

A New Method to Predict Erythrocyte Sedimentation Rate with Natural Geographical Factors and Location by Case-based Reasoning: A Case Study of China

YANG Qingsheng¹, YOU Xibin², ZHANG Hongxian³, Kevin MWENDA⁴, WANG Yuandong², HUANG Ying²

(1. School of Geography and Tourism, Guangdong University of Finance & Economics, Guangzhou 510320, China; 2. Resource and Environment Center, Shaoguan University, Shaoguan 512005, China; 3. School of Management, Guangdong Polytechnic Normal University, Guangzhou 510665, China; 4. Population Studies and Training Center, Brown University, RI 02912, USA)

Abstract: Reference values of erythrocyte sedimentation rate (ESR) are the key to interpret ESR blood test in clinic. The common local reference ESR values are more accuracy in blood test that are established with natural geographical factors by using the multiple linear regression (MLR) model and the artificial neural network (ANN). These knowledge-based methods have limitations since the knowledge domains of ESR and natural geographical factors are limited. This paper presents a new cases-depended model to establish reference ESR values with natural geographical factors and location using case-based reasoning (CBR) since knowledge domain of ESR and geographical factors is weak. Overall 224 local normal ESR values of China that calculated from 13 623 samples were obtained, and the corresponding natural geographical factors and location that include altitude, sunshine hours, relative humidity, temperature, precipitation, annual temperature range and annual average wind speed were obtained from the National Geomatics Center of China. CBR was used to predict the unseen local reference ESR values with cases. The average absolute deviation (AAD), mean square error (MSE), prediction accuracy (PA), and Pearson correlation coefficient (r) between the observed and estimated data of proposed model is 33.07%, 9.02, 66.93% and 0.78, which are better than those of ANN and MLR model. The results show that the proposed model provides higher prediction accuracy than those of the artificial neural network and multiple linear regression models. The predicted values are very close to the observed values. Model results show significant agreement of cases data. Consequently, the model is used to predict the unseen local reference ESR with natural geographical factors and location. In spatial, the highest ESR reference areas are distributed in the southern-western district of China that includes Sichuan, Chongqing, Guangxi and Guizhou provinces, and the reference ESR values are greater than 23 mm/60 min. The higher ESR reference values are distributed in the middle part and northern-eastern of China which include Hubei, Henan, Shaanxi, Shanxi, Jilin and Heilongjiang provinces, and the reference ESR values are greater than 18 mm/60 min. The lowest ESR reference values are distributed in the northern-western of China that includes Tibet and Xinjiang, and the reference ESR values are lower than 5 mm/60 min.

Keywords: erythrocyte sedimentation rate (ESR); natural geographical factors; case-based reasoning (CBR); China

Citation: YANG Qingsheng, YOU Xibin, ZHANG Hongxian, MWENDA Kevin, WANG Yuandong, HUANG Ying, 2020. A New Method to Predict Erythrocyte Sedimentation Rate with Natural Geographical Factors and Location by Case-based Reasoning: A Case Study of China. *Chinese Geographical Science*, 30(1): 157–169. https://doi.org/10.1007/s11769-020-1102-7

1 Introduction

The erythrocyte sedimentation rate (ESR) test is a well-established clinical test for diseased patient that is

commonly used to estimate the body's acute-phase reaction to inflammation and infection (Zacharski and Kyle, 1967; Wyler, 1977). Reference ESR values are usually used to diagnose specific disease severity and assess the

Received date: 2018-12-06; accepted date: 2019-02-22

Foundation item: Under the auspices of National Natural Science Foundation of China (No. 40971060)

Corresponding author: YOU Xibin. E-mail: Youxibin@163.com

© Science Press, Northeast Institute of Geography and Agroecology, CAS and Springer-Verlag GmbH Germany, part of Springer Nature 2020

general sickness index by physicians. In the early 19th century, the Greeks observed the relationship between the sedimentation of red blood cells and fibrinogen, and the concept of ESR appeared (Ropes et al., 1939). In 1918, Fahraeus discovered that erythrocyte sedimentation in plasma occurred more rapidly in pregnant women than in non-pregnant women (Fahraeus, 1929). At the same time, physicians found that the ESR values varied with some diseases. Since then, the ESR test has been used in the evaluation of various diseases because of its disease indicative effects and convenience (da C Sousa et al., 2018; Schäfer et al., 2018; van Atteveld et al., 2019).

The Wintrobe method is the most commonly used standard ESR test method (Olshaker and Jerrard, 1997). Many physicians have measured normal ESR values to compare the difference in the ESR between patients and healthy individuals at local hospitals and research institutes. When researchers collected many local normal ESR values, they found that normal ESR values varied with the features of patients, such as age, gender, weight and other personal habits, such as smoking. For example, Ansell and Bywaters found that the ESR value increased as one became older (Ansell and Bywaters, 1958). Pincherle and Shanks found that the tendency of the ESR values to increase with age flattens out after the age of 60 (Pincherle and Shanks, 1967). They also found that ESR has a significant variation with the weight and gender of patients (Pincherle and Shanks, 1967). Pincherle and Shanks found that the ESR values will significantly rise with increasing degree of obesity (Pincherle and Shanks, 1967). At the same time, physicians also found that smoking tends to cause a rise in the mean ESR values (Pincherle and Shanks, 1967). In addition to the personal features of patients, physicians found that normal ESR values vary with seasonal changes. For example, Pincherle and Shanks observed a general pattern of high ESR values in spring and autumn and low values in summer (Pincherle and Shanks, 1967). Because the ESR values vary with the age, weight and gender of a person, physicians have tried to calculate normal ESR values using age and gender. For example, Miller et al. proposed a formula to calculate the maximum reference ESR at any given age. In their study, the ESR value is calculated as $(\text{age in years} / 2)$ for men and $(\text{age in years} + 10) / 2$ for women (Miller et al., 1983). Consequently, the reference ESR value is expected to be the same for the specific age and

gender, regardless of location. The formula was an old way and it was seldom used in clinic now. Meanwhile, some studies have found that normal ESR values also vary with geographical factors such as altitude, in addition to age, gender and smoking status (Cui et al., 2001; Kuznetsova et al., 2013; Hoiland et al., 2017; Zouboules et al., 2018; Hale et al., 2019). As such, reference ESR values can be calculated locally because geographical factors are known to be associated with location. For example, Ge et al. found that normal ESR values will rise with decreasing altitude—in fact, some formulas were proposed for calculating reference ESR values at a given altitude using multiple linear regression (MLR) with specific age and gender as covariates (Ge et al., 2001; Ge, 2004). Because geographical factors such as temperature and humidity are known to be associated with altitude, the aforementioned observations beg the question: are normal ESR values influenced by natural geographical factors (Kuznetsova et al., 2013)? For example, at Pincherle and Shanks' study, it is not clear if changes in normal ESR values vary with temperature. To determine how such geographical factors affect normal ESR values, Ge et al. studied the relationship among normal ESR values, altitude and humidity using a stepwise regression statistical model (Ge et al., 1999). While the aforementioned study showed that geographical factors improve the prediction accuracy of local reference ESR values, the underlying reasons are not definitive. Obviously, natural geographical factors such as altitude, humidity and temperature, are significantly correlated because humidity and temperature generally decrease as altitude increases. The stepwise regression statistical model has limitations in that the independent variables are correlated (Li and Yeh, 2002). In solving variable correlations, we proposed a method to predict the local ESR value using artificial neural network (ANN) (Yang et al., 2013). After the study was published, many local hospitals requested that we provide the local reference ESR value according to age and gender calculated in the ANN model. Thus, it is necessary and important to locally establish reference ESR values. The methods, regression model and artificial neural network, used to calculate ESR values, are statistical method and machine learning approaches that require the exact causality of the variables. The mechanism underlying how independent variables influence the dependent variable should be clear in the MLR and ANN models. However, physicians do not understand completely the

mechanism regarding how natural geographical factors affect ESR. Researchers have found that the variations in ESR are due to variations in the colloidal state of the plasma with consequent changes in the electric charges on the proteins and red cells. Variations in the concentration of fibrinogen, globulin and other constituents affect ESR through their effect on the colloidal state of the plasma (Greisheimer et al., 1929; Moen and Reimann, 1933; Ropes et al., 1939; Suckling, 1957; Ge et al., 1999). Geographical factors such as altitude and temperature affect red blood cells and plasma proteins because of hypoxia that produces variations of the colloidal state of plasma that then affect the variations of ESR (Kuznetsova et al., 2013). Thus, natural geographical factors affect the ESR value indirectly. As such, the relationship between normal ESR values and geographical factors is indirect and indefinite, and thus complicated. The knowledge domain among ESR values and natural geographical factors is weak. The MLR and ANN models have limitations for predicting ESR values with natural geographical factors because the knowledge domain is weak.

Herein, we present a new method to predict local reference ESR values of China using case-based reasoning (CBR). CBR was inspired by observing human reasoning when learning how to solve new problems by remembering solutions that were applied to similar problems in the past (Kolodner, 1993). In the same way, a CBR system usually find an answer by comparing the problem with old problems and their solutions, which are known as cases that are stored in memory as case library (Holt, 1999). The basic idea of CBR is to solve a new problem by identifying and reusing previous similarity cases based on the heuristic principal that similar problems have a high likelihood of having similar solutions (Kolodner, 1992). Thus, by using CBR, the difficulties of knowledge acquisition and of knowledge representation are often lessened (Montani and Jain, 2010). One of the most important characteristics of CBR is that it does not require explicit domain knowledge but gathering cases. It is not necessary to query experts regarding their method of reasoning (Gierl et al., 2003). Many applications have demonstrated the capability of CBR to solve issues that would be too difficult to manage with other classical artificial intelligence (AI) methods especially in the health sciences, such as rules or models (Montani and Jain, 2010). CBR works well in such domain knowledge is rudimentary, i.e., a weak domain

theory. Because the mechanism underlying how natural geographical factors affect ESR is not ascertained completely and the domain theory is not adequate, CBR is a better choice to predict ESR values from cases with natural geographical factors. For providing scientific basis of establishing regional ESR reference values, this paper introduces the theory of predicting local ESR with CBR and geographical factors systematically, and the local ESR values of China are predicted. The current study first develops a CBR model to predict local reference ESR values with natural geographical factors. Then, the spatial distribution of reference ESR values is analyzed.

2 Materials and Methods

2.1 Normal ESRs and natural geographical factors data

In view of the progressive rise in ESR with age, separate ESR values should be established for each decade of life in males and females (Sharland, 1980; Näyhä, 1987; Jou et al., 2011). The normal ESRs were usually divided into adults and children. For adults, the normal ESR was divided into three groups according to age: younger than 50, 51 to 60 and older than 60. For children, the normal ESR was divided into two groups, newborn and neonatal to puberty (Wetteland et al., 1996). As such, ESR values showed no significance difference between 5 and 21 (Siemons et al., 2014). To determine the effect of natural geographical factors on ESR with sample data, the effect of age and gender on ESR should be minimized. Normal ESR of a specific age and gender should be selected for sample data. Considering data sources and data amount, normal ESR values of men younger than 18 and older than 5 are selected as case data in the study. We obtained 224 normal ESR values of China. The local ESR values of Taiwan, Hong Kong and Macau are not predicted because of lacking data. The 224 sample ESR values of young men from 5 to 18 years old are obtained from Ge's research in which the 224 ESR values were calculated from 13 623 samples (Ge et al., 2001), and the detailed origins of 13 623 samples can be found in the research (Ge et al., 2001). Every normal ESR value was calculated with more than 20 samples and was expressed as the mean ESR \pm standard deviation (SD). The blood samples were local citizens who had undergone normal yearly physical examina-

nations and unhealthy persons were excluded (Ge et al., 2001). All values were collected using the Wintrobe method (Ge et al., 2001). These data were geocoded with the city location of the hospitals to match natural geographical factors. The 224 normal ESR values were matched with 182 cities, and some cities had more than one normal ESR values. It's very hard to correct the difference among ESR values if they locate in the same city because of unknown original samples. It's also hard to choose one ESR value for the city from some ESR values. We hope to keep all ESR values for city if they are not outlier values and the reference ESR value of the city would be calculated with case-based k-NN reasoning way. Outlier statistic way is used to assess the quality of case data for the whole data sets and for ESR values that fall in the same city with SPSS. In 224 cases data, there are not outliers. The same results are found on the cases which locate in the same city. Based on the quality assessment, overall 224 normal ESR values of young men from 5 to 18 years old in 182 cities were collected as case solution data in this research.

Some studies showed that local normal ESR values can be predicted with geographical factors using a multiple linear regression (MLR) model (Ge et al., 1999; 2001). We have observed that local reference ESR values can be predicted with geographical factors using the more complicated artificial neural network (ANN) model (Yang et al., 2013). Referring to past studies, seven natural geographical factors were chosen as case attributes to predict the local reference ESR values (Ge et al., 1999; Yang et al., 2013). The longitude and latitude are the location indicators of the case. There are nine factors for the case attributes as follows: 1) Altitude f_1 (m); 2) Annual sunshine hours f_2 (h); 3) Annual average relative humidity f_3 (%); 4) Annual average temperature f_4 (°C); 5) Annual average precipitation f_5 (mm); 6) Annual temperature range f_6 (°C); 7) Annual average wind speed f_7 (m/s); 8) Longitude f_8 (°); 9) Latitude f_9 (°). The geographical data are obtained from the National Geomatics Center of China.

2.2 Case-based Reasoning

Case-based reasoning (CBR) is a problem solving paradigm that in many respects, is fundamentally different from other major AI approaches. Instead of relying solely on the general knowledge of a problem domain, or making associations along generalized relationships

between problem descriptors and conclusions, CBR is able to utilize the specific knowledge of previously experienced, concrete problem situations (cases). A new problem is solved by finding similar past cases, and reusing it in the new problem situation. Another important difference is that CBR is also an approach to incremental, sustained learning, since a new experience is retained each time a problem has been solved, making it immediately available for future problems (Aamodt and Plaza, 1994).

In CBR, a case usually denotes a problem situation. A previously experienced situation, which has been captured and learned in such way that it can be reused in the solving of future problems, is referred to as a past case, previous case, stored case, or retained case. Correspondingly, a new case or unsolved case is the description of a new problem to be solved. Case-based reasoning is a cyclic and integrated process of solving a problem, learning from this experience, and solving a new problem. A general CBR cycle may be described by the following four processes (Aamodt and Plaza, 1994): 1) RETRIEVE, the most similar case or cases; 2) REUSE, the information and knowledge in that case to solve the problem; 3) REVISE, the proposed solution; 4) RETAIN, the parts of this experience likely to be useful for future problem solving.

A new problem is solved by retrieving one or more previously experienced cases, reusing the case in one way or another, revising the solution based on reusing a previous case, and retaining the new experience by incorporating it into the existing knowledge-base (case-base) (Aamodt and Plaza, 1994).

2.3 CBR-based reference ESR value predicting model

A statistical model or machine learning model such as the regression model and artificial neural network, may not be the best way to reveal relationships between ESR values and natural geographical factors because of the weak domain theory. Instead, CBR can be designed to estimate unseen local reference ESR values at populated areas using known cases.

2.3.1 Establishment of the case library

The first step of this proposed method is to establish the case library. In this proposed model, each case is represented by two parts: 1) the attributes (features) of each location and 2) the reference ESR value of the location

(solution). The first part includes geographical factor variables. The second part is the solution, which is the reference ESR value of the location. Specifically, a case can then be represented as follows:

Case(i): $f_1(i), f_2(i), f_3(i), f_4(i), f_5(i), f_6(i), f_7(i), S(i)$ (1)

where the variables of $f_1(i), f_2(i), f_3(i), f_4(i), f_5(i), f_6(i), f_7(i)$ are the attributes (features) of case i , and variable of $S(i)$ is the solution of case i . The variables of $f_1(i), f_2(i), f_3(i), f_4(i), f_5(i), f_6(i), f_7(i)$ represent the altitude, annual average sunshine hours, annual average relative humidity, annual average temperature, annual average precipitation, annual temperature range and annual average wind speed of case i , respectively. The variable of $S(i)$ represents the reference ESR value of case i .

2.3.2 Retrieving cases and predicting reference ESR value

The second step of the proposed model is to retrieve cases from the case library that match the attributes of new case when finding solution to a new case. The case library stores the experience that describes the natural geographical factors and corresponding normal ESR value. Case matching is a frequently used way to retrieve the cases in the CBR model. The matching is usually based on the similarity between an input (questioned) case i and a known case j in the case library. The similarity can be calculated using the following Euclidean feature distance function:

$$d_{ij} = \sqrt{\sum_{k=1}^m [f_k(i) - f_k(j)]^2} \quad (2)$$

where $f_k(i)$ is the k th feature of the input case i , j is the known case, and m is the number of features. Two cases will be more similar if they have a closer Euclidean distance in the feature space. The similarity is calculated according to several features, which can be treated equally for their importance. In reality, the importance of each feature may be different. It's better to assign a weight to address the contribution of each feature in calculating the similarity. The above equation is then revised as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^m w_k^2 [f_k(i) - f_k(j)]^2} \quad (3)$$

where the variable w_k is the weight for the k th feature. There are many ways to calculate the weights. Entropy

is a frequently used method to calculate the weights when prior knowledge is not sufficient. Entropy is used to represent the content of information and it can be calculated as follows (Theil, 1967):

$$H_k = - \frac{\sum_{i=1}^n p_{ki} \log p_{ki}}{\log(n)} \quad (4)$$

where H_k is the Shannon entropy of factor k , and $p_{ki} = f_k(i) / \sum_{i=1}^n f_k(i)$, and k, n are the total number of factors and observations respectively. Variable $f_k(i)$ need to be standardized into the range of $[0, 1]$.

The value of entropy falls within the range of $[0, 1]$. The smallest value of 0 represents the maximum amount of the information exhibited in the variable. The largest value of 1 indicates the minimum amount of the information. Therefore, the amount of the information is directly proportional to this form: $1-H_k$.

A feature that has a larger amount of information is expected to have a larger weight. Then, the entropy weight for the k th feature can be represented as follows:

$$w_k = \frac{1 - H_k}{m - \sum_{k=1}^m H_k} \quad (5)$$

where m is the number of features. The reasoning process usually locates the new case i to its nearest known case j from the case library. The known case j has a corresponding solution $S(j)$ (reference ESR value). The reasoning assumes that the case closest to j tends to have a target value close to $S(j)$.

Actually, the matching is often carried out by comparing the new case with a number of known cases, its k -nearest neighbours. This is the most popular k -nearest-neighbour (k-NN) algorithm, which works well on many practical problems and is fairly noise-tolerant in CBR applications (Dasarathy, 1991). Intuitively, the k-NN algorithm assigns to each new case the majority value among its k nearest neighbours. It is possible that the use of different k values may have different enquiry results. For example, the predicted value of new case will be different by using five nearest neighbors or by using 10 nearest neighbors. The appropriate value of k is considered to be problem-dependent (Houben et al., 1995). Experiments can be carried out to determine the appropriate

value for an application. It is reasonable that a feature closer neighbour should have more influences than others in predicting. A feature distance-weighted function can be used to treat these neighbours differently. For a new case i , there were some ways to find the adaption solution such as distance weighted averaging of the values of the k cases closest to the input problem (Kibler et al., 1989; Aha et al., 1991), enhancing case-based regression (Jalali and Leake, 2016). In our study, the solution $S(i)$ was estimated by using distance weighted averaging values of neighbor cases as following:

$$S(i) = \sum_{j=1}^k w_{fj} S(j) \quad (6)$$

where $S(i)$ is the solution of the new case i , and w_{fj} and $S(j)$ are the normalized weight and corresponding solution of similar case j , respectively. The original w_{fj} is the reciprocal of the Euclidean feature distance d_{ij} between the new case i and retrieved similar case j .

The above k-NN algorithm only addresses the natural geographical factors influences in the feature space. The distance-weighted function, which is calculated in the feature space, cannot reflect the spatial variations. It is obvious that the influence of a case may not be the same at different spatial locations in the reasoning process. The location information should be included as a part of the attribute for a case. This can be done by adding the coordinates of a case as additional attributes in Equation (6).

Therefore, the distance-weighted function should have two parts: 1) feature-distance weighted; and 2) spatial-distance-weighted. The later element is essential in dealing with spatial variations. The spatial-distance weights (w_{sj}) can be defined to count the spatial influences related to the case locations in original space. w_{sj} is written as follows:

$$w_{sj} = \frac{1}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}} \quad (7)$$

where x and y are the coordinates of a case. Finally, Equation (6) is revised as follows:

$$S(i) = \sum_{j=1}^k w_{nj} S(j) \quad (8)$$

where w_{nj} is the normalized weight of a similar case j . The original w_{nj} is calculated as follows:

$$w_{nj} = w_{fj} w_{sj} \quad (9)$$

2.3.3 Retention and memory update in CBR

Memory update and self-learning is an important feature of CBR. There are some ways to update memory based on the analysis of utility and retention policy in CBR. The simple way is to update case library after solving new cases. The new cases attributes and solutions are added into the case library if the new solutions are successful (Aamodt and Plaza, 1994).

In this study, we used the simple retention way to update memory. After predicted ESR values of a new case, it was then judged whether the solution was successful. We compared the new predicted ESR values with mean ESR values \pm standard deviation (SD) of 224 cases. If the predicted ESR value falls in the interval, it was regarded as successful. The new case was then added into the case library with new index. The attributions and solution of the new predicted case were recorded in the library. Otherwise, it was recorded as a normal case. Fig. 1 indicates the approach to predict reference ESRs using CBR.

3 Results

3.1 Reference ESR value prediction results

In this study, the first step was to establish a case library. There were 224 cases of China in the case library, as we collected the 224 normal ESR values of young men at 182 different locations. The case location and need predicted location spatial distribution is shown in Fig. 2.

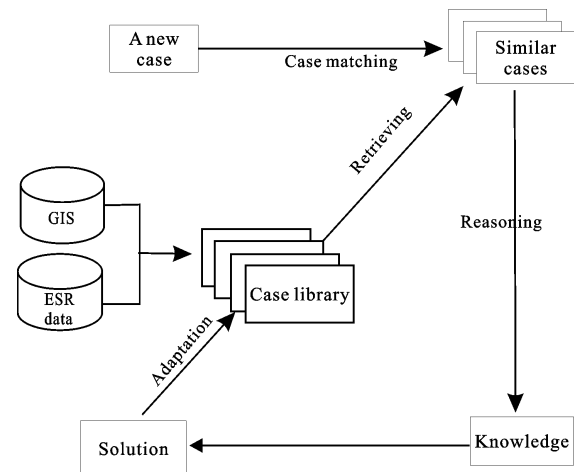


Fig. 1 Approach to predict reference erythrocyte sedimentation rate (ESR) using case-based reasoning (CBR)

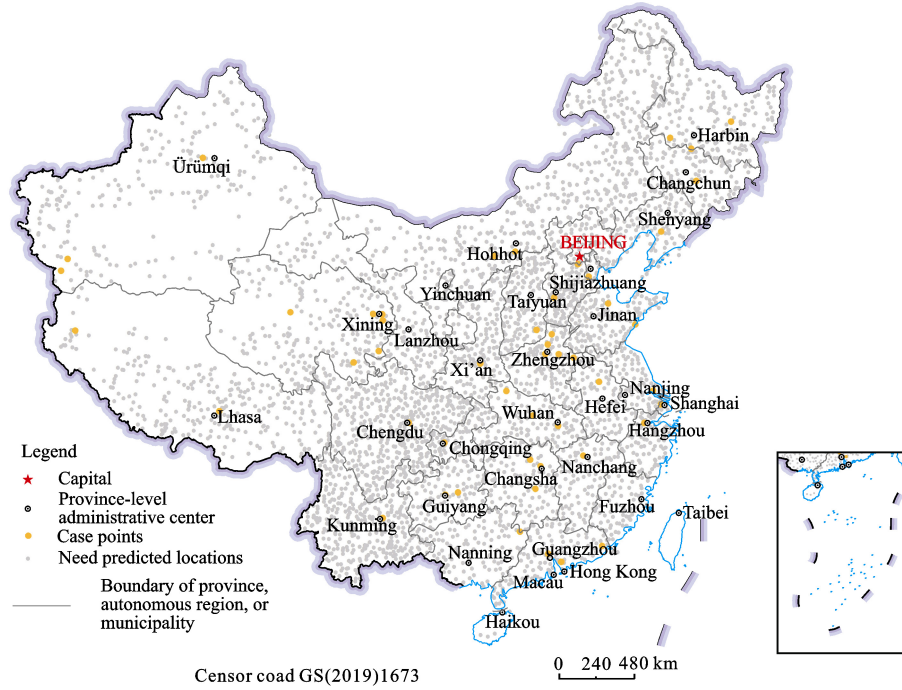


Fig. 2 Case and need predicted location spatial distribution of China

After establishing the case library, the value of k in the k-NN algorithm was determined in the model. Referring to frequently used 10 folder cross-validation methods for determining the k value in k-NN, experiments were carried out to test the influences of various k values (Dasarathy, 1991). The 224 cases were partitioned into 10 equal sized folders. Of the 10 folders, a single folder is retained as the validation data for testing k value and model accuracy, and the remaining $k-1$ folders are used as training data. The model is then run 10 times, with each of the 10 folders used exactly once as the validation data. At last, the predicted results of validation data of 10 folders can then be aggregated into a group to produce a single estimation. The experiments indicated that the increase in k values could improve the prediction accuracy but significantly increased the computation time. The average of the absolutely deviation of validation data became stabilized after k was greater than 15. Therefore, 15 neighbours were used to be the k value in the k-NN algorithm in the model. The performance indices, the average absolute deviation (AAD, %), mean square error (MSE), prediction accuracy (PA), Pearson Correlation coefficient between the observed and estimated data (r) were used to evaluate the accuracy of the whole model. They could be represented as follows:

$$AAD\% = \frac{1}{N} \sum_{i=1}^n \left| 100 \times \frac{V_{\text{obse}} - V_{\text{pred}}}{V_{\text{obse}}} \right| \quad (10)$$

$$MSE = \frac{1}{N} \sum_{i=1}^n (V_{\text{obse}} - V_{\text{pred}}) \times (V_{\text{obse}} - V_{\text{pred}}) \quad (11)$$

$$PA = 1 - \frac{1}{N} \sum_{i=1}^n |V_{\text{obse}} - V_{\text{pred}}| \quad (12)$$

$$r = \frac{\sum_{i=1}^n (V_{\text{obse},i} - \bar{V}_{\text{obse},i})(V_{\text{pred},i} - \bar{V}_{\text{pred},i})}{\sqrt{\sum_{i=1}^n (V_{\text{obse},i} - \bar{V}_{\text{obse},i})^2} \sqrt{\sum_{i=1}^n (V_{\text{pred},i} - \bar{V}_{\text{pred},i})^2}} \quad (13)$$

where N is the number of cases, V_{obse} and V_{pred} stand for the observed and predicted values, respectively.

To evaluate the difference in error deviation, the latter was divided into three grades. If the absolute error difference showed a 5% less difference than the observed value, the case was classified into the small error deviation grade (SED). If the absolute error difference was 5% to 10% compared with the observed value, the case was classified into the moderate error deviation grade (MED). Otherwise, the case was classified into the great error deviation grade (GED).

To ensure whether the CBR based model was the best one, the validation results on 10 folders of the CBR-based model, ANN and MLR models were compared (Fig. 3). Results of some cases were listed in Table 1. In addition, the accuracy assessments of the three models were compared (Table 2).

Regarding the accuracy assessments, the average absolute deviation (AAD%) and mean square error (MSE) were 33.07% and 9.02, respectively in the CBR model, values that were smaller than those of the ANN model (42.87% and 18.95, respectively) and MLR model (68.23% and 45.15, respectively). The prediction accuracy (PA) and the Pearson correlation coefficient between the observed and estimated data (r) in the CBR model were 66.93% and 0.78, respectively, values were greater than those of the ANN model (57.13% and 0.55,

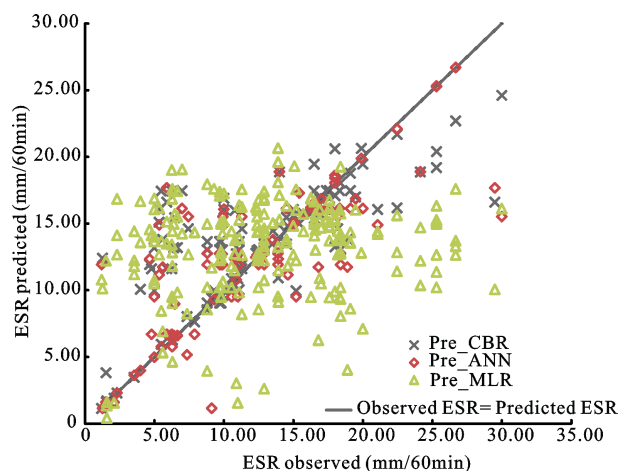


Fig. 3 Predicted results of case-based reasoning (CBR), artificial neural network (ANN) and multiple linear regression (MLR) model

Table 1 Comparison of the observed erythrocyte sedimentation rate (ESR) (V_{obse}) and predicted ESR (V_{predCBR} , V_{predANN} , V_{predMLR}) on some cases with case-based reasoning (CBR), artificial neural network (ANN) and multiple linear regression (MLR) model

No.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	V_{obse}	V_{predCBR}	V_{predANN}	V_{predMLR}
1	3.30	2724.40	12.20	63.00	569.90	30.40	3.00	117.20	39.13	17.38	17.46	16.13	15.26
2	23.50	2058.40	16.60	78.00	1269.00	25.80	2.70	114.31	30.52	12.90	13.50	13.06	16.76
3	100.00	2352.20	14.30	67.00	632.40	26.90	2.80	113.65	34.76	17.60	16.89	16.07	14.82
4	397.50	1646.10	13.70	70.00	553.30	26.70	2.00	108.95	34.27	16.59	16.06	14.90	15.72
5	506.10	1239.00	16.20	82.00	997.60	20.40	1.50	104.06	30.67	11.29	14.61	15.52	17.56
6	3269.00	2800.00	2.80	56.00	635.00	25.10	2.50	101.67	36.92	12.50	12.64	12.77	12.42
7	3500.00	2500.00	-1.50	60.00	615.00	22.70	2.50	101.62	34.75	6.30	6.28	6.32	9.03
8	4000.00	2500.00	-3.20	60.00	460.00	21.90	2.60	99.89	33.95	9.80	9.04	11.92	9.53
9	31.30	2763.70	12.30	60.00	571.90	30.80	2.50	116.46	39.92	10.55	10.07	9.51	14.42
10	1100.00	2400.00	9.00	60.00	600.00	27.50	1.60	113.08	36.18	15.73	15.75	15.71	14.41
11	96.00	2511.00	14.00	67.00	606.70	27.80	2.80	113.85	35.31	17.00	16.99	16.85	10.49
12	262.50	1900.00	15.30	72.00	833.30	24.70	1.80	110.79	32.65	16.27	15.77	16.19	9.48
13	5088.00	3100.00	-9.00	38.00	62.50	21.80	5.40	77.85	35.57	1.55	1.61	1.62	1.35
14	3010.00	2600.00	0.20	55.00	450.00	24.00	3.00	100.99	36.89	9.13	9.41	1.16	11.68
15	3648.90	3021.70	8.00	45.00	462.40	16.00	2.10	91.11	29.97	10.78	10.93	12.35	11.48
16	2.90	2500.00	21.50	81.00	1631.10	12.50	2.60	116.69	23.39	7.35	8.02	5.16	8.23
17	6.30	1906.00	21.80	79.00	1694.10	15.10	2.00	113.27	23.12	16.80	15.82	11.73	6.26
18	46.90	1903.90	17.40	77.00	1550.00	23.80	2.70	115.89	28.68	18.00	17.08	18.41	9.91
19	142.30	2867.00	4.20	65.00	524.30	42.20	3.90	126.63	45.75	14.00	14.87	18.88	15.28

Notes: $f_1, f_2, f_3, f_4, f_5, f_6, f_7$ are seven natural geographical factors representing altitude (m), annual sunshine hours (h), annual average relative humidity (%), annual average temperature ($^{\circ}\text{C}$), annual average precipitation (mm), annual temperature range ($^{\circ}\text{C}$), and annual average wind speed (m/s), respectively; f_8, f_9 represent longitude ($^{\circ}$) and latitude ($^{\circ}$); ESR indicates erythrocyte sedimentation rate; V_{obse} represent the observed normal ESR values; V_{predCBR} represent the predicted normal ESR values with case-based reasoning (CBR); V_{predANN} represent the predicted normal ESR values with artificial neural network (ANN); V_{predMLR} represent the predicted normal ESR values with multiple linear regression (MLR) model

Table 2 Comparison of whole model accuracy among case-based reasoning (CBR), artificial neural network (ANN), and multiple linear regression (MLR) model

	CBR	ANN	MLR
AAD	33.07%	42.87%	68.23%
MSE	9.02	18.95	45.15
PA	66.93%	57.13%	31.77%
<i>r</i>	0.78	0.55	0.37
SED	45.78%	42.22%	11.11%
MED	11.11%	12.89%	7.11%
GED	42.67%	44.44%	81.33%

Notes: AAD means average absolute deviation. MSE means mean square error. PA means prediction accuracy, and *r* represents Pearson correlation coefficient between the observed and estimated data; SED means error deviation difference showed a 5% less difference than the observed value, and MED means error deviation was 5% to 10% difference compared with the observed value; GED means error deviation is 10% greater difference compared with the observed value

respectively) and MLR model (31.77% and 0.37, respectively). The percent values of cases of the LED, MED and GED class were 45.78%, 11.11% and 42.67%, respectively, in the CBR-based model, which were better than those in the ANN model (42.22%, 12.89% and 44.44%, respectively) and MLR model (11.11%, 7.11% and 81.33%, respectively). The model accuracy assessment indicated that the CBR-based model had better prediction ability than the ANN-based model and MLR model.

For checking the spatial reliability of the proposed CBR model, paired-samples *t*-test method is used to check the difference between predicted value and observed value of 224 case data. The value of *t* is -1.229 with $P > 0.05$ that means the paired predicted value and observed value on 224 case data have no significant difference. The paired-sample *t*-test indicates that the proposed CBR model is reliable on the whole case

data.

Because the CBR-based model was reliable, reference ESR values of unseen 4390 locations (Fig. 2) could be predicted using the proposed model. For example, the reference ESR values of Lhasa, Guiyang, Yinchuan, Nanchang and Beijing City were predicted as shown in Table 3.

3.2 ESR values spatial distribution

Fig. 4 shows the spatial difference of reference ESR values. The high ESR reference areas are distributed in the eastern part of Sichuan, Chongqing, north-western part of Guangxi and south-eastern part of Guizhou Province; the reference ESR values are greater than 23 mm/60 min. The western part of Hubei and Henan, eastern part of Shaanxi and Shanxi, eastern part of Jilin and southwestern part of Heilongjiang provinces show the second-highest ESR reference values; the reference ESR values are greater than 18 mm/60 min. The western part of Tibet and southwestern part of Xinjiang show the lowest ESR reference values; the reference ESR values are lower than 5 mm/60 min.

4 Discussion

It is necessary to establish ESR reference values for interpreting whole blood tests in clinical settings. The variations in ESR are due to variations in the colloidal state of the plasma with consequent changes in the electric charges on the proteins and red cells (Greisheimer et al., 1929). Variations in age and gender, and residence location affect ESR through their effect on the colloidal state of the plasma indirectly. It is difficult to establish the ESR reference because ESR could be affected by many factors.

Table 3 Predicted reference erythrocyte sedimentation rate (ESR) values

City	Natural geographical factors									CBR solution
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	V_{pred}
Lhasa	3648.90	3021.70	8.00	45.00	462.40	16.00	2.10	91.11	29.97	10.93
Guiyang	1223.80	1371.00	15.30	77.00	1174.70	19.40	2.20	106.71	26.57	17.42
Yinchuan	1110.90	2867.00	9.00	53.00	186.30	33.00	1.80	106.27	38.47	12.51
Nanchang	46.90	1903.90	17.40	77.00	1550.00	23.80	2.70	115.89	28.68	17.08
Beijing	31.30	2763.70	12.30	60.00	571.90	30.80	2.50	116.46	39.92	10.07

Notes: V_{pred} represents the predicted normal ESR values with case-based reasoning (CBR); $f_1, f_2, f_3, f_4, f_5, f_6, f_7$ are seven natural geographical factors representing altitude (m), annual sunshine hours (hours), annual average relative humidity (%), annual average temperature ($^{\circ}\text{C}$), annual average precipitation (mm), annual temperature range ($^{\circ}\text{C}$), and annual average wind speed (m/s), respectively; f_8, f_9 represent longitude ($^{\circ}$) and latitude ($^{\circ}$); CBR means case-based reasoning

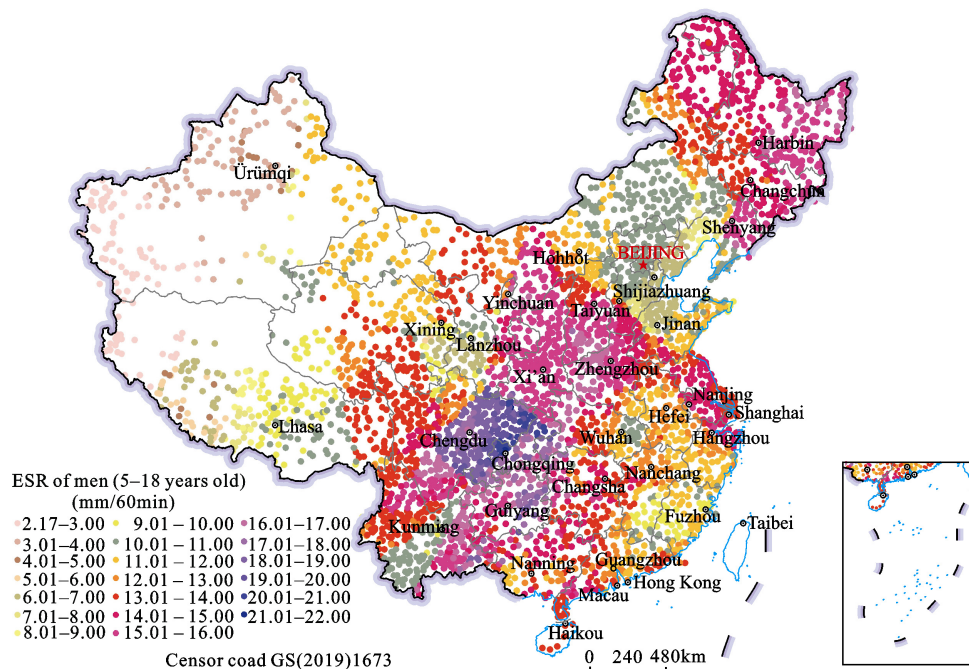


Fig. 4 Spatial distribution of erythrocyte sedimentation rate (ESR) reference values

In this study, we proposed a case-dependent method to establish reference ESR values with natural geographical factors and location in China. Compared with the multiple linear regressions (Ge et al., 1999; 2001; Ge, 2004), and the artificial neural network (Yang et al., 2013), CBR had a higher prediction accuracy. It was a better method to predict reference ESR values than the ANN and MLR models. Moreover, the mechanism underlying how geographical factors affect ESR cannot be understood completely and the relationship between ESR and geographical factors is indirect and nonlinear. The knowledge domain is weak and limited. The similarity-based ‘lazy’ reasoning CBR method is better than rule-based model because the knowledge domain is weak (Kolodner, 1993). Based on the prediction accuracy and model running mechanism, CBR is a better model than the ANN and MLR models to establish reference ESR values.

Compared with the MLR model, the CBR-based model and ANN model are nonlinear models and local optimized nonlinear models, making them provide better prediction accuracy for nonlinear problems (Li and Yeh, 2002). Our research reveals the similar results, and the results of the CBR and ANN models are all better than that of the MLR model in this study area. We also found that the CBR-based model has the same predic-

tion trend as the ANN model. If CBR-based model has a great error deviation for a case, the ANN also has great error deviation, and the error deviation of the ANN model is almost greater than that of the CBR-based model. The error deviation of some cases was greater than 0.10 in the CBR-based model, with similar results using the ANN model. The CBR-based model provides higher prediction accuracy than the ANN model mainly because the former measures the weight of the case by the spatial Euclidean distance and the feature weight in feature space, while the latter only measures the feature weight. For checking the reliability of the proposed model, the predicted ESR values are compared with that of measured in other places. The results indicate that the predicted ESR values fall in the interval of measured values (Jou et al., 2011; Kawashiri et al., 2011; Sonmez et al., 2014; Isiksacan et al., 2016; Kratz et al., 2017) that means the proposed CBR model is reliable.

We also find that some cases have the highest error deviations of all three models. The reason is that it has more than 5 cases in one location with the same attributes having different ESR values in the case library. For example, seven cases have the smallest ESR value of 1.52 and the greatest ESR value of 2.10, resulting in a great error deviation at the location of Beijing City. The situation is the result of the data sources. We col-

lected cases from published studies, and the normal ESR values (224) of those released were calculated with samples (13 623) using the mean ESR \pm standard deviation (SD). This problem may be solved by using the original 13 623 samples that were not previously reported. Another problem that has also emerged is: how to judge which ESR is the reference ESR value when researchers report different normal ESR values for the same location with a specific age and gender.

It is important to explore the ways in which ESR values can be influenced by natural geographical factors. Researchers have proven that the ESR values decrease with the increase in altitude because air becomes thin, i.e., the oxygen content gradually decreases while the altitude rises. As a result, the amount of red blood cells increases. This induces a fall in the ESR reference value in healthy subjects (Ge et al., 2001; Ge, 2004; Yang et al., 2013). However, the exact mechanisms surrounding the influence of ESR values by such natural geographical factors need to be further studied.

Because ESR values can be affected by natural geographical factors, the reference ESR values according to age and gender should be geographically established. Physicians and researchers have found differences in ESR values in different cities, and they have begun to interpret the ESR test in the clinic (Pincherle and Shanks, 1967; Ge et al., 1999; Ge et al., 2001; Ge, 2004; Yang et al., 2013). Spatial difference ESR reference values are more accuracy in clinic.

5 Conclusions

Currently, normal ESRs are measured as reference values at most hospitals and medical institutions. The reference ESR values have limitations regardless of the differences in age, gender and location. In this study, a case-based model was developed to predict the local normal ESR values by considering the corresponding local natural geographical factors. The model was tested using 10 folder cross validation. The results showed that the predicted values are very close to the observed values. Based on paired-samples *t*-test analysis between predicted ESR values and observed ESR values on 224 case data of China, the proposed CBR model is reliable though the case data are rare in western district of China. CBR could be an effective method to predict reference ESR values because of its complexity and limited domain knowledge. Although the reference ESR

values can be established with natural geographical factors and location using cases, the mechanism and process that how natural geographical factors affect reference ESR values should be further explored.

In the current study, we tried to geographically establish reference ESR values of young men aged 5 to 18 using a CBR-based model. A standard system of ESR values according to whole age and gender should be established geographically in the future.

References

- A Aamodt A, Plaza E, 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1): 39–59. doi: 10.3233/AIC-1994-7104
- Aha D W, Kibler D, Albert M K, 1991. Instance-based learning algorithms. *Machine Learning*, 6(1): 37–66. doi: 10.1007/BF00153759
- Ansell B, Bywaters E G L, 1958. The ‘Unexplained’ high erythrocyte sedimentation rate. *British Medical Journal*, 1(5067): 372–374. doi: 10.1136/bmj.1.5067.372
- Cui J H, Zhang X Z, Zhang F et al., 2001. Change of erythropoietin on high altitude hypoxia adaptation mechanism. *Clinical Journal of Medical Officer*, 29(3): 45–47. (in Chinese)
- da C Sousa J V, dos Santos M N N, Magna L A et al., 2018. Validation of a fractional model for erythrocyte sedimentation rate. *Computational and Applied Mathematics*, 37(5): 6903–6919. doi: 10.1007/s40314-018-0717-0
- Dasarathy B V, 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos: IEEE Computer Society Press.
- Fahraeus R, 1929. The Suspension stability of the blood. *Physiological Reviews*, 9(2): 241–274. doi: 10.1152/physrev.1929.9.2.241
- Ge M, Yan Y, Ren Z Y et al., 1999. The relationship between normal erythrocyte sedimentation rate of Chinese young people and geographical factors. *Clinical Hemorheology and Microcirculation*, 20(3): 151–157.
- Ge M, Ren Z Y, Yang Q S et al., 2001. The relationship between reference value of old people’s erythrocyte sedimentation rate and altitude. *Clinical Hemorheology and Microcirculation*, 24(3): 155–159.
- Ge M, 2004. Reference value of younger people’s erythrocyte sedimentation rate and altitude. *Journal of Laboratory and Clinical Medicine*, 143(6): 367–368. doi: 10.1016/j.lab.2004.03.006
- Gierl L, Steffen D, Ihracky D et al., 2003. Methods, architecture, evaluation and usability of a case-based antibiotics advisor. *Computer Methods and Programs in Biomedicine*, 72(2): 139–154. doi: 10.1016/S0169-2607(02)00121-9
- Greisheimer E M, Warwick M, Walton M, 1929. The relationship between fibrin content and sedimentation index in orthopedic cases. *American Journal of Diseases of Children*, 37(5):

- 953–956. doi: 10.1001/archpedi.1929.01930050063007
- Hale A J, Ricotta D N, Freed J A. 2019. Evaluating the erythrocyte sedimentation rate. *JAMA*, 321(14): 1404–1405. doi: 10.1001/jama.2019.1178
- Hoiland R L, Smith K J, Carter H H et al., 2017. Shear-mediated dilation of the internal carotid artery occurs independent of hypercapnia. *American Journal of Physiology Heart and Circulatory Physiology*, 313(1): H24–H31. doi: 10.1152/ajpheart.00119.2017
- Holt A, 1999. Applying case-based reasoning techniques in GIS. *International Journal of Geographical Information Science*, 13(1): 9–25. doi: 10.1080/136588199241436
- Houben I, Wehenkel L, Pavella M, 1995. Coupling of K-NN with decision trees for power system transient stability assessment. In: *Proceedings of International Conference on Control Applications*. Albany, NY, USA: IEEE. doi: 10.1109/CCA.1995.555856
- Isiksacan Z, Erel O, Elbuen C, 2016. A portable microfluidic system for rapid measurement of the erythrocyte sedimentation rate. *Lab on A Chip*, 16(24): 4682–4690. doi: 10.1039/c6lc01036a
- Jalali V, Leake D, 2016. Enhancing case-based regression with automatically-generated ensembles of adaptations. *Journal of Intelligent Information Systems*, 46(2): 237–258. doi: 10.1007/s10844-015-0377-0
- Jou J M, Lewis S M, Briggs C et al., 2011. ICSH review of the measurement of the erythrocyte sedimentation rate. *International Journal of Laboratory Hematology*, 33(2): 125–132. doi: 10.1111/j.1751-553X.2011.01302.x
- Kawashiri S Y, Kawakami A, Iwamoto N et al., 2011. Disease activity score 28 may overestimate the remission induction of rheumatoid arthritis patients treated with tocilizumab: comparison with the remission by the clinical disease activity index. *Modern Rheumatology*, 21(4): 365–369. doi: 10.3109/s10165-010-0402-7
- Kibler D, Aha D W, Albert M K, 1989. Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5(2): 51–57. doi: 10.1111/j.1467-8640.1989.tb00315.x
- Kolodner J, 1993. *Case-based Reasoning*. San Mateo: Morgan Kaufmann.
- Kolodner J L, 1992. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1): 3–34. doi: 10.1007/bf00155578
- Kratz A, Plebani M, Peng M et al., 2017. ICSH recommendations for modified and alternate methods measuring the erythrocyte sedimentation rate. *International Journal of Laboratory Hematology*, 39(5): 448–457. doi: 10.1111/ijlh.12693
- Kuznetsova D A, Sizova E N, Tuliakova O V, 2013. Influence of high altitude on the functional status and morbidity of teenagers. *Gigiena I Sanitariia*, (3): 77–80.
- Li X, Yeh A G O, 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, 16(4): 323–343. doi: 10.1080/13658810210137004
- Miller A, Green M, Robinson D, 1983. Simple rule for calculating normal erythrocyte sedimentation rate. *British Medical Journal*, 286(6361): 266. doi: 10.1136/bmj.286.6361.266
- Moen J K, Reimann H A, 1933. Plasma protein changes and suspension stability of the blood in lobar pneumonia. *The Journal of Clinical Investigation*, 12(3): 589–598. doi: 10.1172/JCI100522
- Montani S, Jain L C, 2010. *Successful Case-Based Reasoning Applications*. Berlin: Springer Verlag.
- Näyhä S, 1987. Normal variation in erythrocyte sedimentation rate in males over 50 years old. *Scandinavian Journal of Primary Health Care*, 5(1): 5–8. doi: 10.3109/02813438709024179
- Olshaker J S, Jerrard D A, 1997. The erythrocyte sedimentation rate. *The Journal of Emergency Medicine*, 15(6): 869–874. doi: 10.1016/S0736-4679(97)00197-2
- Pincherle G, Shanks J, 1967. Value of the erythrocyte sedimentation rate as a screening test. *British Journal of Preventive & Social Medicine*, 21(3): 133–136. doi: 10.1136/jech.21.3.133
- Ropes M W, Rossmeisl E, Bauer W, 1939. The relationship between the erythrocyte sedimentation rate and the plasma proteins. *The Journal of Clinical Investigation*, 18(6): 791–798. doi: 10.1172/JCI101096
- Schäfer V S, Weiß K, Krause A et al., 2018. Does erythrocyte sedimentation rate reflect and discriminate flare from infection in systemic lupus erythematosus? Correlation with clinical and laboratory parameters of disease activity. *Clinical Rheumatology*, 37(7): 1835–1844. doi: 10.1007/s10067-018-4093-3
- Sharland D E, 1980. Erythrocyte sedimentation rate: the normal range in the elderly. *Journal of the American Geriatrics Society*, 28(8): 346–348. doi: 10.1111/j.1532-5415.1980.tb01096.x
- Siemons L, ten Klooster P M, Vonkeman H E et al., 2014. How age and sex affect the erythrocyte sedimentation rate and C-reactive protein in early rheumatoid arthritis. *BMC Musculoskeletal Disorders*, 15(1): 368. doi: 10.1186/1471-2474-15-368
- Sönmez Ç, Guntas G, Kaymak A Ö et al., 2014. Comparison of erythrocyte sedimentation rate results of test-1 and automatic westergren device with reference westergren method. *Gazi Medical Journal*, 25(2): 52–54. doi: 10.12996/gmj.2014.15
- Suckling P V, 1957. The erythrocyte sedimentation rate in kwashiorkor in Cape coloured children. *South African Journal Of Laboratory And Clinical Medicine. Suid-Afrikaanse Tydskrif vir Laboratorium-en Kliniekwerk*, 3(4): 308–315.
- Theil H, 1967. *Economics and Information Theory*. Chicago: North-Holland Pub. Co.
- van Atteveld V A, van Ancum J M, Reijnierse E M et al., 2019. Erythrocyte sedimentation rate and albumin as markers of inflammation are associated with measures of sarcopenia: a cross-sectional study. *BMC Geriatrics*, 19: 233. doi: 10.1186/s12877-019-1253-5
- Wetteland P, Røger M, Solberg H E et al., 1996. Population-based erythrocyte sedimentation rates in 3910 subjectively healthy Norwegian adults. A statistical study based on men and women

- from the Oslo area. *Journal of Internal Medicine*, 240(3): 125–131. doi: 10.1046/j.1365-2796.1996.30295851000.x
- Wyller D J, 1977. Diagnostic implications of markedly elevated erythrocyte sedimentation rate: a reevaluation. *Southern Medical Journal*, 70(12): 1428–1430. doi: 10.1097/00007611-197712000-00015
- Yang Q S, Mwenda K M, Ge M, 2013. Incorporating geographical factors with artificial neural networks to predict reference values of erythrocyte sedimentation rate. *International Journal of Health Geographics*, 12(1): 11. doi: 10.1186/1476-072x-12-11
- Zacharski L R, Kyle R A, 1967. Significance of extreme elevation of erythrocyte sedimentation rate. *JAMA*, 202(4): 264–266. doi: 10.1001/jama.1967.03130170064008
- Zouboules S M, Lafave H C, O'Halloran K D et al., 2018. Renal reactivity: acid-base compensation during incremental ascent to high altitude. *The Journal of Physiology*, 596(24): 6191–6203. doi: 10.1113/JP276973