Profiling Wellness Product Consumers in Short-Video Livestreaming Commerce: A Big Data Cluster Analysis

CAO Hongyu[1], GAN Weixiang[1,2,*]

1 Langfang Zhenguigu Technology Co., Ltd., Beifeng Road, Guangyang District, Langfang, Hebei 065000, China
2 Lincoln University College, Petaling Jaya, Selangor Darul Ehsan 47301, Malaysia

Abstract: Drawing on 186,442 users of wellness-category livestreaming channels on a leading short-video platform, this study merges 90-day transaction records, viewing and interaction logs, and 4.126 million comments into a feature set that spans two modalities, behavior and text. An LDA topic model condenses the comments into three demand-topic intensities (tonic diet, fitness shaping, sleep and mood), the entropy weight method assigns feature weights, and K-means++ performs the clustering. The average silhouette coefficient and the variance ratio criterion both peak at $k = 5$, yielding five segments: senior tonic (24.1%), white-collar functional (22.0%), high-engagement repeat (8.9%), promotion-sensitive (18.0%), and lurker (26.9%). The high-engagement repeat segment delivers 40.8% of gross merchandise volume with 8.9% of users. In a negative binomial regression of purchase counts over a 30-day holdout window, every segment dummy is significant at the 1% level, and purchase intensity in the high-engagement repeat segment is 23.9 times that of the lurker segment. Clustering on transaction features alone blurs the white-collar functional and senior tonic segments and drags the adjusted Rand index down to 0.624. The demand information carried by comment text proves indispensable to segmentation.

Keywords: Livestreaming commerce; Consumer profiling; Cluster analysis; Latent Dirichlet allocation; Negative binomial regression; Wellness products

I. Introduction

The short-video sector reaches 1.068 billion users in China, 95.1% of the online population, and livestreaming reaches 833 million (CNNIC, 2025a, 2025b). Wellness consumption is growing faster still inside livestream channels: between July 2022 and June 2023, the paid GMV, buyer count, and number of active items of health food on Douyin's e-commerce arm each grew by more than 200% year on year, and GMV of the tonic and supplement category rose 57% in the first half of 2023 (Douyin E-commerce & Yicai Business School, 2023). The wider health industry generated CNY 8.0 trillion in revenue in 2021 (iiMedia Research, 2022). Traffic of this scale pours into the same livestreaming channels, where viewers in their fifties shopping for tonics sit alongside young consumers with irregular sleep patterns seeking to improve their sleep. A merchant who serves all these groups with a single product lineup and a uniform sales script suffers significant conversion losses.

The idea of dividing a market by differences

* *Corresponding author*

in demand goes back to Smith (1956). Wedel and Kamakura (2000) sorted segmentation methods into four families, geo-demographic, lifestyle, response-based, and conjoint-based, and clustering algorithms have been the standard workhorse ever since. In livestreaming commerce, Meng et al. (2020) used structural equation modeling to show that the credibility, expertise, and interactivity of sales-oriented streamers have the strongest effect on arousing purchase intention; Xie et al. (2019) proposed that social presence steers online conformity consumption through information trust and evaluation apprehension; survey evidence from Wongkitrungrueng and Assarut (2020) traces utilitarian and hedonic value to customer engagement through product trust and then seller trust. These studies map how purchase intention forms during a livestream, yet their samples rarely reach beyond questionnaires and experiments. To the best of our knowledge, no published work, in Chinese or in English, has segmented wellness buyers with platform-scale behavioral data or written the health demands voiced in comments into the segmentation features.

What makes the wellness category challenging is demand heterogeneity. The same box of ejiao cake may be ordered as a traditional tonic, as a fitness supplement, or as a sleep aid, and the transaction records show no difference among the three. This paper responds with three design choices. An LDA topic model (Blei et al., 2003) condenses 4.126 million comments into three demand-topic intensities, which join the transaction and viewing features in a dual-modality feature set. Feature weights come from the entropy weight method rather than equal weighting, so the data, not the analyst, decide how much each feature counts. Clustering draws only on a 90-day observation window, and a 30-day holdout window is reserved for a negative binomial regression that tests whether the segments predict subsequent purchasing.

II. Data and Features

A. Data

The data cover users of wellness-category livestreaming channels (tonic foods, functional foods, and wellness devices) on a leading short-video platform, anonymized by the platform prior to data release. The observation window runs from March 1 to May 29, 2025, a span of 90 days; the holdout window covers the following 30 days. Every user who watched at least three sessions during the observation window enters the sample, 186,442 in total, spread across a transaction table, a behavioral log table, and a comment table. Identity fields retain only age, gender, and city tier, and all analysis is run at the aggregate level.

B. Dual-Modality Features

On the behavioral side, the RFM framework is extended to seven indicators: recency (R), frequency (F), average order value (M), daily watch time, average dwell time per session, interaction rate, and coupon redemption rate. On the textual side, the comments are tokenized with jieba, stripped of stop words, and fed to an LDA model; a search over 2 to 8 topics based on perplexity and topic coherence identifies three as optimal. Judged by their top words, the topics are labeled traditional tonics (ejiao, bird’s nest, qi and blood), body shaping (protein, meal replacement, body fat), and sleep and mood (melatonin, insomnia, anxiety). A user’s demand-topic intensity is the mean posterior topic probability across that user’s comments. Table 1 reports variable definitions and descriptive statistics. The variance of holdout-window purchase counts (2.53) is 2.6 times the mean (0.98), a clear sign of overdispersion.

Table 1. Variable definitions and descriptive statistics (N = 186,442)

Variable	Definition and Unit	Mean	SD
R	Days since last purchase, 90 if none (days)	34.71	34.49
F	Purchase count in the observation window	3.51	3.17
M	Average order value (CNY)	158.31	132.19
Watch time	Daily watch time of wellness content (min)	17.27	7.88
Stay per session	Average stay per livestream session (min)	7.17	3.20
Interaction rate	Interactions per session watched (%)	2.57	1.68
Coupon redemption	Coupons redeemed over coupons claimed (%)	33.63	23.47
Tonic diet	Mean posterior probability of topic 1	0.44	0.22
Fitness shaping	Mean posterior probability of topic 2	0.29	0.17
Sleep and mood	Mean posterior probability of topic 3	0.28	0.12
Age	User age (years)	41.79	12.81
Holdout purchases	Purchase count in the 30-day holdout window	0.98	1.59

C. Entropy Weighting

Each feature j is first min-max standardized into; its information entropy and weight follow Eqs. (1) and (2):

$$e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij}\ln p_{ij}, \qquad p_{ij} = x_{ij}^{'}/\sum_{i=1}^{n} x_{ij}^{'} \tag{1}$$

$$w_j = \frac{1-e_j}{\sum_{m=1}^{J}(1-e_m)} \tag{2}$$

In Eqs. (1) and (2), $n = 186{,}442$ is the sample size, $J = 10$ the number of features, and p_{ij} the share of observation i in feature j. The weights depend only on how dispersed each feature is; no subjective weighting is imposed. Recency, frequency, and order value take the three largest weights (0.196, 0.186, and 0.174), coupon redemption receives 0.105, and the three demand topics together receive 0.170, about one sixth of the total. Clustering operates on the standardized features z_{ij} scaled by $\sqrt{Jw_j}$, which amounts to a weighted Euclidean metric with the entropy weights on the diagonal.

III. Models

A. Entropy-Weighted K-means++

The K-means algorithm (MacQueen, 1967) minimizes the within-cluster sum of squares in Eq. (3):

$$\min_{C_1,\ldots,C_k} \sum_{t=1}^{k}\sum_{i\in C_t} \|\sqrt{\boldsymbol{w}} \odot \boldsymbol{z}_i - \boldsymbol{\mu}_t\|^2 \tag{3}$$

In Eq. (3), $\boldsymbol{z}_i$ is the standardized feature vector, the entropy weight vector, $\boldsymbol{w}$ the element-wise product, and C_t and μ_t the t-th cluster and its centroid. Initial centroids follow the D^2-weighted random seeding of k-means++ (Arthur & Vassilvitskii, 2007), whose expected approximation ratio is $\Theta(\log k)$. The number of clusters is judged by three indicators read together: the SSE elbow, the average silhouette coefficient (Rousseeuw, 1987), and the variance ratio criterion (Caliński & Harabasz, 1974). The silhouette value is defined in Eq. (4):

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{4}$$

In Eq. (4), $a(i)$ is the mean distance from sample i to the other members of its own cluster and $b(i)$ the mean distance to the nearest foreign cluster. Pairwise distances over the full sample are too costly, so the latter two indicators are computed on a random subsample of 20,000.

B. Negative Binomial Regression

Once a count outcome is overdispersed, the Poisson model understates standard errors (Cameron & Trivedi, 2013), so the conditional moments of holdout-window purchase counts take the negative binomial form:

$$E(y_i|\boldsymbol{x}_i) = \mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}),$$
$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \mu_i + \alpha\mu_i^2 \quad (5)$$

In Eq. (5), $\boldsymbol{x}_i$ contains four segment dummies (the lurker segment, which converts least, serves as the base group), age, gender, city tier, and daily watch time; $\boldsymbol{\beta}$ is the coefficient vector and α the overdispersion parameter. Exponentiated coefficients are incidence rate ratios (IRR).

IV. Results

A. Choosing the Number of Clusters

Figure 1 plots SSE and the average silhouette coefficient for k from 2 to 8. SSE bends sharply at k = 5, falling 26.9% from k = 4 and less than 6% per step thereafter; the silhouette coefficient reaches its global maximum of 0.466 at k = 5, and the variance ratio criterion peaks there as well (18,104). Three pieces of evidence converge on the same value, so k is set to 5.

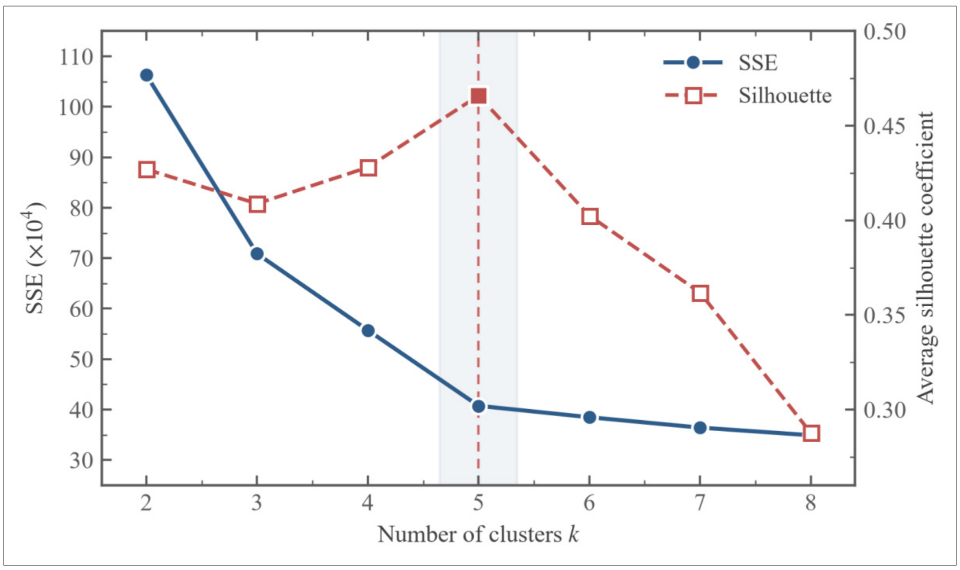

Figure 1 SSE and average silhouette coefficient across cluster numbers

B. Five Consumer Segments

Table 2 reports cluster centroids and demographic structure, and Figure 2 lays the five profiles side by side as a heat map. The segment profiles are clearly differentiated. The senior tonic segment (C1) averages 57.8 years of age and is 62.0% female; its tonic-diet intensity reaches 69.2%, it repurchases every 9 days, and its order value averages CNY 267.9: need-driven buyers in the classic sense. The white-collar functional segment (C2) is far younger (30.3 on average), sits 70.0% in tier-1 and tier-2 cities, and scores 51.2% on fitness shaping; it watches little and buys quickly. The high-engagement repeat segment (C3) holds a mere 8.9% of users yet delivers 40.8% of GMV on 9.9 purchases, a CNY 411.6 order value, and a 6.8% interaction rate. The promotion-sensitive segment (C4) is characterized primarily by its coupon redemption rate of 74.1%; its order value of CNY 78.0 is the lowest among the purchasing segments. The lurker segment (C5) watches the most (26.1 minutes a day) and buys the least (0.2 purchases), with the three demand topics split evenly at about one third each: indicating broad interest that has not yet focused on a specific category.

Table 2. Cluster centroids and profile characteristics of the five segments

Indicator	C1 Senior Tonic	C2 White-Collar Functional	C3 High-Engagement Repeat	C4 Promotion-Sensitive	C5 Lurker
Segment share (%)	24.1	22.0	8.9	18.0	26.9
Recency (days)	9.0	18.2	5.1	24.9	87.6
Frequency (count)	5.2	3.5	9.9	3.1	0.2
Order value (CNY)	267.9	185.2	411.6	78.0	8.0
Watch time (min/day)	14.1	11.6	22.0	13.0	26.1
Stay per session (min)	6.8	5.3	11.2	4.6	9.4
Interaction rate (%)	3.1	2.2	6.8	2.0	1.4
Coupon redemption (%)	30.9	36.7	29.0	74.1	8.1
Tonic diet (%)	69.2	15.5	51.9	54.6	33.4
Fitness shaping (%)	13.5	51.2	14.7	22.7	33.2
Sleep and mood (%)	17.4	33.3	33.3	22.7	33.4
Mean age (years)	57.8	30.3	46.2	42.0	35.2
Female (%)	62.0	57.7	59.7	65.9	49.1
Tier-1/2 city (%)	38.3	70.0	54.4	33.0	51.9
GMV share (%)	37.5	16.3	40.8	4.9	0.5

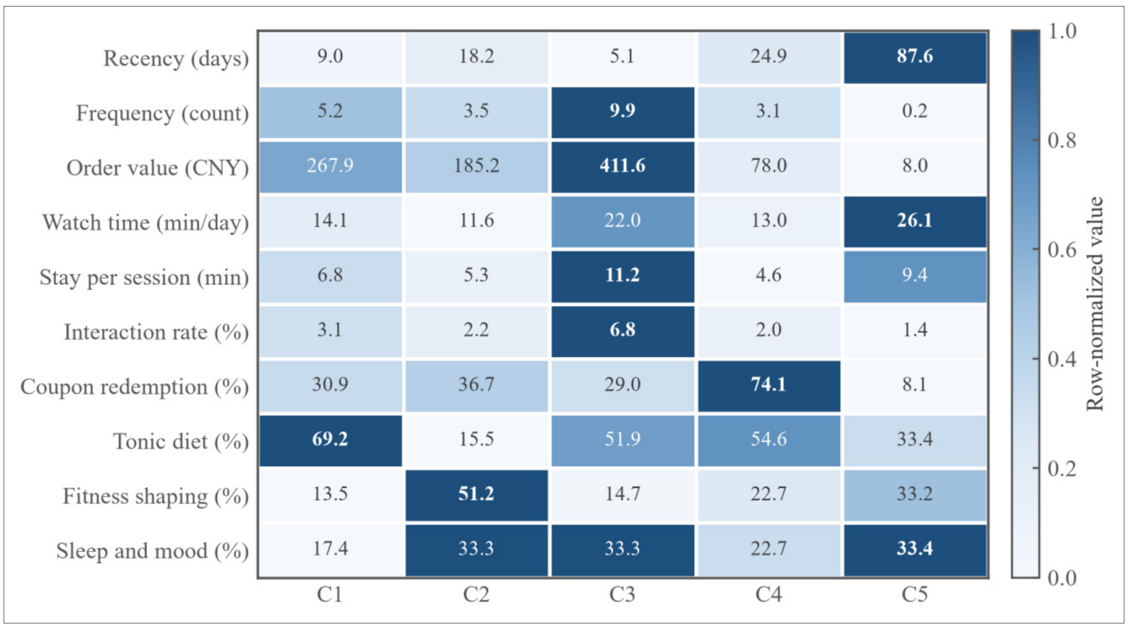

Figure 2. Heat map of cluster centroids for the five segments

The contribution of the text features can be quantified. Re-clustering on R, F, and M alone (k = 5) drags the adjusted Rand index against the dual-modality benchmark down to 0.624: one cluster mixes 60.6% C2 members with 39.3% C1 members, while another lumps together 75.2% of C1, 16.5% of C2, and 7.8% of C3. The transaction ranges of these groups overlap heavily; coupon redemption and demand-topic intensity are what differentiate them.

C. Holdout Validation

Figure 3 shows 30-day holdout conversion by segment, with incidence rate ratios relative to the base group on the right axis. Conversion rates reach 84.3% (C3), 68.3% (C1), 48.6% (C2), 43.7% (C4), and 12.2% (C5).

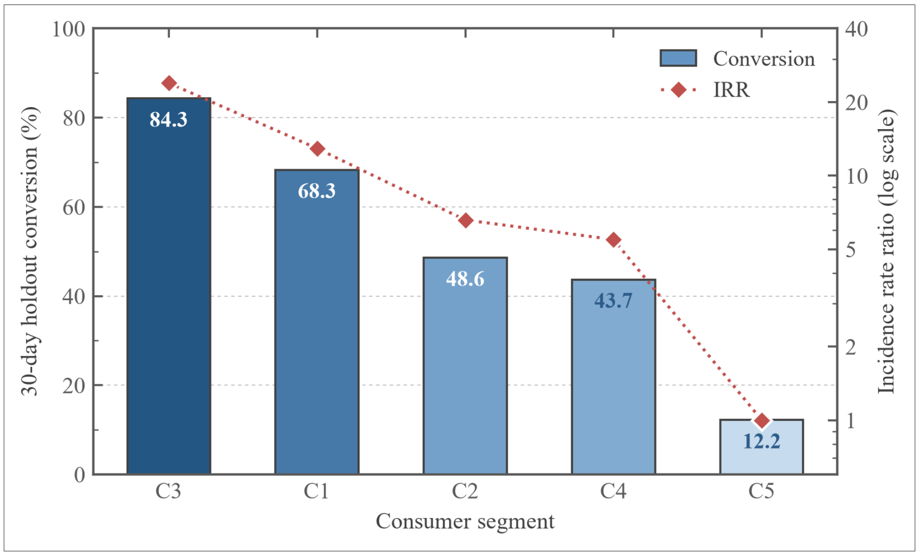

Figure 3. Holdout conversion and incidence rate ratios of the five segments

The negative binomial estimates in Table 3 are consistent with the ranking. Against C5, the IRR is 23.94 for C3, then 12.93, 6.59, and 5.48 for C1, C2, and C4, with all four dummies significant at the 1% level; ten extra minutes of daily watch time lifts purchase counts by 7.2%. Notably, demographics matter little once segment membership is controlled: ten additional years of age cut purchase counts by a mere 1.2%, and city tier loses significance altogether, so the surface differences across age and region are absorbed almost entirely by segment structure. The estimate is significant at the 1% level, confirming overdispersion and favoring the negative binomial over the Poisson; pseudo R-squared is 0.131 and the likelihood ratio test gives $p < 0.001$.

Table 3. Negative binomial regression of holdout-window purchase counts (N = 186,442)

Variable	Coefficient	SE	z	IRR
C1 Senior tonic	2.560***	0.019	135.98	12.93
C2 White-collar functional	1.885***	0.018	107.01	6.59
C3 High-engagement repeat	3.175***	0.016	204.79	23.94
C4 Promotion-sensitive	1.701***	0.018	96.22	5.48
Age/10	-0.012***	0.004	-2.72	0.99
Female	0.011*	0.006	1.75	1.01
Tier-1/2 city	0.004	0.006	0.69	1.00
Watch time/10	0.070***	0.007	10.20	1.07
Constant	-2.155***	0.027	-78.91	
α	0.511***	0.006	90.74	

D. Robustness

Two checks examine stability. Replacing entropy weights with equal weights ($k = 5$) returns an adjusted Rand index of 0.968 against the benchmark, while the silhouette coefficient decreases from 0.466 to 0.414, so the entropy-weighted partition is tighter within clusters. Re-clustering a random half of the sample reproduces the full-sample partition with an adjusted Rand index of 1.000.

E. Conclusions and Operational Responses

The evidence can be summarized in three statements. Comment text is not merely supplementary to transaction data: strip out the demand topics and the adjusted Rand index falls to 0.624, with white-collar functional and senior tonic buyers conflated. T The market is highly polarized: 26.9% of the audience contributes 0.5% of GMV while 8.9% takes 40.8%; RFM carries a combined entropy weight of 0.556, which leaves nearly half the information in the coupon, interaction, and demand-topic features, so a transactions-only view captures only part of the underlying demand structure. Segment labels predict holdout purchasing across a 5.5-to-23.9-fold IRR range and absorb the explanatory share of age and city tier; they sit closer to the purchase decision than demographic labels do.

Each segment requires a tailored operational strategy. Senior tonic buyers respond to fit and trust: stock traditional tonics, emphasize ingredient provenance and dosage in the presentation, and avoid deep discounts, since redemption in this group is only 30.9% to begin with. White-collar functional buyers want efficiency: protein, meal replacements, sleep gummies, broadcasts between 21:00 and 23:00, and ingredient tables up front. High-engagement repeat buyers merit membership schemes and early access to new items; a holdout conversion of 84.3% can carry frequent contact. Promotion-sensitive buyers can be served with volume deals and group buying, but coupon depth and frequency need tightening, or the 74.1% redemption rate will erode profit margins and reinforce a dependency on coupons for purchase decisions. Passive viewers should not be targeted with intensive promotions, because 12.2% conversion cannot repay the spend; short-video content and topic testing should first clarify their demand preferences before directing them to livestream sessions.

Three limitations deserve mention. The sample is drawn from a single platform and a single category, and the generalizability of the segment structure requires external validation. The selection of three topics represents a dimensionality reduction choice; categories with a finer demand spectrum may warrant more. The negative binomial model omits supply-side variables such as price promotions and broadcast schedules, so the segment effects inevitably reflect some influence of targeting strategy. Multi-platform data and supply-side instruments are the next step.

References

[1] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding[C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[3] Caliński T, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics, 1974, 3(1): 1-27.

[4] Cameron A C, Trivedi P K. Regression analysis of count data[M]. 2nd ed. Cambridge: Cambridge University Press, 2013.

[5] CNNIC. The 55th statistical report on China's internet development[R]. Beijing: China Internet Network Information Center, 2025.

[6] CNNIC. The 56th statistical report on China's internet development[R]. Beijing: China Internet

Network Information Center, 2025.

[7] Douyin E-commerce, Yicai Business School. White paper on health food consumption trends[R]. 2023.

[8] iiMedia Research. Big data and decision analysis report on the global and Chinese health industry, 2022-2023[R]. 2022.

[9] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297.

[10] Meng L, Liu F J, Chen S Y, et al. Can I arouse you: the influence of information source characteristics of different types of live streamers on consumers' purchase intention[J]. Nankai Business Review, 2020, 23(1): 131-143.

[11] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.

[12] Smith W R. Product differentiation and market segmentation as alternative marketing strategies[J]. Journal of Marketing, 1956, 21(1): 3-8.

[13] Wedel M, Kamakura W A. Market segmentation: conceptual and methodological foundations[M]. 2nd ed. Boston: Kluwer Academic Publishers, 2000.

[14] Wongkitrungrueng A, Assarut N. The role of live streaming in building consumer trust and engagement with social commerce sellers[J]. Journal of Business Research, 2020, 117: 543-556.

[15] Xie Y, Li C Q, Gao P, et al. The influence and mechanism of social presence on online herd consumption in livestream marketing: a behavioral and neurophysiological perspective[J]. Advances in Psychological Science, 2019, 27(6): 990-1004. DOI:10.3724/SP.J.1042.2019.00990.