BMC Public Health

RESEARCH

Open Access

Questionnaire and LGBM Model for Assessing Health Literacy levels of Mongolians in China

Yan Hong[1*†] and Xiaoda Zhang[2†]

Abstract

Background: It is difficult to accurately assess the health literacy(HL) level of Mongolians by using Chinese conventional HL questionnaire, due to their particularity in language, culture and living environment. Therefore, it is very important to design an exclusive HL questionnaire for them. In addition, the existing statistical models cannot meet the requirement of HL assessment with high precision, so it is necessary to study a new HL assessment model.

Methods: A HL questionnaire with 68 questions is designed by combing the HLS-EU-Q47and the characteristics of Mongolians in China. 742 Mongolians aged 18 to 87 in Inner Mongolia of China answered the questionnaire. A data set with 742 samples is constructed, where each sample has 68 features and 1 target. Based on it, the XGB and LGBM regression models are respectively constructed to assess the HL levels of respondents, and their evaluation effects are compared. The impact of each question on the HL level is quantitatively analyzed by using the feature-importance function in LGBM model to verify the effectiveness of the questionnaire and to find the key factors for affecting HL.

Results: The HL questionnaire has the high reliability, which is reflected by the high internal consistency (Cronbach's coefficient=0.807) and test-retest reliability (Mutual Information Score= 0.803). The validity of the HL questionnaire is obtained by solving KMO and Bartlett Spherical Test Chi-square Value, which are 0.765 and 2486 ($p < 0.001$), respectively. R^2 index and the absolute error obtained by using the HL assessment model based on LGBM are 0.98347 and 11, which are better than ones by applying the model based-XGB, respectively. The quantitative analysis results show that all 68 questions have influence on HL level, but their degree are different. The first three factors are age, salary level, the judgment ability for the HL information in media, respectively. The HL level distribution of the respondents was 66.71% excellent, 25.74% good and 7.54% poor, respectively.

Conclusions: The presented HL questionnaire with 68 questions and LGBM regression model can obtain the HL level assessment results with high precision for Mongolians in China. The impact of each question in the questionnaire on the final assessment results can be quantified by using the feature-importance function in LGBM model, which is better than the existing qualitative analysis methods.

Keywords: Health literacy, Assessment model, LGBM regression model, Questionnaire design, Quantitative analysis

Background

Health literacy (HL) was a complex and multidimensional concept related to literacy [1]. It was defined as "an ability that people maintain and promote themselves health by acquiring, understanding, and using health information" by International Union for Health Promotion

[†] Yan Hong and Xiaoda Zhang contributed equally to this work.

*Correspondence: 03hongyan@163.com

[1] School of Nursing, Inner Mongolia Minzu University, 028000 Tongliao, China
Full list of author information is available at the end of the article

and Education [2]. Higher-level HL also included critical thinking, analysis, decision-making, and problem-solving in health-related matters. It was important to improve HL, because it not only could promote peoples' heathy and application awareness in health services but also could reduce the risk of disease, and thus reduce social burden. It was necessary that public libraries with empowerment, equitable, inclusive, collaborative and integrated characteristics were constructed to improve the HL of individuals, communities, organizations and countries [3].

In recent years, the HL had been widely investigated, in which interviewees included adults, adolescents, children, and patients. The low HL and medication literacy were main risk factors for health of children and adults [4]. A cross-sectional study in [5] showed that about 28-38 ones among 120 students had poor health literacy. It was concluded in [6] that adolescents with psychological symptoms and low HL had non-suicidal risks, which would be solved by intervening their mental health and behavior problems. The digital HL survey for college students found that the students with different educational levels had the different usage levels for digital HL [7].

The relationships between HL and some diseases were investigated widely. It was discovered that the high digital HL could increase the number of cancer survivors [8]. A systematic review for HL in individuals at risk for alzheimers dementia was developed in [9], and concluded that alzheimer's disease patients couldn't use HL skills, so it was very important to analyze alzheimer's disease patients' needs, and to give them some essential information, because it could help them to make decisions during specific medical situations. By investigating the echinococcosis-specific HL of the Tibet Plateau in China, it was found that the echinococcosis-specific HL was a key factor to prevent echinococcosis [10]. The importance of HL for preventing or curing borderline personality disorder [11], hypertension [12], chronic disease [13], atherosclerosis, ischemic heart disease [14], and type II diabetes mellitus [15] were also investigated, respectively. However, the above studies used the statistics methods.

Information technology was used as a data collection tool applied to HL [16–19], but it wasn't found that information technology was applied to assess HL level. However, with the rapid development of information science, big data analysis and machine learning technology had been applied widely in medical field to solve statistical problems. For example, the early detection of breast cancer based on CNN and light gradient boosting machine (LGBM) [20], the recognition of cancer cells in blood based on GBDT algorithm[21], classifing and predicting for the survival probability of patients with cancer comorbidities by LGBM [22], analyzing for 12 characteristics of breast cancer by random forest methods [23], and diagnosis for thyroid cancer, colon cancer, liver cancer by SVM approach [24], and predicting disease progression of breast cancer by XGBoost (XGB) [25], and so on. Therefore, we think that it is a good idea that big data analysis and machine learning technique were used to predict or assess HL for an individual or group and quantitatively analyze the effectiveness of every factor on their hearth literacy level.

China has 56 ethnic groups, in which the Mongolian is an Ethnic minority, and it accounts for 8.89% of the total population in China. It is difficult to accurately assess their HL level by using existing Chinese conventional HL questionnaire, due to the particularity of language, writing and living environment of ethnic minorities, Therefore, it is very important to design an exclusive HL questionnaire and assessment model for them.

Therefore, 742 Mongolian in Inner Mongolia, China are surveyed in this paper. The HL questionnaire with 68 questions, the LGBM assessment model with high precision, and a quantitative analysis method for every question are presented.

The innovations of this paper are as follows:

i) From four dimensions, the HL questionnaire with 68 questions is designed by both improving the HLS-EU-Q47 and analyzing Mongolian's characteristics in Inner Mongolia, China. Four dimensions include health concepts and knowledge literacy, healthy lifestyle and behavior, and healthy skills, as well as health status and disease history.
ii) The data set based on the HL questionnaire is constructed, and the LGBM HL assessment model is presented, which can obtain the higher assessment accuracy than the presented XGB HL assessment model and the statistical models.
iii) The impact of each question in the questionnaire on the HL level is quantitatively analyzed one by one by using the feature-importance function in LGBM model in order to verify the effectiveness of the questionnaire and to find the key factors for affecting HL levels.
iv) The above approaches can provide a new idea for investing HL level of other ethnic minorities in China or ethnic minorities in other countries.

Methods

Design and setting

The cross-sectional study was carried out for a period of six months between November 2018 to April 2019 in Inner Mongolia, China. The participants was Mongolian, over 18 years old, with no history of psychiatric disorders, and able to read and comprehend the Chinese language. Each participant introduced with the informed consent, upon their approval, the health literacy questionnaire was provided. 742 participants were invited to the survey. The baseline characteristics of participants are shown in Table 1.

HL questionnaire design

The health literacy assessment survey questionnaire is a tool to assess the health literacy level of respondents, which is designed by a health organization or researchers. The survey questionnaire have a lots of questions, and each question is assigned a score. The health literacy levels of respondents are decided by the respondents' scores. It is very important to design a suitable survey questionnaire for the health literacy assessment. The HLS-EU-Q47 was developed in 2011, which has 86 questions and mainly investigated the peoples' abilities to understand the health-related issues and to get the health-related knowledge in the complicated situations caused by the inadequate health literacy. However, the HLS-EU-Q47 can't be applied directly to Chinese. It is necessary to design a health literacy assessment survey questionnaire. In order to reduce participants' workload, we combined some similar items in the HLS-EU-Q47, such that 86 items are changed as 47 items. And according to the characteristics of Monggolians in Inner Mongolia, China, 21 questions are added. Therefore, we construct a HL survey questionnaire with 68 questions from four dimensions. Four dimensions include health concepts and knowledge literacy, healthy lifestyle and behavior, and healthy skills, as well as health status and disease history. 68 questions are divided into three parts. The first part is the questions about the respondents' general situation, which are age area, gender, and territory; The second part is the respondents' own health-related questions, namely health status and disease history; The third part is the health-related issues. Some similar items among 86 items in the EHLSQ are combined as 47 items in order to reduce the workload of participants. 40 questions in the 47 questions are put in the third part, and others are given in the second part. The scoring method for each question are:

i) Age, height, and weight of respondents are recorded according to their actual values.
ii) The scores in the other 65 items of respondents are assessed by using a five-point, self-reported Likert type scale, such as very easy, fairly easy, fairly difficult, very difficult, and unknown . The lowest score is 1, and the highest score is 5.

The score for each of the 68 questions is added to produce the HL level score for individual.

Table 1 Baseline characteristics of participants

Characteristics	Participants (%)	Men (%)	Women(%)
Age group			
18-30	67(9.0)	29(8.1)	38(9.9)
31-40	286(38.6)	108(30.1)	178(46.5)
41-50	164(22.1)	94(26.2)	70(18.3)
51-60	164(22.1)	93(25.9)	71(18.5)
≥61	61(8.2)	35(9.7)	26(6.8)
Total	742 (100)	359(100)	383(100)
Education(Edu)			
Higher Edu.	88(11.9)	49(13.6)	39(10.2)
Vocational Edu.	447(60.1)	201(56.0)	246(64.2)
Secondary Edu.	105(14.2)	58(16.2)	47(12.3)
Elementary Edu.	80(10.8)	41(11.4)	39(10.2)
Not schooled	22(3.0)	10(2.8)	12(3.1)
Marriage			
Married	606(81.7)	302(84.1)	304)79.4)
Not married	85(11.5)	33(9.2)	52(13.6)
Divorced	51(6.8)	24(6.7)	27(7.0)
Occupation			
State sector	392(52.9)	171(47.6)	221(57.7)
Private sector	322(43.4)	179(49.9)	143(37.3)
Retired	28(3.7)	9(2.5)	19(5.0)
Health insurance			
National	651(87.7)	318(88.6)	333(86.9)
Private	91(12.3)	41(11.4)	50(13.1)
Living with			
Alone	93(12.5)	41(11.4)	52(13.6)
Family	614(81.8)	302(84.1)	312(81.4)
Relatives	35(4.7)	10(4.5)	19(5.0)

Reliability and validity of the HL questionnaire analysis

According to the HL questionnaires completed by 742 Mongolians, the reliability of the designed HL questionnaire is analyzed by calculating the Cronbach's α value and the Mutual Information Score. They can be obtained by using the Python programs designed by ourselves.

The validity of the HL questionnaire is verified by solving the KMO value and Bartlett Spherical Test Chi-square of the HL questionnaires completed by 742 Mongolians. The KMO value and Bartlett Spherical Test Chi-square can be obtained by using a Python program designed by ourselves.

Data set construction

According to the HL questionnaires completed by 742 Mongolians, the following data set is constructed.

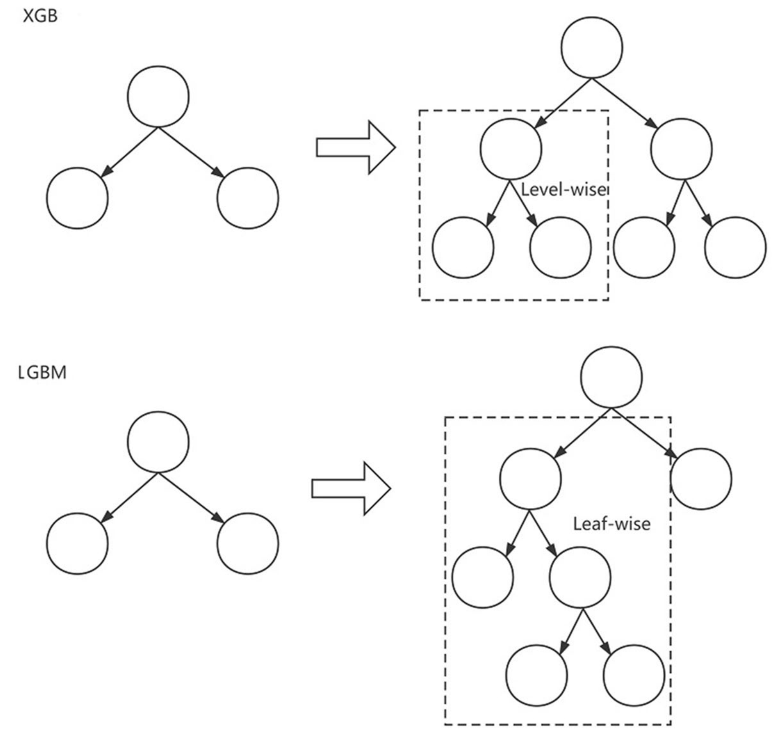

Fig. 1 The difference between level-wise and leaf-wise

$$\begin{aligned} &\Xi = \{(\varpi_i^1, \varpi_i^2, \cdots, \varpi_i^{68}, H_i)\} \\ &for\ i = 1, 2, \cdots, 742 \end{aligned} \tag{1}$$

where Ξ describes the data set that have 742 samples with 68 features. where i is the the i^{th} sample, ϖ_i^1-ϖ_i^{68} are 68 characteristics of the i^{th} sample; H_i is the target value of the i^{th} sample, and H_i describes the scores obtained by the i^{th} respondent.

XGB and LGBM model construction

The GBDT, XGB and LGBM are machine learning models. The XGB model developed from the GBDT model. Compared with the GBDT model, the XGB uses the second-order Taylor expansion for the loss function, such that the prediction accuracy is improved. However, there is lower efficiency in the features selection and growth of the decision tree due to XGB uses hierarchical leaf node selection method. In order to solve the problem, the Histogram algorithm and growing leafs with maximum split gain method were applied in LGBM model, which can improve greatly the prediction accuracy and efficiency. In addition, the maximum depth limit is added to the growth of the algorithm, which can avoid over-fitting under guaranteeing the high training efficiency. Therefore, XGB and LGBM regression models are constructed to assess the HL levels of respondents in this paper. The comparison diagram of XGB and LGBM is shown in the Fig. 1.

The following evaluation index is considered when the LGBM model is applied to predict the considered target, which is called as R^2 index.

$$R^2 = 1 - \frac{\sum(y_{pred} - y_{true})^2}{\sum(y_{true} - \bar{y})^2} \tag{2}$$

where y_{pred} is the predicted value, y_{true} is the true value, and $\bar{y}$ is the average value of the samples. $0 < R^2 < 1$, and a large R^2 value indicates a high prediction accuracy.

Based on the data set 1, the XGB regression model and LGBM regression model to assess the HL for 742 respondent are constructed, respectively. 80% samples in the data set Ξ are designed as training samples, and others are looked as testing samples. The flow diagram to evaluate HL by using LGBM regression model is given in Fig. 2.

Quantifing the impact of each question on the final assessment results

By using the 'feature-importance' function in LGBM model, we analyze quantitatively the influences of 68 features on the HL assessment results, and find the key

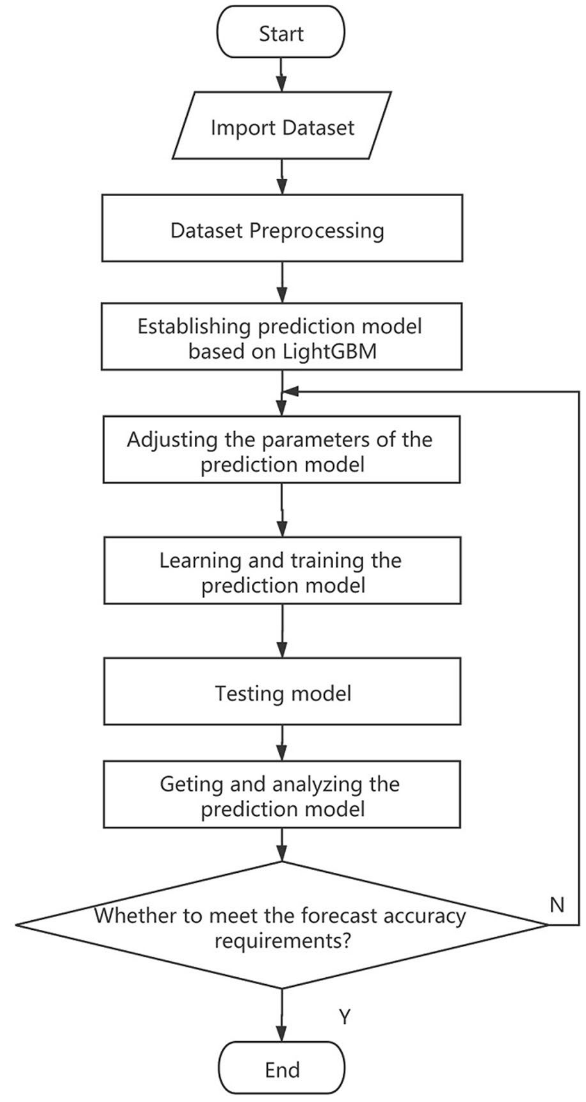
Start
Import Dataset
Dataset Preprocessing
Establishing prediction model based on LightGBM
Adjusting the parameters of the prediction model
Learning and training the prediction model
Testing model
Geting and analyzing the prediction model
Whether to meet the forecast accuracy requirements?
N
Y
End

Fig. 2 Flow chart of health literacy prediction model

factors to affect HL level of Mongolian in Inner Mongolia, China.

The above all processes are completed by our own Python programming.

Results

Reliability and validity of the HL questionnaire results

According to the HL questionnaires completed by 742 Mongolians, we calculate the Cronbach's α value and the Mutual Information Score to analyze the reliability of the designed HL questionnaire. We have obtained that the Cronbach's α is 0.807 and the Mutual Information Score equals 0.803 by using the Python programs designed by ourselves. It can be seen that the designed HL questionnaire has the high reliability.

Table 2 R^2 indexes obtained by the health literacy prediction models based on XGB and LGBM

Model	XGB	LGBM
R^2 Score	0.97553	0.98347

In order to verify the validity of the HL questionnaire, the KMO and Bartlett Spherical Test Chi-square Value of the HL questionnaires completed by 742 Mongolians are solved by using a Python program designed by ourselves., which are 0.765 and 2486 ($p < 0.001$), respectively. These results show that the designed HL questionnaire has good validity.

HL assessment results

The HL assessment models based on XGB and LGBM are trained and tested by training samples and testing samples, respectively. R^2 indexes obtained by XGB and LGBM regression models are shown in the following Table 2, respectively. From Table 2, it can be seen that the LGBM regression model has more higher HL assessment accuracy than the XGB regression model, and its R^2 value is 0.98347, which can meet the actual demands for the HL evaluation.

The Fig. 3 shows that the comparison between the results predicted by XGB and LGBM regression models and the true values. The red line describes the true values; The blue line represents the values predicted by LGBM; The green line is the results predicted by XGB. From the Fig. 3, we know that the high prediction results can be obtained by using XGB and LGBM, respectively. However, the prediction errors from two models can't be found. Therefore, we draw the absolute error curves obtained by using XGB and LGBM, which are given in Fig. 4. In Fig. 4, the blue line represents the absolute errors between the values predicted by LGBM and the true values. The green line is the absolute errors between the values predicted by XGB and the true values. It can be seen from Fig. 4 that the absolute errors between the values predicted by LGBM and the true values are less than 11, while the absolute errors between the values predicted by XGB and the true values are less than 15. Therefore, the health literacy prediction model based on the LGBM is more effective than one based on the XGB.

The HL level distribution of the respondents was 66.71% excellent, 25.74% good and 7.54% poor, respectively, which are given in Fig. 5. The percentage of men respondents who scored good and excellent is 58.2 % , which is better than that of women respondents(28.1 %). The HL scores of the urban respondents are higher

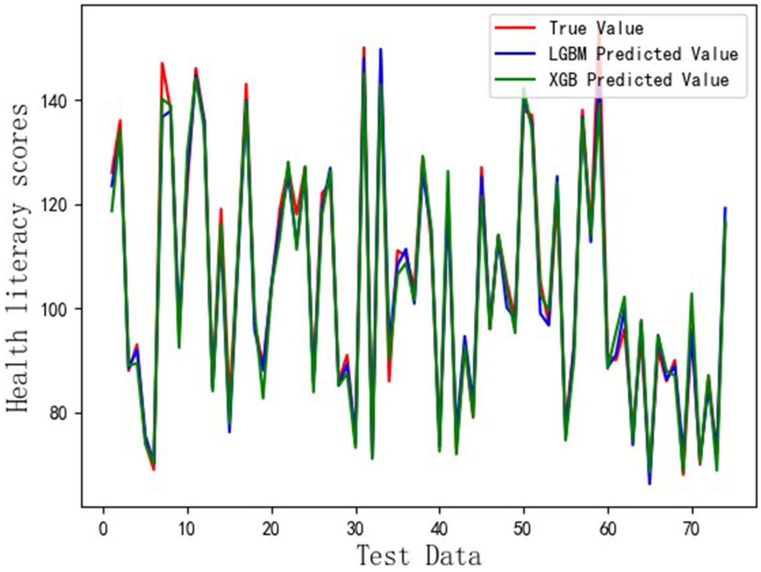

Fig. 3 HL scores assessed by using XGB and LGBM

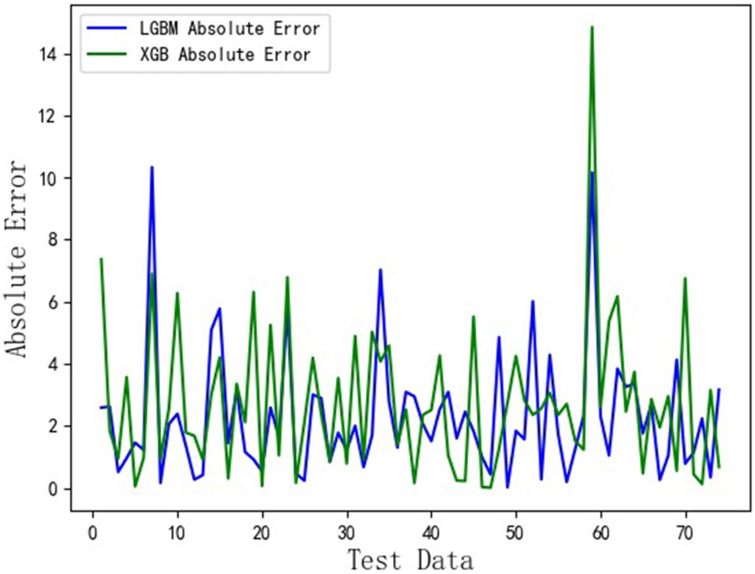

Fig. 4 Absolute error rate between the HL assessment results and true values

than ones of the rural residents. In addition, we find that there is a positive linear correlation between the level of HL and the educational background of the respondents.

Calculating and analyzing the influences of each question on HL assessment result

The influence of each question in 68 questions on the HL assessment results is respectively calculated by the 'feature-importance' function in LGBM, which are shown in Fig. 6. It can be seen that the biggest impact factor is 1105, and the smallest impact factor is 23. The numbers in Fig. 6 are dimensionless. The *Column*_16 (age of the respondents) has the highest influence on the HL level. The *Column*_27 (the salary level of the respondents) is second. The *Column*_36 is third, which is the ability of the interviewees to judge relevant health information in the media. The forth factor, the fifth factor, and the sixth factor are *Column*_25 (probability of medical

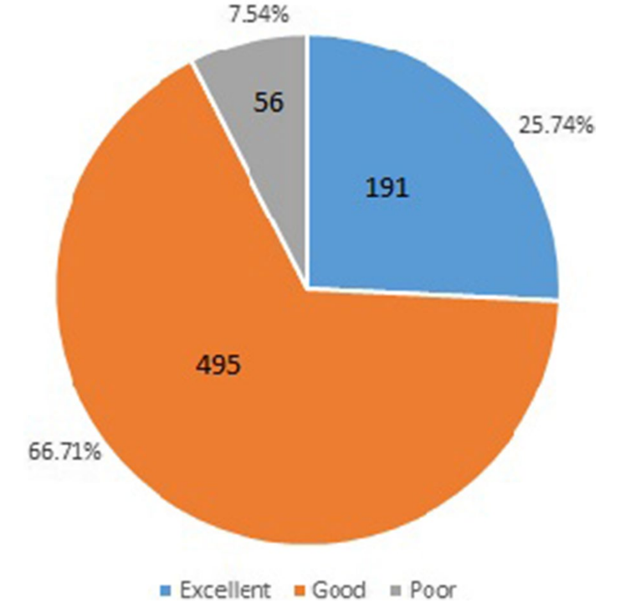

Fig. 5 HL level distribution of the respondents

attendance), $Column\_43$ (knowing about vaccinations and checkups), and $Column\_53$ (obtaining healthy eating information), respectively. The influence of Gender ($Column\_1$) on the HL level is 69. The impact indexes of the Territory ($Column\_2$), Education background ($Column\_20$), and Professional ($Column\_21$) are 96, 69, and 71, respectively. The forth dimension (health status and disease history) of the HL questionnaire is reflected by the $Column\_3, 4, 5, 7, 8, 9, 10$ in Fig. 6, where the impact index of the health status ($Column\_3$) is the largest, which is 168. The least influence on the final health index is the $Column\_6$, which describes the insurance type used by the respondents. According to the above analysis, it also can be seen that all questions in the designed questionnaire are reasonable, because that they affect the HL assessment results by varying degrees.

The above results can be summarized as:

i) The reliability and validity of the designed HL questionnaire are high, which are respectively verified by 'Cronbach's $\alpha = 0.807$', 'Mutual Information Score=0.803', 'KMO =0.765', and 'Bartlett Spherical Test Chi-square Value $= 2486$ ($p < 0.001$)' .
ii) According to the HL questionnaires completed by 742 Mongolians, the data set with 742 samples and 68 features is constructed to provide. data basis for the HL assessment model based on LGBM or XGB.
iii) Both LGBM-based HL assessment model and XGB-based HL assessment model can accurately predict the HL levels of respondents, and the former's accuracy is higher than that of the latter, which is 0.98347. Therefore, LGBM-based HL assessment model can be used as an intelligent tool to predict people's HL levels, which can decrease greatly manual calculations.
iv) Assessment results obtained by applying LGBM-based HL assessment model show that the HL levels of the Mongolian in Inner Mongolia, China are high. Because 92.45% of the respondents have scored above the level of Good, according to Fig,5. The percentage of men respondents who scored good and excellent is 58.2 % , which is better than that of women respondents(28.1 %). The HL scores of the urban respondents are higher than ones of the rural residents. There is a positive linear correlation between the level of HL and the educational background of the respondents.
v) The influences of each question in the HL questionnaire on the HL assessment results are quantitatively calculated by the 'feature-importance' function in LGBM. The results reveal the following points:

- It can be seen that the impact index of 'Age' is biggest, and the impact index of 'the insurance type used by the respondents' is the smallest, which are 1105 and 23, respectively.
- The influence indexes of 'Salary level', 'ability to judge relevant health information in the media', 'probability of medical attendance', 'knowledge about vaccinations and checkups', and 'ability to obtain the healthy eating information' are the second, the third, the forth, the fifth, and the sixth, respectively.
- The influence indexes of 'Gender', 'Territory', 'Education background', and 'Professional' on the HL levels are 69, 96, 69, and 71, respectively.
- The fourth dimension (health status and disease history) of the HL questionnaire is reflected by the $Column\_3, 4, 5, 7, 8, 9, 10$ in Fig. 6, where the impact index of ' health status' is the largest, which is 168. These results reveal that the forth dimension (health status and disease history) shouldn't be ignored during investigating HL assessment problems, which provides a new idea for the existing HL questionnaire design with three dimensions.

According to the above analysis, it also can be seen that all questions in the designed questionnaire are reasonable, because that they affect the HL assessment results by varying degrees.

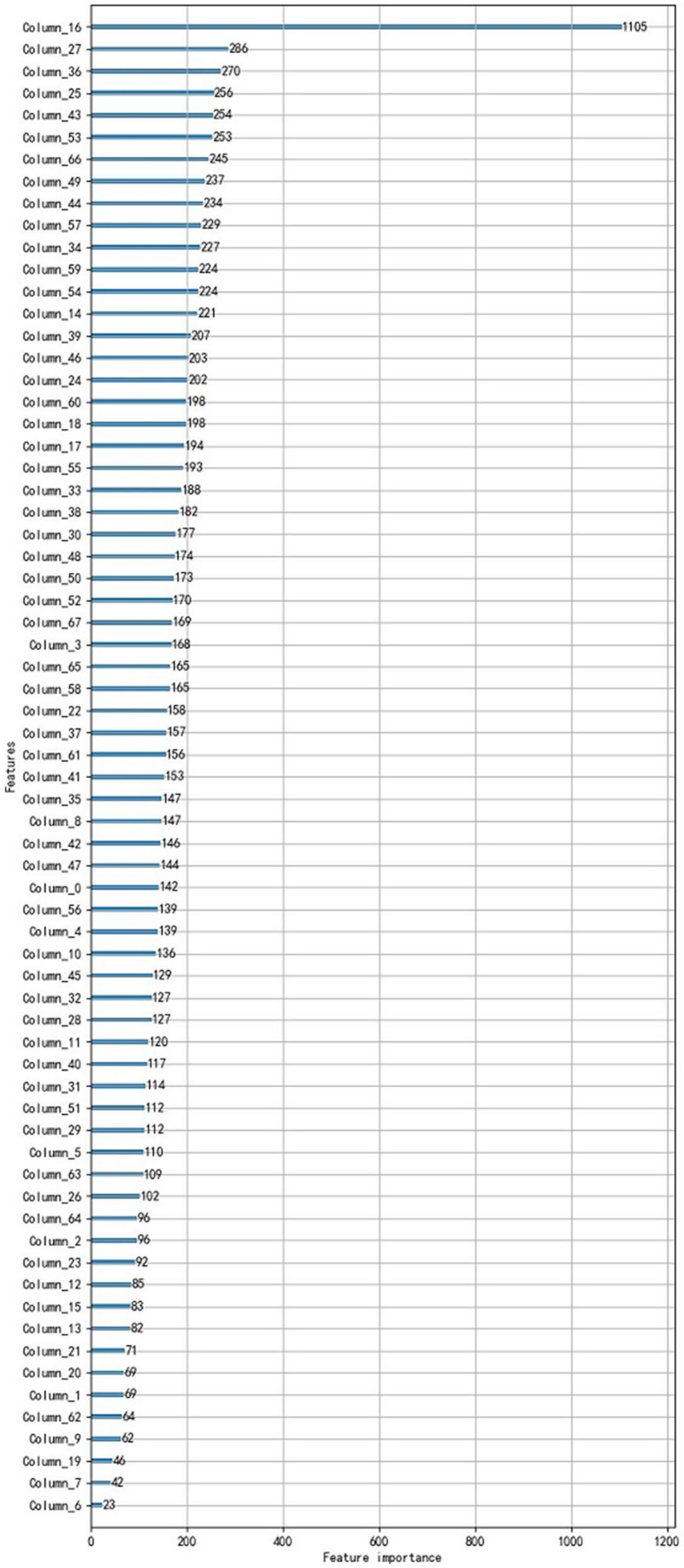

Fig. 6 The influence of 68 features on health literacy

Discussion

HL is an essential factor that affects health [26]. People with low HL have low self-management skills [27]. Poor HL can also lead to high health care costs. This paper aims at that an exclusive HL assessment questionnaire and LGBM model for Mongolians in China are presented to improve the Mongolians' HL level assessment accuracy and to find influence factors on HL by analyzing quantitatively every questions, which can provide a new idea for the HL assessment of other ethnic minorities in China or ethnic minorities in other countries.

Four dimensions are considered during the HL questionnaire's design, which are health concepts and knowledge literacy, healthy lifestyle and behavior, and healthy skills, as well as health status and disease history. It is different from the existing three dimensions methods [28–30] and five dimensions approach [31] in China, because the health status and disease history of respondents aren't considered in [28–31]. The HL questionnaire with 68 questions are designed by both improving the HLS-EU-Q47 and analyzing the characteristics in Mongolians in China. In order to verify the presented HL assessment method by a set of cross - sectional data, 742 Mongolians in Inner Mongolia of China are invited to answer the above HL questionnaire.

Based on the HL questionnaires completed by 742 Mongolians, the reliability and validity of the designed HL questionnaire are analyzed by using Cronbach's α coefficient, Mutual Information Score (MIS), KMO and Bartlett Spherical Test Chi-square Value (BSTCV). The results show that the designed HL questionnaire has the high reliability and validity, because we get Cronbach's $\alpha = 0.807$, MIS=0.803, KMO=0.765, and BSTCV=2486 ($p < 0.001$) by using our Python programs. The MIS method is better than Pearson correlation coefficient approach [32], because the latter can only handel linear correlations, however, the former can not only deal with linear correlation but also nonlinear correlation.

A data set with 742 samples is constructed, where each sample has 68 features and 1 target. 68 features correspond to 68 questions in the HL questionnaire, and 1 target corresponds to the HL score that each respondent obtained by answering the questionnaire. Based on this data set, the XGB and LGBM regression models to predict HL are constructed, respectively. 80% samples in the above data set are designed as training samples, and others are looked as testing samples. The XGB and LGBM regression models are trained by 594 (80%) samples, respectively. Then the XGB and LGBM regression models are tested by 148 (20%) samples, respectively. The $R^2(0 < R^2 \leq 1)$ index is chosen as an evaluation accuracy index. The large $R^2(0 < R^2 \leq 1)$ means the high assessment accuracy. The results show that R^2 index and the absolute error by using LGBM regression model are 0.98347 and 11, respectively, which are better than ones by applying XGB. It can be seen that the HL assessment model based on LGBM can achieve the assessment results with high accuracy.

In addition, the existing correlation analysis methods, such as Covariance method, Pearson correlation coefficient, and MIS approach, can only give quantitative results for analyzing the correlation problem among questions of questionnaires. This does not meet the growing demand for HL assessments with high-precision. Therefore, we quantitatively analyze the influence of each question in the questionnaire on the HL assessment results by using the feature-importance function in the HL assessment model based on LGBM. The quantitative results for correlation analysis among all questions are given in Fig. 6. It can be seen that the biggest impact factor is 1105, and the smallest impact factor is 23. The age has the highest influence on the HL level. It shows there is a strong correlation between age and HL levels, which is consistent with other studies [28–31, 33]. For example, Japanese HL survey [33] concluded that the HL level for Japanese increased with age; The HL survey in European countries and Turkey demonstrated that older people tended to have lower HL [33]. The impact index of the salary level of the respondents ($Column\_27$) is 286, which is the second, but it is much smaller than one of age. This result is consistent with the conclusions from [28–30]. The impact index of the ability of the interviewees to judge relevant health information in the media ($Column\_36$) is 270, which is the third. The impact indexes of the probability of medical attendance ($Column\_25$), the knowing about vaccinations and checkups ()$Column\_43$, and the obtaining healthy eating information($Column\_53$) are the forth, the fifth, and the sixth, which are 256,254, and 253, respectively. These analysis aren't found in the existing results. The influence of Gender ($Column\_1$) on the HL level is 69. The scores of the respondents show that Men's HL is higher than Women's HL, which is consistent with ones in [29, 34], but the quantification of influencing factors wasn't investigated in [29, 34]. The impact indexes of the Territory ($Column\_2$), Education background ($Column\_20$), and Professional ($Column\_21$) are 96, 69, and 71, respectively. And the scores of the respondents show that the HL levels of respondents living in cities are higher than ones of the residents in villages; there is a positive linear correlation between the level of HL and the educational background of the respondents. These results for Territory and Education background are consistent with ones in [29]. The fourth dimension (health status and disease history) of the HL questionnaire is reflected by the $Column\_3, 4, 5, 7, 8, 9, 10$ in Fig. 6, where the impact index of the health status (*Column*_3) is the largest, which is 168. However, they aren't considered in [28–30]. The impact indexes of other questions aren't addressed individually, which can be found in Fig. 6. It is worth mentioning that the least influence question on the final HL assessment result is the insurance type (*Column*_6), and its value is 23. However, this factor isn't investigated in other papers.

From Fig. 6 and the above discussion, it can be seen that the designed questionnaire is reasonable, because there are no the features that do not contribute to the health literacy assessment. It is worth mentioning that the HL assessment LGBM model and the quantitative analysis method for each question are suitable for the HL assessment for anyone else.

Conclusions

The presented HL questionnaire with 68 questions has the high reliability and validity, which are verified by using Cronbach's α, MIS, KMO, and BSTCV theories. The HL level assessment model based on LGBM can assess accurately the HL levels of Mongolians in China. The impact of each question in the questionnaire on the final assessment results can be quantified by using the 'feature-importance' function in LGBM model, which is better than the existing qualitative analysis methods. It is worth mentioning that The HL level assessment model based on LGBM and the quantitative calculation based on 'feature-importance' method for the influence index of each question on the final assessment results can also be applied to other assessment studies based on scales or data sets.

Acknowledgements

The authors wish to thank the anonymous reviewers and the Area Editor for their constructive comments and helpful suggestions.

Authors' contributions

Yan Hong and Xiaoda Zhang contributed equally to this work. The author(s) read and approved the final manuscript.

Funding

This research was supported by Doctoral Research Foundation of Inner Mongolia University for Nationalities (Grant No. BS611) and Mongolian Medicine Standardization Research International Cooperation Science and Technology Innovation Project (Grant No. MDK2020017).

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to the respondents' privacy concerns, but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All respondents gave their written informed consent to take part in the study. Research ethical permission was obtained from the Ethics Committee of Mongolia National University of Medical Sciences (Decision numb2018/3-16) and Ethics Committee of Affiliated Hospital of the Inner Mongolia University for the Nationalities (Decision number MM-LL-2018-10-13-01). All methods were performed in accordance with the relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Nursing, Inner Mongolia Minzu University, 028000 Tongliao, China. [2]Micron Intelligent Manufacturing Systems Science and Technology (Beijing) Co., Ltd, 100086 Beijing, China.

Received: 23 February 2022 Accepted: 19 October 2022
Published online: 05 November 2022

References

1. Sϕrensen K, Van DBS, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health. 2012;12:1-5.
2. Bakker MM, Putrik P, Rademakers J, et al. Addressing health literacy needs in rheumatology: which patient health literacy profiles need the attention of health professionals? Arthritis Care Res. 2021;73:100–9.
3. Naccarella L, Horwood J. Public libraries as health literate multi-purpose workspaces for improving health literacy. Health Promot J Aust. 2021;32(1):29–32.
4. Vaillancourt R, D Cameron J. Health literacy for children and families. Br J Clin Pharmacol. 2021;21(6):1-9.
5. Guo S, Davis E, Armstrong R, Yu X, Naccarella L. A pilot study of adolescent health literacy research in Melbourne: Implementation and reflections. Health Promot J Aust. 2021;32(1):128–34.
6. Shichen Z, Fangbiao T, Xiaoyan W, Shuman T, Jun F. Low health literacy and psychological symptoms potentially increase the risks of non-suicidal self-injury in Chinese middle school students. BMC Psychiatry. 2016;16(1):327–35.
7. Adnan A, Ahmed U, Nasir MK, Faria IM. Adolescent health literacy: factors effecting usage and expertise of digital health literacy among universities students in Pakistan. BMC Public Health. 2021;21(1):107–12.
8. Kemp E, Trigg J, Beatty L, Christensen C, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Aust. 2021;32:104–14.
9. Rostamzadeh A, Stapels J, Genske A, et al. Health literacy in individuals at risk for alzheimers dementia: a systematic review. J Prev Alzheimers Dis. 2020;7(1):47–55.
10. Jie Z, Yangzong D, Kezhong A, et al. Association between echinococcosis-specific health literacy and behavioural intention to prevent echinococcosis among herdsmen on the Tibet Plateau in China: a cross-sectional study. BMC Infect Dis. 2021;21(1):101–10.
11. Adrian F, Vanessa L, Vladimir K. Mental health literacy and borderline personality disorder (BPD): what do the public make of those with BPD? Soc Psychiatry Psychiatr Epidemiol. 2015;50(1):317–24.
12. Di S, Jiangbo L, Yong W, et al. Association between health literacy and hypertension management in a Chinese community: a retrospective cohort study. Intern Emerg Med. 2017;12(1):765–76.
13. Lefan L, Xujun Q, Zhuo C, Tianfeng H. Health literacy and its effect on chronic disease prevention: evidence from China's data'. BMC Public Health. 2020;20(1):690–703.
14. Madison DA, Sharon SM, Susan AER, et al. Health literacy within a diverse community-based cohort: the multiethnic study of atherosclerosis. J Immigr Minor Health. 2021;23(1):659–67.
15. Jessie C, Huaping W, Adam WA, et al. Health literacy, processing capacity, illness knowledge, and actionable memory for medication taking in type 2 diabetes: cross-sectional analysis. J Gen Intern Med. 2020;36(7):1921–7.
16. Mokhtar IA, Majid S, Foo S. Using information technology to improve health information literacy in singapore- an exploratory study. In: ITI 4th International Conference on Information and Communications Technology, Cairo, Egypt. 2006. p. 1–2.

17. Thapa-Chhetry B, Keck T. A chrome app for improving reading comprehension of health information online for individuals with low health literacy. In: IEEE/ACM 1st International Workshop on Software Engineering for Healthcare (SEH), Montreal, QC, Canada. 2019. p. 57–64.
18. Zhang X, Li Y. Health information literacy: conceptual discrimination and related research progress. J Lit Data. 2020;2(1):78–88.
19. Ateeq A, Alkadi K. Assessing the quality of online health information portals (languages other than arabic or english) in asian countries in meeting health on the net standards. In: International Conference on Informatics, Health, Technology (ICIHT), Riyadh, Saudi Arabia. 2017. p. 1–8.
20. Ojha N, Kumar A. A comparison based breast cancer high microscopy image classification using pre-trained models. In: IEEE Students Conference on Engineering and Systems (SCES), Prayagraj, India. 2020. p. 1–6.
21. Mandal S, Daivajna V, Rajagopalan V. Machine learning based system for automatic detection of leukemia cancer cell. In: IEEE 16th India Council International Conference (INDICON), Rajkot, India. 2020. p. 1–4.
22. Liu P, Fei S. A model to predict cancer comorbid patients survivability based on improved Chi2 model. In: 39th Chinese Control Conference (CCC), Shenyang, China. 2020. p. 3206–10.
23. Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, Podder E. Breast cancer risk prediction using XGBoost and random forest algorithm. In: 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India. 2020. p. 1–41.
24. Turki T. An empirical study of machine learning algorithms for cancer identification. In: 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China. 2018. p. 1–5.
25. Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Trans Biomed Eng. 2021;68(1):148–60.
26. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav. 2007;31(1):19–26.
27. Schillinger D, Grumbach K, Piette J, et al. Association of health literacy with diabetes outcomes. JAMA. 2002;288(1):475–82.
28. Nie X, Li Y, Li L. Statistic analysis of 2012 Chinese residents health literacy monitoring. Chin J Health Educ. 2014;30(2):178–81.
29. Gao M. Survy and analysis on health literacy among residents in Jilin Province, Master's Thesis. China: Jilin University; 2011. p. 1–79.
30. Zhang X, Menghe G, Ju W, Qi Q, Zhang Z, Liu X, Li H. Analysis of the health literacy level and its influence factors among residents Inner Mongolia Autonomous Region. Chin J Health Educ. 2016;32(12):1059–62.
31. Liu C. Conceptual model and measurement of Chinese health literacy. Doctoral Dissertation, Huazhong University of Science and Technology, China. 2018. p. 1–276.
32. Zhang Y, Li M, Jiang H, Shi H, Xu B, Atkins S, Qian X. Development and validation of a Chinese parental health literacy questionnaire for caregivers of children 0 to 3 years old. BMC Pediatr. 2019;19(1):293–302.
33. Nakayama K, Osaka W, Togari T, et al. Comprehensive health literacy in Japan is lower than in Europe: a validated Japanese-language assessment of health literacy. BMC Public Health. 2015;15(1):505.
34. Duong VT, et al. Health literacy in Taiwan: a population-based study. Asia Pac J Public Health. 2015;27(1):871–80.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.