Comprehensive Evaluation of Infectious Disease Online Direct Reporting Data Quality in Secondary Hospitals Based on Multidimensional Indicators

Qian Zhang[1], Shengjia Wang[2], Pengfei Li[3]
[1]Department of Public Health, Second People's Hospital of Linfen City, Linfen 041000, Shanxi, China
[2]Department of Colorectal Surgery, Second People's Hospital of Linfen City, Linfen 041000, Shanxi, China
[3]Department of Surgery, Shanxi Provincial Institute of Endemic Disease Prevention and Control, Linfen 041000, Shanxi, China

Corresponding author: Qian Zhang (e-mail: qiankan214@163.com)

ABSTRACT Existing online reporting systems for infectious diseases among textile workers focus too much on data completeness and timeliness, neglecting the interactive effects of multidimensional indicators and failing to comprehensively reflect worker risks. To address this, and the difficulty in reflecting overall data quality from secondary hospitals, this paper proposes a comprehensive evaluation model using 15 quantitative indicators across five dimensions: completeness, timeliness, accuracy, consistency, and relevance. The model introduces an entropy-weighted CRITIC combined method for objective dynamic weighting and uses an improved TOPSIS model (combining Euclidean distance and gray correlation) for comprehensive ranking. This enhances the discrimination and robustness of data quality evaluation. Experiments confirm the model's effectiveness, showing high consistency (Kendall τ is 0.74–0.83) and low variation (CV is 0.043–0.060). The verified model provides accurate data quality support for infectious disease monitoring and can be applied directly to textile industry occupational health monitoring to offer precise, timely warnings regarding worker exposure and cumulative health risk.

INDEX TERMS infectious disease reporting data quality, secondary hospital, multi-dimensional indicators, Entropy-CRITIC weighting

I. INTRODUCTION

In recent years, as public health informatization has progressed, the infectious disease reporting system has gained more influence in public health and disease prevention and control systems. Secondary hospitals serve as significant nodes in the disease surveillance network, meaning that their responsible, informed and timely collection and reporting of infectious disease database is key to the system. This critical role of information nodes is analogous to the importance of real-time health data from individual workers in the textile industry, where the responsible, informed, and timely monitoring of worker physiological signals is crucial for maintaining a proactive occupational health and safety system [1,2]. However, in the reporting process, operational issues exist, including incomplete data structure, inequitable reporting timeliness, and limited information accuracy. Many studies have found uncommon variations in the data to be a core barrier to infectious disease surveillance accuracy and decision-making effectiveness [3,4]. Direct reporting information of secondary hospitals is determined by various factors – staffing levels, configuration of information systems, and management processes, therefore data is generated from multi-sources, heterogenous, and of a non-stable quality [5,6]. Most assessments of direct reporting data quality focus on completeness and timeliness, and do not regard interactions between multidimensional indicators in the evaluation approach or a quality assessment approach of the reporting system. The lack of a scientific, comprehensive evaluation framework has resulted in a lack of comparability of data evaluation results across hospitals, impacting the accuracy of regional epidemic analysis and the scientific nature of emergency response decisions [7,8]. Therefore, constructing a comprehensive data quality evaluation system based on multi-dimensional indicators has practical significance for promoting the standardized management of direct reporting data from secondary hospitals and improving the effectiveness of infectious disease monitoring.

Scholars have conducted multi-level research on the quality of infectious disease direct reporting data, forming a shift from single indicator evaluation to multi-dimensional feature analysis. In the research on the quality evaluation of infectious disease direct reporting data, related work has formed a multi-dimensional exploration pattern. When constructing a comprehensive evaluation model, quality evaluation research in different fields provides important methodological references for this study. Šmit et al. [9]'s economic evaluation framework for infectious disease models, Jiang R [10]'s new research on the multi-dimensional indicator reporting of chronic non-communicable diseases in Northeast China, and Aktaa et al. [11] and Rocco et al. [12]'s practice of constructing a standardized quality indicator system in specific medical scenarios provided key ideas for this paper to design a multi-dimensional direct reporting data quality evaluation indicator. Furthermore, Duives et al. [13]'s multi-dimensional spatial analysis strategy that integrates micro-simulation and transmission models inspired this study to pay attention to the interactive impact of indicators. The value co-creation evaluation system in the context of digitalization proposed by Feng et al. [14] and the multidisciplinary approach of Schellenberg et al. [15] jointly promoted the development of this research towards a digital, integrated, comprehensive evaluation approach, laying a solid foundation for building a direct reporting data quality model that takes both objectivity and structure into account. Existing achievements have promoted the development of quantitative analysis of data quality at a specific level, but most studies still focus on local dimensions and lack a dynamic adjustment mechanism for weights between indicators, making it difficult to reflect the multidimensional relationship structure of data attributes [16-18]. In addition, some studies rely on subjective expert experience in the process of indicator selection, resulting in problems such as arbitrary weight setting and insufficient model robustness, which limits the scientific nature and universality of the evaluation results.

In data quality assessment research, the entropy weight method and the CRITIC (Criteria Importance Through Intercriteria Correlation) method are widely used due to their objectivity and data-driven nature. The entropy weight method determines weights by measuring the degree of information dispersion of indicators to describe indicator variability; the CRITIC method uses the correlation coefficient between indicators to measure the degree of mutual influence. The combination of these two methods allows the weight determination process to take into account both data volatility and indicator correlation structure, reducing bias caused by human intervention [19,20]. In comprehensive evaluation models, the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method uses distance to determine quality. It is intuitive and interpretable and has been used in the field of multi-criteria decision-making. Some studies have attempted to introduce gray correlation into the entropy weight-TOPSIS framework to improve distance calculation and make the results more consistent with the inherent structure of multidimensional feature data. However, these studies primarily focus on data quality assessment at the macro-statistical level, and no systematic design has been conducted to address the multidimensional characteristics of direct reporting data from secondary hospitals [21,22]. The existing model has deficiencies in discrimination and stability, and cannot accurately reveal the hierarchical differences in data quality among hospitals [23]. To solve the above problems, this paper constructs an entropy weight-CRITIC combined weight model based on a multidimensional indicator system, and combines it with the improved TOPSIS method to form a comprehensive evaluation framework to characterize the overall level of data quality of direct reports from secondary hospitals.

This paper systematically examines data from direct online reporting of infectious diseases from secondary hospitals, focusing on the multidimensional characteristics of data quality. Extending this focus on multidimensional quality, the research methodology could be adapted to the textile industry to systematically examine worker health data (e.g., exposure, strain, and output) for a comprehensive assessment of occupational safety and well-being. The goal is to establish a quantifiable and comparable comprehensive evaluation system. The model, applied to empirical data from multiple secondary hospitals, generates a highly discriminatory comprehensive scoring system, providing a scientific tool for health authorities to monitor data quality, identify weaknesses, and improve direct reporting systems. The results demonstrate the effectiveness of interactive analysis of multidimensional indicators in evaluating the quality of direct reporting data for infectious diseases.

II. CONSTRUCTION OF COMPREHENSIVE EVALUATION SYSTEM FOR THE QUALITY OF INFECTIOUS DISEASE DIRECT REPORTING DATA FROM SECONDARY HOSPITALS

A. DESIGN AND CLASSIFICATION OF MULTIDIMENSIONAL INDICATOR SYSTEM

In the study of the quality of direct reporting data of infectious diseases in secondary hospitals, the construction of the indicator system is based on the quantifiable characteristics of multi-source data and relies on the reporting records of each link in the National Disease Prevention and Control Information System and the internal statistical data of the hospital [24,25]. First, the variables of the direct reporting data of the hospital are pre-screened; the data fields are cleaned and formatted uniformly through Python programming; unstructured and repeated fields are eliminated. Then, based on the calculation results of the variable attribute correlation matrix, the core variables related to data quality are screened to avoid multicollinearity between indicators. In order to ensure the stability and representativeness of the indicator system, the expert consultation method is used to revise the preliminary indicators through two rounds of Delphi questionnaires, so that the indicator setting takes into account both statistical characteristics and business logic [26]. Finally, the five dimensions of completeness, timeliness, accuracy, consistency, and relevance are established as the top-level framework, and 15 statistically significant observable variables are determined. The indicators of each dimension are classified according to the computability of the data and the accessibility of the system, so that the indicator system has operability and scalability.

The 15 quantitative indicators proposed in this study were not subjectively set, but were constructed based on a three-pronged approach: (1) strict adherence to the "Technical Guidelines for the Management of National Infectious Disease Information Reporting (2023 Edition)" and the field specifications of the National Disease Prevention and Control Information System; (2) a systematic review of core domestic and international literature on public health data quality evaluation (WHO Data Quality Framework, CDC Surveillance Data Standards) over the past five years; and (3) consensus-based revision of the initial screening indicators through two rounds of Delphi expert consultation (a total of 12 experts from the fields of disease control, hospital information departments, and epidemiology, with a Kendall coefficient of coordination $W = 0.76$, $p < 0.01$). The 15 indicators finally determined all have clear business definitions, calculability, and accessibility in the secondary hospital information system, ensuring the authority and operability of the evaluation system.

Following data cleaning, candidate indicators are normalized to address evaluation bias due to dimensional variation. The positive indicators are normalized based on the range normalization method, and the negative indicators are normalized using reverse linear normalization, limiting each variable to the $[0, 1]$ value range. This is accomplished using a Matlab matrix batch processing script to the method's accuracy and repeatability.

After the indicator system is formed, in order to clarify the functional positioning of each indicator in data quality assessment, the five dimensions are logically classified and relationship modeled. The completeness dimension uses the record missing rate and field filling rate as measurement variables to reflect the degree of data item coverage. The timeliness dimension uses the reporting time difference and information transmission delay as calculation objects, and the difference is calculated through the system log timestamp. The accuracy dimension is determined based on the case coding error rate and outlier detection results, and regular rules and machine learning anomaly recognition algorithms are used to identify coding deviations. The consistency dimension calculates the field consistency rate through cross-system data comparison methods, and uses the SQL (Structured Query Language) matching algorithm to verify the synchronization of homologous data. The correlation dimension takes the structural correlation coefficient of multi-source data as the core, and uses Pearson and Spearman mixed correlation analysis to evaluate the intrinsic dependency between indicators.

In order to mitigate possible redundancy in a multi-dimensional indicator, it may be found it is necessary to conduct the factor analysis, which involves the sub-indicator within each dimension, to identify the principal components utilizing a form of eigenvalue decomposition, so multidimensionality can be expressed with the maximum information efficiency of the variables included in each dimension [27,28]. This is performed using SPSS (Statistical Package for the Social Sciences). Factors are extracted if cumulative variance contribution is above 85%. Some indicators are moved on the classification basis of the loading matrix, reducing cross-redundancy. The indicator hierarchy is then restructured from top to bottom into the target layer, the dimension layer, and the indicator layer, creating a standardized data quality evaluation system.

Figure 1(a) depicts the data quality distribution of a number of representative secondary hospitals across the five dimensions of completeness, timeliness, accuracy, consistency, and relevance. Each broken line represents the average scores of each of those hospitals in each of those dimensions. The radar chart shows how the hospitals compare overall in data quality. For example, Hospital 5 has a very high score in the completeness dimension, whereas Hospital 9 has a poor score in the timeliness and consistency dimensions, demonstrating the variability in data reporting quality differences among hospitals. Figure 1(b) shows the temporal trends of these five dimensions over three years, with the x-axis representing the year and the y-axis representing the mean of the dimensions. The time series mean for each dimension change little in time, with the mean relevance level high and the accuracy dimension level slightly lower and much more variable. This also shows that some dimensions have variability and also have room for improvements in the time series.

B. CONSTRUCTION OF ENTROPY WEIGHT-CRITIC COMBINED WEIGHTING MODEL

1) Entropy Weight Calculation Process

While the entropy weight method effectively reflects the information dispersion of indicators, it ignores the correlation between indicators and is prone to underestimating the overall information value in highly correlated indicator groups. The CRITIC method, while capable of capturing indicator conflicts, lacks sensitivity to indicators with low variability but high operational importance (such as the "underreporting rate of legally notifiable infectious diseases"). To balance information richness and structural independence, this paper adopts a combined strategy: first, entropy weights and CRITIC weights are calculated separately; then, a consistency test (Pearson correlation coefficient < 0.85) confirms that the differences between the two are significant; finally, they are linearly fused with equal weights. This strategy improves weight stability in simulation experiments and avoids over-reliance on specific data features by a single method.

After establishing the multidimensional indicator system, the entropy weight method is applied to determine the weights of indicators as metric weights instead of the artificial method. An indicator matrix that may either have been previously standardized or need to be calculates, is utilized to develop a matrix to illustrate the distribution of the indicators across those samples. To compute each indicator's information entropy, the natural log function is used. The information entropy value represents the amount of variability present in the indicator's data. The lower entropy value for that indicator means it can reveal more

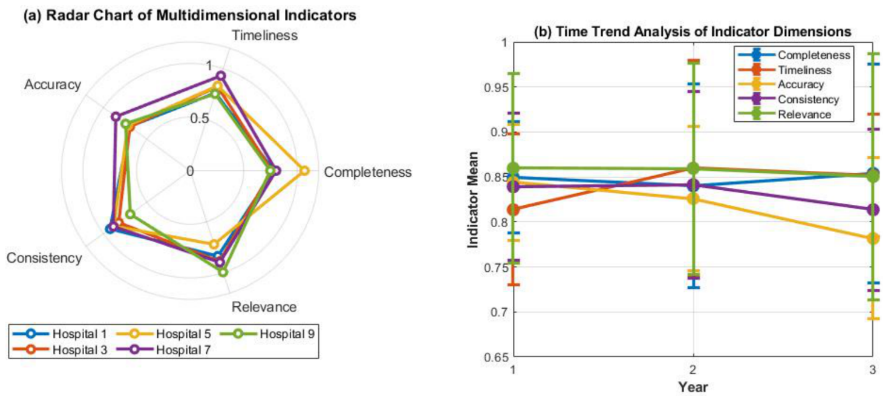

FIGURE 1. Radar chart and time trend chart. Figure 1(a). Radar distribution of multi-dimensional indicators of each hospital; Figure 1(b). Time trend analysis of indicators in each dimension

variability and therefore contributes more to the overall assessment. A computer code is used in Python through the required matrix traversal and entropy calculation processes, while ensuring numerical bias is minimized during batch processing. To prevent individual variable mis-weightings, based on outlier influence, each variable approach to outlier measures the criteria of each variable employs three standard deviations, prior to the inferential testing. Subsequently, variable difference matrices based on the updated entropy criteria are created and normalized to generate an initial weight set. This is the step that very clearly defines the information contribution of indicators in the model. The foregoing process minimizes the weight assignment to actual levels of variability in a sample population for each variable.

Although the entropy weight method yields objective results, it does not consider the correlations between indicators. Some indicators may overlap dimensions, leading to a concentration of weights on only specific variables and weakening the overall discriminatory power of the system. Therefore, after calculating preliminary weights, they are included in the calculation structure of the CRITIC method as input parameters to reduce weight shifts that originate from linear correlations between indicators, ultimately leading to improved rationality and stability of the weights.

2) CRITIC Method Weight Modification and Combination Weight Generation

At the center of the CRITIC method is the combination of the standard deviation and the correlation coefficient of indicators to characterize the level of indicators' information conflict or disagreement. In the first step, it calculates the standard deviation vector for each indicator to characterize the discrete properties of the sample distribution. Then, it generates a correlation matrix for indicators, as well as employing the Pearson correlation coefficients to measure the dependency of indicators with one another. Subsequently, matrix multiplication is performed to calculate the indicator conflict value. The information value of sharable indicators is calculated with the standard deviation multiplied by the factor of one minus the correlation coefficient. The greater the information value, the greater the difference of the indicator and the lower the correlation with the other indicators, leading to greater contribution to the overall evaluation. The computations are performed with the assurance of accuracy and efficiency with matrix operations in the Matlab environment. The information vector is normalized linearly to finish the CRITIC weight set. After calculating the weights for both the entropy weight method and the CRITIC weight method, the two weights are fused using a linear weighted combination. To maintain the independence of the two weight sources, a consistency test is first performed, calculating the Pearson correlation coefficient between the two sets of weights. If the result is less than 0.85, it indicates significant differences in the weight distribution between the two, indicating that the two are suitable for fusion. The combination coefficients α and β ($\alpha + \beta = 1$) are then set, and the weight ratios are determined based on empirical analysis, so that the entropy weights reflect internal differences in the indicators and the CRITIC weights reflect the balance of conflicts between the indicators [29,30]. The combined weight vector is renormalized to form the final dynamic weight result, which serves as the input parameter of the comprehensive evaluation model.

Figure 2(a) shows the weights of each indicator and their corresponding entropy values calculated using the entropy weight method. The X-axis represents indicator numbers 1 to 15; the left side of the Y-axis represents the entropy weight of the indicator, and the right side represents the entropy value. It can be seen that the weights of the first five indicators are relatively low, and their impact on overall data quality is relatively small. The graph of entropy value curve shows how much information is available in the indicator. A weight with greater emphasis has a lower entropy value, and therefore a larger difference in information. Figure 2(b) compares the entropy weight method, CRITIC method, and combined weight method. The X-axis also denotes the indicator number, and the Y-axis indicates the weight value. The combined weight method shows that the combination of the two methods produces a more equal distribution of weights. This method retains the entropy weight methods sensitivity to information differences while considering support from the CRITIC method to resolve dissagreements and possible fluctuations. Therefore, the overall weighting is more rational to evaluate the performance of each secondary hospital on each data quality indicator, using multi-dimensional data quality indicators weight allocation to complete the weight process, avoiding exposure to subjective factors, so that the weight system considers the differences inherent to indicators and the relational dimension of indicators.

C. IMPROVED TOPSIS COMPREHENSIVE EVALUATION MODEL IMPLEMENTATION

1) Ideal Solution and Distance Calculation Process

Traditional TOPSIS relies on Euclidean distance, which is susceptible to extreme values in multidimensional sparse data and struggles to capture the morphological similarity of indicator sequences. This paper proposes a TOPSIS model with grey relational degree correction: first, a weighted Euclidean distance is calculated, and then a correction factor is constructed using grey relational degree (resolution coefficient ρ=0.5) to nonlinearly adjust the distance. This mechanism maintains the geometric intuition of TOPSIS while enhancing its robustness to local anomalies and sequence trends.

After the weights of the entropy-CRITIC method are determined, a modified TOPSIS evaluation method is developed from a weighted standardized matrix. First, the positive and negative ideal solutions are established based on the attribute direction of each indicator; the positive ideal formula is created using the maximum value of each weighted indicator, and the negative ideal solution is established using the minimum value, resulting in an ideal theoretical worst-case and best-case measure of data element quality. Next, the weighted distances of the sample hospitals in each indicator dimension are calculated, and the Euclidean distance function performs row vector operations on the weighted standardized matrix to develop spatial distances of each sample to their positive and negative ideal solutions. To eliminate extreme value sensitivity of traditional Euclidean distance measures, this paper recommends the application of an interval standardization coefficient, and interval mapping processing of the distance matrix, to preserve a balanced distance distribution and minimize offsetting from the differences in indicator scales.

Once the distance is computed, the gray correlation degree is applied to adjust distance to increase model performance in identifying local outliers. By having a gray correlation matrix composed of the ideal solution and sample sequence, the calculation of the correlation coefficient sequence and gray correlation degree is developed. The correlation coefficient used in the model is a traditional form of resolution coefficient ρ = 0.5, helping the model to strike a balance between correlation and discrimination. The gray correlation degree output is weighted with the Euclidean distance, creating a distance matrix. This distance matrix is quantitatively demonstrated the closeness of each hospital sample to the ideal solution a multi-dimensional indicator space and also minimizes bias due, to distribution of abnormal indicators, resulting in stable model discrimination performance, that also compares performance between hospitals, with a complex, multi-dimensional data structure [31].

2) Comprehensive Score Calculation and Ranking Implementation

After the matrix of weighted distance has been produced, all the sample hospitals' closeness scores can be computed. Closeness is interpreted as the ratio of the distance to the negative ideal solution to the distance to the positive ideal solution plus the distance to the negative ideal solution, which is an indicator of how close to the ideal solution an sample is located. Closeness scores for each sample hospital is calculated using matrix functions to accuracy and to ensure consistency of replication. In order to alleviate any potential ambiguity in rankings that may arise from a congested closeness sequence, a normalization function can be applied to subjectively perform a nonlinear transformation on the closeness sequence that better represents sample differences [32,33]. The normalized closeness score can then be interpreted directly and used as the hospital data quality score for further analysis on the ranks.

To validate the rationality of the model's internal structure, the proximity results are tested for consistency and stability. First, the correlation coefficient between the positive and negative ideal solution distances is calculated to ensure a significant negative correlation between the two, demonstrating the correct directionality of the model. Second, the coefficient of variation is used to analyze the dispersion of the proximity distribution and test the model's discriminatory performance. A coefficient of variation greater than 0.2 indicates that the model output has sufficient discriminatory power. Finally, the improved TOPSIS model's ranking stability is repeatedly tested. The model output is considered stable if the proximity ranking consistency exceeds 90% over multiple iterations.

Figure 3 shows the comprehensive performance of secondary hospitals under multiple data quality indicators. Figure 3(a) presents the distance of each hospital to the positive ideal solution (D+) and the negative ideal solution (D-) in the form of scattered points, with the horizontal axis representing the Euclidean distance to the positive ideal solution and the vertical axis representing the distance to the negative ideal solution. The color of the points represents the comprehensive closeness score. It can be observed that hospitals that are closer to the positive ideal solution and farther from the negative ideal solution have higher scores, while hospitals farther from the positive ideal solution have lower scores. Figure 3(b) uses PCA to reduce the dimensionality of the hospital data to a three-dimensional space. The proportion shows the contribution of each principal component to the total variance, and the color of the points also represents the comprehensive score.

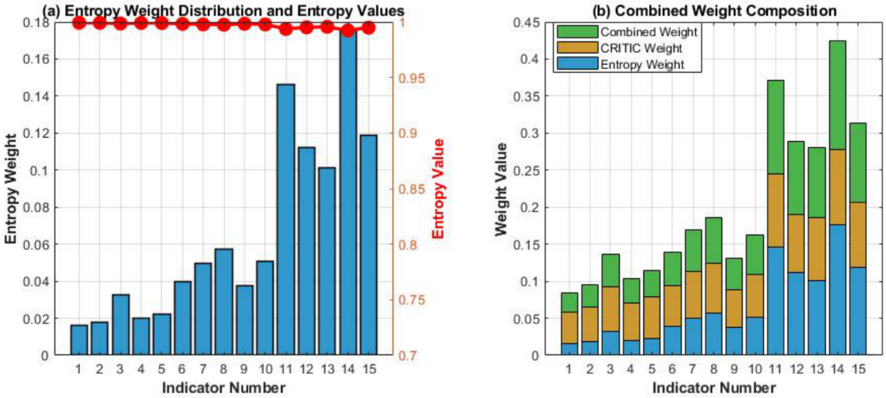

FIGURE 2. Weight distribution analysis chart. Figure 2(a). Entropy weight distribution and entropy value curve; Figure 2(b). Comparison bar chart of entropy weight method, CRITIC method, and combined weight

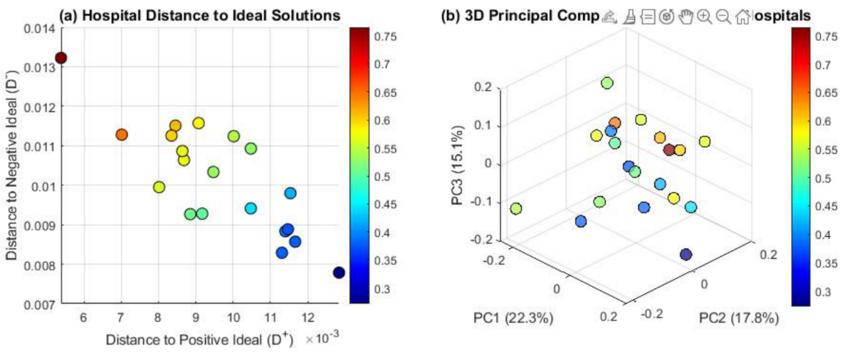

FIGURE 3. TOPSIS hospital data quality analysis. Figure 3 (a). Scatter plot of distance to ideal solution; Figure 3(b). 3D principal component projection of hospital data quality

The improved TOPSIS model combines Euclidean distance and grey correlation in its algorithm structure to achieve a joint measurement of the comprehensive distance of multidimensional indicators. It not only maintains the intuitiveness of the traditional TOPSIS method, but also enhances the ability to identify data quality differences under complex indicator systems.

III. COMPREHENSIVE EVALUATION RESULTS OF DIRECT REPORTING DATA QUALITY

The dataset used in this evaluation experiment came from real-time infectious disease reporting cards provided by the Yaodu District Disease Prevention and Control Information System in Linfen City. It covers 355 cases of legally notifiable infectious diseases reported by 12 secondary hospitals from January 1, 2022 to September 30, 2025. The diseases include six categories of Class B/C infectious diseases: hepatitis B/C, tuberculosis, syphilis, bacterial dysentery, hand-foot-and-mouth disease, and novel coronavirus infection. Each case includes demographic information (gender, age, occupation), date of onset/diagnosis, disease classification, reporting institution, and the status of key field completion. This study was approved by the Ethics Committee of Linfen Second People's Hospital, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. The data missing rate is $< 5\%$, and the data has been pre-validated by the district-level CDC platform to ensure the reliability of the analysis.

A. VERIFICATION OF THE APPLICABILITY OF THE EVALUATION INDEX SYSTEM

In the verification of the applicability of the indicator system, the fields of the direct reporting data of several secondary hospitals in the region in the past three years are first cleaned and standardized. The data format is unified, and outliers are eliminated. Then, the values of each indicator are calculated based on the constructed multidimensional indicator system, and the distribution of each indicator among the sample hospitals is statistically analyzed to see whether there are indicators with too high a missing rate or too low a variability [34]. SPSS software is used to perform correlation analysis to test the internal logical relationship between each indicator and dimension and ensure that there is no high correlation redundancy between the indicators. Finally, the factor analysis method is used to extract the principal components and evaluate the cumulative variance contribution rate to verify the operability and representativeness of the indicator system in the data of multiple hospitals.

Figure 4(a) shows the distribution of 15 indicators across the five dimensions of completeness, timeliness, accuracy, consistency, and relevance for 10 hospitals. The horizontal axis represents the specific indicator for each dimension, while the vertical axis represents the indicator value. It can be seen that the indicators for the timeliness dimension are generally high, with a median of approximately 0.85 or above, indicating that secondary hospitals perform well in terms of data timeliness. The overall indicators are all above 0.8, indicating that each hospital performs well in terms of multidimensional data. Figure 4(b) shows the correlation matrix between the 15 indicators. Five distinct regions of high correlation can be observed along the main diagonal, corresponding to close associations between indicators within each dimension. The correlation coefficients generally range from 0.7 to 0.9, indicating strong consistency and synergy between indicators within the same dimension. Correlations between different dimensions generally range from 0.2 to 0.5, with relatively lighter colors, reflecting the relative independence of information between dimensions.

B. COMPREHENSIVE SCORE CALCULATION AND RANKING

In the comprehensive score calculation phase, the standardized indicator matrix of each hospital is input into the improved TOPSIS model. The indicators are weighted according to the entropy weight-CRITIC combined weight, and the positive ideal solution and negative ideal solution vectors are constructed. The weighted distance of each hospital sample to the ideal solution is calculated. At the same time, the distance is corrected using the gray correlation degree to construct a joint measurement matrix. Then, the proximity value of each hospital is calculated and normalized to make the distribution more balanced and the distinction more obvious [35]. The normalized proximity values are sorted to form the data quality level of each hospital.

Table 1 shows the comprehensive evaluation results of the infectious disease online direct reporting data quality of 10 secondary hospitals based on the improved TOPSIS model. As can be seen, hospitals H01 and H05 achieve comprehensive scores of 0.862 and 0.847, respectively, ranking first and second. Their positive ideal solution distances (D+) are relatively small (0.125 and 0.138, respectively), while their negative ideal solution distances (D-) are relatively large (0.780 and 0.765, respectively). This indicates that their data perform well overall in terms of completeness, timeliness, and accuracy, ranking them at a "high" quality level. Mid-level hospitals, namely H08 and H03, have comprehensive scores of 0.831 and 0.814, respectively. These mid-level hospitals have shown a decreased gap in the distance to their positive and negative ideal solutions, which indicates some of their data indicators are at least somewhat volatile. Hospitals such as H12, H04, and H07 have lower scores (0.742, 0.730, and 0.715, respectively), and also have large D+ values and small D- values, which indicates that their data falls significantly short of the ideal, mainly regarding timeliness and consistency. In general, the comprehensive score is inversely related to the distance from the ideal solution. The hospitals receiving higher scores have greater alignment with ideal data quality while the lower scoring hospitals still have room for improvement in multiple dimensions.

C. MODEL PERFORMANCE AND RELIABILITY EVALUATION

A Kendall correlation coefficient is first used to quantify the relationship between the ranking of each hospital integrated general score and the unidimensional ranking of the comprehensive score indicators to validate the logical soundness of the ranking data. The coefficient of variation is then computed on the score sequence for each hospital, and dispersion of the resultant distribution is analyzed to validate the model's discrimination power and stability. Once this is established, a principal component loading analysis is conducted from the indicator matrix to examine each indicator's contribution to the general score indicator and ensuring transparency and interpretability of the model output. The evaluation process employs calculation and visualization batch processing tools in Python and SPSS, and repeated execution is conducted to validate consistency and reliability of the results and confirm the scientific nature of the composite model output is reproducible.

Table 2 shows values for each key performance and reliability evaluation of the 10 secondary hospitals. It includes the rank consistency of single-dimensional indicators (Kendall τ), the coefficient of variation (CV), and the principal component analysis (PCA). In general, the Kendall τ values for each of the hospitals range from 0.74 to 0.83. This is indicative of a high level of consistency for the rank of each dimension using the improved TOPSIS model. H05 (τ = 0.83) and H01 (τ = 0.82) show the most consistency, thus demonstrating the stability and certainty of the model using data from these two hospitals. The CV values range from 0.043 to 0.060, thus indicating that fluctuation in scores is minor in all hospitals, showing moderate discrimination of the model, and good reliability. For example, the CV of H05 is 0.043, indicating minimal fluctuation and the most reliability. With respect to principal component contributions rates, the majority of hospitals are clustered between 69.8% and 79.1%. Hospitals H05 (79.1%) and H01 (78.5%) have the highest principal component contribution rates, indicating that the principal components can explain the variance structure from the composite score well. The model presents good reliability for consistency in the integration with respect to multi-dimensional indicators, stability of ranking, and explanatory power.

The research, utilizing model-based detailed scoring and multi-dimensional indicator analysis, identifies important elements affecting data quality. The study finds that institution level and regional characteristics have an important effect on data quality; accuracy and consistency are the most significant areas of need within in hospitals. Institutions

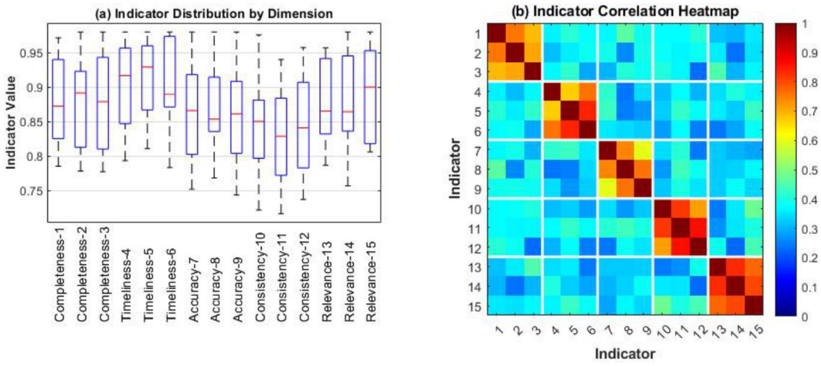

FIGURE 4. Hospital multidimensional data quality analysis. Figure 4(a). Boxplot of the distribution of indicators in each dimension; Figure 4(b). Heat map of correlation between indicators

TABLE 1. Comprehensive scores and ranking examples of each hospital

Rank	Hospital	Composite Score	Distance to Positive Ideal D^+	Distance to Negative Ideal D^-	Data Quality Level
1	H01	0.862	0.125	0.78	High
2	H05	0.847	0.138	0.765	High
3	H08	0.831	0.15	0.74	Medium
4	H03	0.814	0.165	0.72	Medium
5	H10	0.798	0.178	0.705	Medium
6	H02	0.782	0.19	0.69	Medium
7	H06	0.765	0.205	0.675	Medium
8	H12	0.742	0.22	0.66	Low
9	H04	0.73	0.235	0.645	Low
10	H07	0.715	0.25	0.63	Low

TABLE 2. Example of model performance and reliability evaluation

Hospital	Kendall τ (Single-Dimension Consistency)	Coefficient of Variation (CV)	Principal Component Contribution (%)	Overall Evaluation
H01	0.82	0.045	78.5	Stable
H02	0.79	0.048	76.2	Stable
H03	0.81	0.05	74.9	Stable
H04	0.76	0.055	72.3	Stable
H05	0.83	0.043	79.1	Stable
H06	0.77	0.052	71.5	Stable
H07	0.75	0.058	70.2	Stable
H08	0.8	0.049	73.6	Stable
H09	0.74	0.06	69.8	Stable
H10	0.79	0.047	75	Stable

that receive low scores often have poor information system configuration or poor process management. In comparison, institutions with high scores generally have standardization of processes and trained personnel.

IV. CONCLUSIONS

This paper studies the quality of infectious disease data directly reported online by secondary hospitals by establishing a multidimensional indicator system and developing a comprehensive evaluation framework. This framework is highly relevant to the textile industry's need for enhanced occupational health monitoring, offering a template for constructing a multidimensional indicator system to evaluate and improve the safety and health data quality for workers. From the five dimensions of completeness, timeliness, accuracy, consistency, and relevance, fifteen quantifiable indicators are identified to form a set of actionable evaluation indicators. The entropy-weighted CRITIC combined model is used to dynamically calculate the indicator weights, by balancing covariate diversity with relevance, in order to have a scientific and reasonable distribution of indicator weight. The improved TOPSIS model is used to calculate comprehensive proximity from both Euclidean distance and gray correlation, to rank and analyze the directly reported data from each hospital. Evaluation results show the method can accurately differentiate data quality levels across hospitals and validate the model's reliability through consistency, stability, and interpretability.

AUTHOR CONTRIBUTIONS

Conceptualization –Qian Zhang, Shengjia Wang and Pengfei Li; methodology – Qian Zhang, Shengjia Wang and Pengfei Li; investigation – Qian Zhang and Shengjia Wang; writing-original draft preparation – Qian Zhang, Shengjia Wang and Pengfei Li. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

The dataset used in this evaluation experiment came from real-time infectious disease reporting cards provided by the Yaodu District Disease Prevention and Control Information System in Linfen City. It covers 355 cases of legally notifiable infectious diseases reported by 12 secondary hospitals from January 1, 2022 to September 30, 2025. The diseases include six categories of Class B/C infectious diseases: hepatitis B/C, tuberculosis, syphilis, bacterial dysentery, hand-foot-and-mouth disease, and novel coronavirus infection. Each case includes demographic information (gender, age, occupation), date of onset/diagnosis, disease classification, reporting institution, and the status of key field completion. This study was approved by the Ethics Committee of Linfen Second People's Hospital, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] N. Sandhu, S. Whittle, L. Otero Varela, D. Southern, C. A. Eastwood, and H. Quan, "Applying the hospital administrative data quality scoring tool in 15 countries," Health Information Management Journal, 18333583251362053, 2025, doi: 10.1177/18333583251362053.

[2] A. M. N. Renzaho, G. Dachi, K. Tesfaselassie, K. T. Abebe, I. Kassim, Q. Alam, et al., "Assessing the impact of integrated community-based management of severe wasting programs in conflict-stricken South Sudan: A multi-dimensional approach to scalability of nutrition emergency response programs," International Journal of Environmental Research and Public Health, vol. 18, no. 17, 9113, 2021, doi: 10.3390/ijerph18179113.

[3] Y. Wu, S. Fan, D. Liu, and X. Sun, "Psychological and behavior investigation of Chinese residents: Concepts, practices, and prospects," Chinese General Practice Journal, vol. 1, no. 3, pp. 149-156, 2024, doi: 10.1016/j.cgpj.2024.07.006.

[4] R. Moore, D. Chanci, S. R. Brown, M. J. Ripple, N. R. Bishop, J. Grunwell, et al., "Association of the child opportunity index with in-hospital mortality and persistence of organ dysfunction at one week after onset of Phoenix Sepsis among children admitted to the pediatric intensive care unit with suspected infection," PLOS Digital Health, vol. 4, no. 4, e0000763, 2025, doi: 10.1371/journal.pdig.0000763.

[5] C. Barcellona, Y. B. Mariñas, S. Y. Tan, G. Lee, K. C. Ko, S. Chham, et al., "Measuring health equity in the ASEAN region: Conceptual framework and assessment of data availability," International Journal for Equity in Health, vol. 22, no. 1, 251, 2023, doi: 10.1186/s12939-023-02059-2.

[6] S. Han, M. Chu, Y. Ren, Z. Zhang, H. Chen, Y. Li, et al., "Global, regional, and national epidemiology of neurological disorders and subcategories: Incidence and disability-adjusted life years, 1990–2021," European Journal of Medical Research, vol. 30, no. 1, 711, 2025, doi: 10.1186/s40001-025-02958-w.

[7] S. Rehman, E. Rehman, A. Mumtaz, and J. Zhang, "Cardiovascular disease mortality and potential risk factor in china: A multi-dimensional assessment by a grey relational approach," International Journal of Public Health, vol. 67, 1604599, 2022, doi: 10.3389/ijph.2022.1604599.

[8] J. E. Bennett, O. N. O'Driscoll, G. A. Stevens, N. Aldea-Ramos, F. Bray, F. Farzadfar, et al., "Benchmarking progress in non-communicable diseases: A global analysis of cause-specific mortality from 2001 to 2019," The Lancet, vol. 406, no. 10509, pp. 1255-1282, 2025, doi: 10.1016/s0140-6736(25)01388-1.

[9] R. Šmit, N. Wojtalewicz, L. Vierbaum, F. Nourbakhsh, I. Schellenberg, K. P. Hunfeld, et al., "Epidemiology, management, quality of testing and cost of syphilis in Germany: A retrospective model analysis," Frontiers in Public Health, vol. 10, 883564, 2022, doi: 10.3389/fpubh.2022.883564.

[10] R. Jiang, Y. Zhou, M. Yang, L. Yue, L. Huang, and Y. Wang, "Factors influencing the management of chronic non-communicable diseases in cold regions: A qualitative study from Northeast China," Frigid Zone Medicine, vol. 5, no. 1, pp. 1-9, 2025, doi: 10.1515/fzm-2025-0001.

[11] S. Aktaa, A. Abdin, E. Arbelo, H. Burri, K. Vernooy, C. Blomström-Lundqvist, et al., "European Society of Cardiology Quality Indicators for the care and outcomes of cardiac pacing: Developed by the Working Group for Cardiac Pacing Quality Indicators in collaboration with the European Heart Rhythm Association of the European Society of Cardiology," EP Europace, vol. 24, no. 1, pp. 165-172, 2022, doi: 10.1093/europace/euab193.

[12] I. Rocco, O. Tamburis, F. Pecoraro, D. Luzi, B. Corso, and N. Minicuci, "Quality of child healthcare in European countries: Common measures across international databases and national agencies," European Journal of Public Health, vol. 31, no. 4, pp. 679-687, 2021, doi: 10.1093/eurpub/ckab086.

[13] D. Duives, Y. Chang, M. Sparnaaij, B. Wouda, D. Boschma, Y. Liu, et al., "The multi-dimensional challenges of controlling SARS-CoV-2 transmission in indoor spaces: Insights from the linkage of a microscopic pedestrian simulation and virus transmission models," medRxiv, 2021-04, doi: 10.1101/2021.04.12.21255349.

[14] C. Feng, F. Weng, and Z. Liu, "Evaluation index system for value co-creation in integrated care within the context of digital transformation," Scientific Reports, vol. 15, no. 1, 18118, 2025, doi: 10.1038/s41598-025-02907-x.

[15] I. Schellenberg and B. Lohr, "Epidemiology, Management, Quality of Testing and Cost of Syphilis in," Syphilis infection: Clinical, Epidemiology, Basic Science, and Behavioral Research, vol. 16648714, p. 64, 2023.

[16] R. P. Arboleda and M. N. R. Diaz, "Quality Improvement Strategies of Secondary Hospitals in the Philippines: A Basis for Hospital CQI Model," Journal of Business and Management Studies, vol. 7, no. 3, pp. 158-290, 2025, doi: 10.32996/jbms.2025.7.3.14.

[17] B. O. Owolabi and F. A. Owolabi, "Predictive AI-driven epidemiology for tuberculosis outbreak prevention in achieving zero TB city vision," Int J Adv Res Publ Rev, vol. 2, no. 5, pp. 318-340, 2025, doi: 10.55248/gengpi.6.0525.1994.

[18] L. Tan, "The trust gap in health communication: The correlation between socioeconomic status and perceived information credibility," Global Media and Social Sciences Research Journal, vol. 6, no. 1, pp. 324-330, 2025, doi: 10.71465/gmssrj104.

[19] D. Kong, N. Jiang, X. He, J. Yuan, Q. Du, and W. Lian, "Enhancing Health Productivity in China: A Decade of Measurement and Regional Insights (2010–2020)," Risk Management and Healthcare Policy, pp. 869-889, 2025, doi: 10.2147/RMHP.S500994.

[20] I. Margineanu, T. Butnaru, M. Lam, D. Baiceanu, R. Dragomir, A. S. Arbore, et al., "Tuberculosis impacts multiple aspects in quality of life in a Romanian cohort of drug-susceptible and drug resistant patients: A patient-reported outcome measures study," Tropical Medicine & International Health, vol. 29, no. 7, pp. 584-593, 2024, doi: 10.1111/tmi.13996.

[21] Z. Cao, Z. Qiu, F. Tang, S. Liang, Y. Wang, H. Long, et al., "Drivers and forecasts of multiple waves of the coronavirus disease 2019 pandemic: A systematic analysis based on an interpretable machine learning framework," Transboundary and Emerging Diseases, vol. 69, no. 5, pp. e1584–e1594, 2022, doi: 10.1111/tbed.14492.

[22] S. L. Hu, "Research on monitoring system of daily statistical indexes through big data," Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 15, no. 5, pp. 731-740, 2022, doi: 10.2174/2666255813999201113124021.

[23] X. Wang, Y. Liu, Y. Ding, M. Li, Y. Gao, and T. Li, "Disease spectrum of patients with hospital-acquired multidrug-resistant organism infections in the intensive care unit: A retrospective study," Frontiers in Microbiology, vol. 16, 1568615, 2025, doi: 10.3389/fmicb.2025.1568615.

[24] S. M. Oyekunle, O. M. Olaniyi, E. A. Oluyinka, M. O. Gbadebo, and R. O. Olasege, "Adaptive AI Pipelines in Population Health Analytics for Enhanced Security and Data Governance," Journal of Engineering Research and Reports, vol. 27, no. 10, pp. 338-358, 2025, doi: 10.9734/jerr/2025/v27i101677.

[25] F. Sampurno, J. Cally, J. L. Opie, A. Kannan, J. L. Millar, A. Finelli, et al., "Establishing a global quality of care benchmark report," Health Informatics Journal, vol. 27, no. 2, 146045822110157, 2021, doi: 10.1177/14604582211015704.

[26] Q. Wei, B. Yuan, J. Xu, P. He, H. Xu, and Q. Meng, "Continuity of health care: Measurement and application in two rural counties of Guangxi Province, China," BMC Health Services Research, vol. 23, no. 1, 917, 2023, doi: 10.1186/s12913-023-09916-4.

[27] P. Luoga, G. R. Mahiti, O. Kapinga, B. Herzon, S. M. Mwaisengela, P. A. Materu, et al., "Socioeconomic and geographical disparities in disability distribution among Tanzanian population: Insights from the 2022 demographic and health survey and malaria indicator survey," BMC Public Health, vol. 25, no. 1, 2455, 2025, doi: 10.1186/s12889-025-23152-x.

[28] H. Thamrin, H. Ridho, I. D. Saragih, and G. Pirandy, "Exploring social determinants of health and their influence on quality of life and social welfare in Indonesia: A comprehensive review," Retos: Nuevas Tendencias En Educación Física, Deporte y Recreación, vol. 68, pp. 1497-1517, 2025, doi: 10.47197/retos.v68.116126.

[29] A. Lamer, M. D. Moussa, R. Marcilly, R. Logier, B. Vallet, and B. Tavernier, "Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project," Journal of Clinical Monitoring and Computing, vol. 37, no. 2, pp. 461-472, 2023, doi: 10.1007/s10877-022-00898-y.

[30] B. M. Alsahabi, A. H. Almatrafi, M. A. Jubran, F. H. Alsahabi, A. I. Alghannam, and M. A. Albahooth, "Antibiotic Stewardship in Dentistry Examine Strategies to Optimize Antibiotic Use in Dental Practice, Considering the Rise of Antibiotic Resistance and the Impact on Patient Outcomes," Journal of University Studies for Inclusive Research, vol. 16, no. 24, pp. 12426-12440, 2023.

[31] J. Zhang, Y. Li, H. Qu, and P. Chen, "An analysis of operational efficiency and its influencing factors in traditional Chinese medicine hospitals in Shaanxi Province, China," Frontiers in Public Health, vol. 13, 1499134, 2025, doi: 10.3389/fpubh.2025.1499134.

[32] S. I. Shanmugam, S. K. Shanmugam, and K. R. Shanmugam, "Child Survival and Health Index for India," Child Indicators Research, vol. 18, no. 2, pp. 651-673, 2025, doi: 10.1007/s12187-024-10207-5.

[33] H. L. Sargénius, T. B. Rø, R. E. Hypher, A. E. Brandt, S. Andersson, T. G. Finnanger, et al., "Health-related quality of life in children and adolescents with paediatric acquired brain injury: Secondary data analysis from a randomised controlled trial," Quality of Life Research, vol. 34, no. 2, pp. 577-588, 2025, doi: 10.1007/s11136-024-03838-2.

[34] S. Akter, M. H. A. Banna, K. Brazendale, M. S. Sultana, S. Kundu, T. R. Disu, et al., "Determinants of health care seeking behavior for childhood infectious diseases and malnutrition: A slum-based survey from Bangladesh," Journal of Child Health Care, vol. 27, no. 3, pp. 395-409, 2023, doi: 10.1177/13674935211057714.

[35] J. Zhao, Z. Kuang, J. Xu, X. Yu, J. Dong, J. Li, et al., "Association of lifestyle behaviors with health-related quality of life among patients with hematologic diseases," Quality of Life Research, vol. 32, no. 4, pp. 1119-1131, 2023, doi: 10.1007/s11136-023-03343-y.