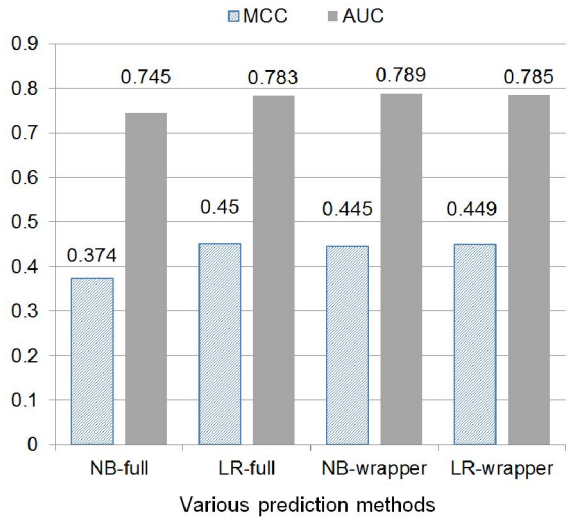

Figure 1. Comparison of prediction performance among various methods using the area under the receiver operating characteristic curve (AUC) and Matthews correlation coefficient (MCC).

Table 3. Detailed prediction results using different methods (Sens.: sensitivity, 1-spec.: 1-specificity, Prec.: precision, F-mea.: F-measure).

Method	Class	Sens.	1-spec.	Prec.	F-mea.
NB-full	High	0.659	0.286	0.698	0.678
	Low	0.714	0.341	0.677	0.695
LR-full	High	0.706	0.256	0.734	0.719
	Low	0.744	0.294	0.717	0.73
NB-wrapper	High	0.685	0.241	0.74	0.711
	Low	0.759	0.315	0.707	0.732
LR-wrapper	High	0.708	0.259	0.732	0.72
	Low	0.741	0.292	0.717	0.729

4. Discussions

In previous studies on the association of hematocrit level with disease, several researchers evaluated the relationship between hematocrit levels and type 2 diabetes, blood pressure, and mortality. For example, Tong and colleagues (Tong et al., 2006) reported that in Chinese individuals with type 2 diabetes, a low level of hematocrit served as an independent risk factor of cardiovascular disease. Additionally, Wannamethee and colleagues (Wannamethee et al., 1996) suggested that hematocrit levels were related to noninsulin-dependent diabetes mellitus in the British Regional Heart Study, and Yanagawa and colleagues (Yanagawa et al., 1997) reported that hematocrit levels were associated with blood pressure in both Japanese men and women. Kosiborod and colleagues further demonstrated that each 1% decrease in hematocrit level was related to a 2% increase in mortality after 1 year (Foley, 2003; Kosiborod et al., 2003). Additionally, Balducci (Balducci, 2003) proposed that anemia is associated with age because the prevalence and incidence of anemia increase with age, and our results are consistent with findings of Balducci. In analyses of the predictive powers of individual features, we found that the age feature was the best indicator of hematocrit level; therefore, we believe that hematocrit levels are strongly associated with age.

Several studies have investigated the relationship between blood hematocrit levels and body mass index or weight (Fricker et al., 1990; Paknahad et al., 2008; Famodu and Awodu, 2009). For instance, a study by Paknahad et al. found that the hematocrit level was significantly different between BMI quartiles and between weight quartiles ($p = 0.01$) (Paknahad et al., 2008). In another study (Famodu and Awodu, 2009), the authors reported that the hematocrit level increased in parallel with BMI in Nigerian women and suggested that BODY6 served as a better indicator of the risk of hemorheological cardiovascular disease than BMI. However, previous studies have only focused on the association of hematocrit levels with measures such as BMI, waist circumference, waist-to-hip ratio, and weight. In contrast, our study focused on the predictive powers of hematocrit levels using a variety of morphological features, and our results are the first to report the prediction of hematocrit levels using combined morphological measurements.

5. Conclusion

In the present study, we demonstrated the potential to predict high and low blood hematocrit levels according to body shape. Our results indicated that naive Bayes with wrapper-based variable subset selection demonstrated the best prediction accuracy. To our knowledge, this is the first report to predict hematocrit levels using basic body shape measurements and data mining techniques. The findings of this study may provide clinical knowledge for primary health screening and support helpful tools for prediction of high and low hematocrit levels in both on-site and remote-site healthcare services. However, the present study was limited by regional and ethnic differences as well as gender because we only used data from Korean men.

Acknowledgements:

This research was supported by the National Research Foundation of Korea (NRF) and received funding from the Ministry of Science, ICT & Future Planning (No.2006-2005173) (NRF-2012-0009830 and NRF-2009-0090900).

Corresponding Author:
Dr. Jong Yeol Kim
Medical Research Division, Korea Institute of Oriental Medicine, Deajeon 305-811, Republic of Korea
E-mail: bjlee@kiom.re.kr

References

1. Alleyne M, Horne MK, Miller JL. Individualized Treatment for Iron-deficiency Anemia in Adults. Am J Med 2008;121(11):943–948.
2. Balducci L. Epidemiology of anemia in the elderly: information on diagnostic evaluation. J Am Geriatr Soc 2003;51(3):S2–S9.
3. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321–357.
4. Cusick SE1, Mei Z, Freedman DS, Looker AC, Ogden CL, Gunter E, Cogswell ME. Unexplained decline in the prevalence of anemia among US children and women between 1988-1994 and 1999-2002. Am J Clin Nutr 2008;88(6):1611–1617.
5. Famodu AA, Awodu OA. Anthropometric indices as determinants of haemorheological cardiovascular disease risk factors in Nigerian

adults living in a semi-urban community. Clin Hemorheal Microcirc 2009;43(4):335–344.
6. Foley R: Anaemia and the heart: what's new in 2003? Nephrol Dial Transplant 2003;18(8):viii13–viii16.
7. Fricker J, Le Moel G, Apfelbaum M. Obesity and iron status in menstruating women. Am J Clin Nutr 1990;52(5):863–866.
8. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explor 2009;11(1):10–18.
9. Jang E, Kim JY, Lee H, Kim H, Baek Y, Lee S. A Study on the Reliability of Sasang Constitutional Body Trunk Measurement. Evid Based Complement Alternat Med 2012;604842.
10. Khusun H, Yip R, Schultink W, Dillon DH. World Health Organization hemoglobin cut-off points for the detection of anemia are valid for an Indonesian population. J Nutr 1999;129(9):1669–1674.
11. Kosiborod M, Smith GL, Radford MJ, Foody JM, Krumholz HM. The prognostic importance of anemia in patients with heart failure. Am J Med 2003;114(2):112–119.
12. Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE Journal of Biomedical and Health Informatics doi: 10.1109/JBHI.2013.2264509, 2013, In press.
13. Lee BJ, Kim JY. A Comparison of the Predictive Power of Anthropometric Indices for Hypertension and Hypotension Risk. PLoS ONE 2014;9(1):e84897.
14. Lee BJ, Kim JY. Prediction of hematocrit levels from body shape in adult men using data mining. Proceedings of 2014 FTRA International Conference on Advanced Computing and Services (ACS-14) 2014;42–44.
15. Mwanri L, Worsley A, Ryan P, Masika J. Supplemental Vitamin A Improves Anemia and Growth in Anemic School Children in Tanzania. J Nutr 2000;130(11):2691–2696.
16. Paknahad Z, Mahboob S, Omidvar N, Ebrahimi M, Ostadrahimi A, Afiatmilani S. Body mass index and its relationship with hematological indices in Iranian women. Pakistan Journal of Nutrition 2008;7(2):377–380.
17. Scholl TO. Iron status during pregnancy: setting the stage for mother and infant. Am J Clin Nutr 2005;81(5):1218S–1222S.
18. Tamariz LJ, Young JH, Pankow JS, Yeh HC, Schmidt MI, Astor B, Brancati FL. Blood viscosity and hematocrit as risk factors for type 2 diabetes mellitus: the atherosclerosis risk in communities (ARIC) study. Am J Epidemiol 2008;168(10):1153–1160.
19. Tong PC, Kong AP, So WY, Ng MH, Yang X, Ng MC, Ma RC, Ho CS, Lam CW, Chow CC, Cockram CS, Chan JC. Hematocrit, independent of chronic kidney disease, predicts adverse cardiovascular outcomes in chinese patients with type 2 diabetes. Diabetes Care 2006;29(11):2439–2444.
20. Trost LB, Bergfeld WF, Calogeras E. The diagnosis and treatment of iron deficiency and its potential relationship to hair loss. J Am Acad Dermatol 2006;54(5):824–844.
21. Wannamethee SG, Perry IJ, Shaper AG. Hematocrit and risk of NIDDM. Diabetes 1996;45(5):576–579.
22. Wannamethee G, Perry IJ, Shaper AG. Haematocrit, hypertension and risk of stroke. J Intern Med 1994;235(2):163–168.
23. WHO, Report of the UNICEF/WHO Regional Consultation, Prevention and Control of Iron Deficiency Anaemia in Women and Children Geneva, Switzerland, 1999
24. Yanagawa T, Yoshida Y, Wada N, Nakao E, Ogiwara H, Uyama I, Takahara T, Nomura T, Gomi K, Saruta T. Blood pressure, insulin, and haematocrit values in Japanese subjects over 60 years of age. J Hum Hypertens 1997;11(6):355–359.

5/26/2014