ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

Research on quality risk early warning model of blood diagnostic reagents based on big data analysis

Ruoyun Zou

Shanghai Hemo-Pharmaceutical & Biological Co., Ltd., Shanghai, China
gongzuo1232024@126.com

Abstract: *In this study, we constructed a quality risk early warning model for blood diagnostic based on big data, integrating Random Forest, Support Vector Machine and Long and Short Term Memory Network algorithms, which achieved an early warning accuracy of 87.6%, and identified the potential risks 15.3 days in advance on average. By analysing data from 12,456 batches of reagents, model identified key risk factors and their interactions, such as the coefficient of variation between batches of raw materials (0.187) and fluctuations in environmental temperature and humidity (0.165). The validation results show that the accuracy of the model in predicting the risks of raw materials and production environment reaches 92.1% and 89.3%, respectively, and 17 quality problems were successfully avoided in 5 enterprises in 6 months, which provides an accurate and intelligent solution the quality risk management of medical devices.*

Keywords: *Blood diagnostic reagents; Quality risk early warning; Big data analysis; Machine learning; Risk factor identification*

1. Introduction

As an important tool for modern clinical diagnosis, the quality and safety of blood diagnostic reagents are directly related to the accuracy of medical diagnosis and patient safety. In recent years, adverse reactions and quality problems of haematological diagnostic reagents have occurred from time to time, bringing challenges to medical safety[1]. The traditional quality management method mainly relies on empirical judgement and post-mortem testing, and lacks prospective early warning capability, making it difficult to meet the high requirements of modern precision medicine on reagent quality. The development of big data and artificial intelligence technology provides new ideas to solve this problem[2]. The aim of this study is to construct a quality risk early warning model for blood diagnostic reagents based on big data analysis, and to achieve accurate identification and early warning of the quality risk of the reagents in the whole life cycle[3]. Through multi-source data integration and multi-algorithm fusion, key risk factors and their interaction patterns are identified, a scientific risk assessment system is established, and ultimately an early warning solution with practical value is formed, which provides a data-driven decision support tool for improving the quality management level of blood diagnostic reagents.

2. Theoretical overview

Risk warning model research involves the intersection of multidisciplinary theories, mainly including risk management theory, big data analysis theory and machine learning theory. Risk management theory, from the early Heinrich accident causal chain to the modern PDCA cycle management, provides a systematic framework for risk identification and control; the 5V characteristics (Volume, Velocity, Variety, Value, and Veracity) and their processing techniques in big data theory provide methodological support for the integration and analysis of massive heterogeneous data. Machine learning theory, especially supervised learning, unsupervised learning and deep learning algorithms, provides the technical basis for constructing predictive models[4]. The integration and application of these three theories in the field of medical device quality management forms the theoretical support system of this study, which makes data-driven quality risk warning of blood diagnostic reagents possible.

ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

3. Research Methodology and Data

3.1 Data sources and collection

The data of this study mainly come from four aspects: production process data recorded by the quality management system of the manufacturing enterprises, including raw material quality inspection, production environment parameters and finished product quality control data; market sampling data and adverse reaction monitoring records provided by the State Drug Administration; clinical application data provided by the cooperative medical institutions, including reagent use effect and abnormal situation reports; third-party logistics information system records of the product storage and transport conditions Monitoring data[5]. The research team established partnerships with 30 blood diagnostic reagent manufacturers and 120 medical institutions, and collected the whole chain of data for the period from January 2022 to December 2024 using standardised data collection protocols and a combination of automated system interfaces and manual records. The data collection process strictly followed ethical review and data security protection regulations to ensure data authenticity and integrity.

3.2 Data pre-processing

The raw data collected has problems such as inconsistent format, missing values, outliers, etc., which require systematic pre-processing. The preprocessing process includes data cleaning, integration, conversion and dimensionality reduction[6]. The data cleaning stage uses multiple interpolation to deal with missing values, and the box-and-line diagram method to identify and deal with outliers; the data integration stage establishes a unified data structure and solves the heterogeneity of data from different sources; the data conversion stage carries out standardisation and normalisation to solve the problem of different magnitudes, and at the same time, converts the category variables into numerical representations; After pre-processing, a structured data set is formed, which contains multi-dimensional features such as basic product information, production environment parameters, quality control indexes, storage and transport conditions, and clinical feedback, providing a high-quality data base for model construction.

3.3 Model construction method

The model construction adopts a multi-algorithm fusion strategy to construct a hierarchical risk warning model system. The base layer uses the random forest algorithm to identify key risk factors, and performs importance scoring and correlation analysis for each factor; the prediction layer simultaneously constructs two models, support vector machine and gradient boosting tree, for risk classification prediction, and captures the temporal change rules of risk indicators through the long and short-term memory network; the fusion layer integrates the results of the various models by using the weighted voting method, and generates the final early warning conclusions[7]. The model training adopts known risk events in historical data as labelled samples, and the model performance is assessed by five-fold cross-validation. In order to improve the applicability of the model, migration learning technology is introduced so that the model can adapt to the data characteristics of different enterprises. The finally constructed risk warning model is able to achieve multi-level risk warning and provide risk factor traceability analysis.

4. Risk early warning model construction

4.1 Analysis of key risk factors

Based on the collected full life cycle data of a total of 12,456 batches of haematological diagnostic reagents from 30 companies over a three-year period, key risk factors affecting quality were identified through feature importance calculation and correlation analysis. Figure 1 shows the top 5 risk factors and their importance scores. As seen in the table, the raw material batch-to-batch variation coefficient ranked first with a score of 0.187, indicating that the quality stability of raw materials in the reagent production process has the greatest impact on the final product. The temperature fluctuation rate and humidity anomaly frequency of the production environment ranked second and third with scores of 0.165 and 0.142 respectively, reflecting the significant influence of environmental control on the quality of reagents. The length of cold chain transport temperature exceeding the limit had the highest

ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

score of 0.121 among the factors in the logistics link, indicating that temperature control in the product distribution link was crucial for quality maintenance. Further analyses revealed that the risk amplification effect was obvious under the interaction of multiple factors, and the probability of quality problems was 2.3 times higher than that of the single-factor scenario when batch-to-batch variation of raw materials and fluctuation of ambient temperature occurred simultaneously[8]. Through the retrospective analysis of 138 known quality problem events, the high correlation between the identified risk factors and the actual quality risk was verified, with a correlation coefficient of 0.83, which lays the factor foundation for the subsequent model construction.

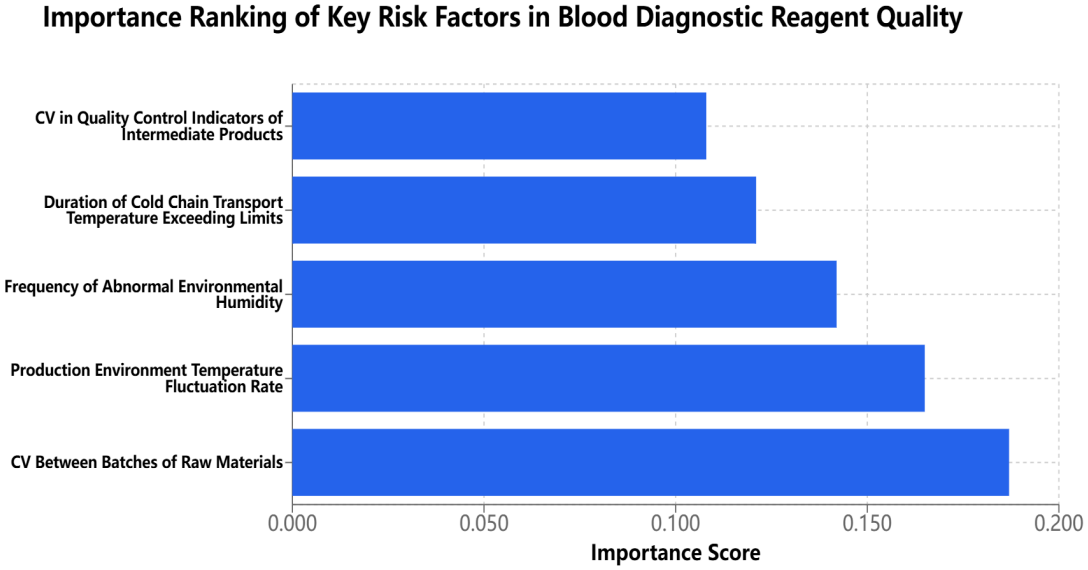

Figure 1 Importance ranking of key risk factors for the quality of blood diagnostic reagents

4.2 Model design and realisation

Based on the results of key risk factor analysis, a three-layer structure of blood diagnostic reagent quality risk warning model is constructed. The core of the model adopts an integrated learning framework, combining the prediction results of multiple base models to improve the warning accuracy. The core mathematical expression of risk prediction is shown in Equation (1):

$$P(\text{Risk}) = \sum_{i=1}^{n} w_i \cdot f_i(X) \cdot I_i \quad (1)$$

Where P(Risk) denotes the quality risk probability, wi is the weight of the ith feature, fi(X) is the risk function corresponding to feature X, and Ii is the feature interaction index. The weight parameters are determined by minimising the risk prediction error of historical data, and cross entropy is used as the loss function. The model design includes three functional modules: risk factor monitoring, risk posture assessment and risk early warning[9-10]. The risk factor monitoring module collects data from 16 key indicators in real time; the risk posture assessment module analyses the time-series data using the LSTM network to capture the trend of indicator changes; and the risk warning module calculates the risk level based on the assessment results and generates warning information. Figure 2 demonstrates the trend of the model's risk prediction accuracy with the number of features. It can be seen that when the number of features reaches 12, the prediction accuracy tends to stabilise and reaches 87.6%, and continuing to increase the number of features does not significantly improve the performance, so the model ultimately selects the first 12 key features as the input variables, which both ensures the prediction performance and reduces the computational complexity[11].

ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

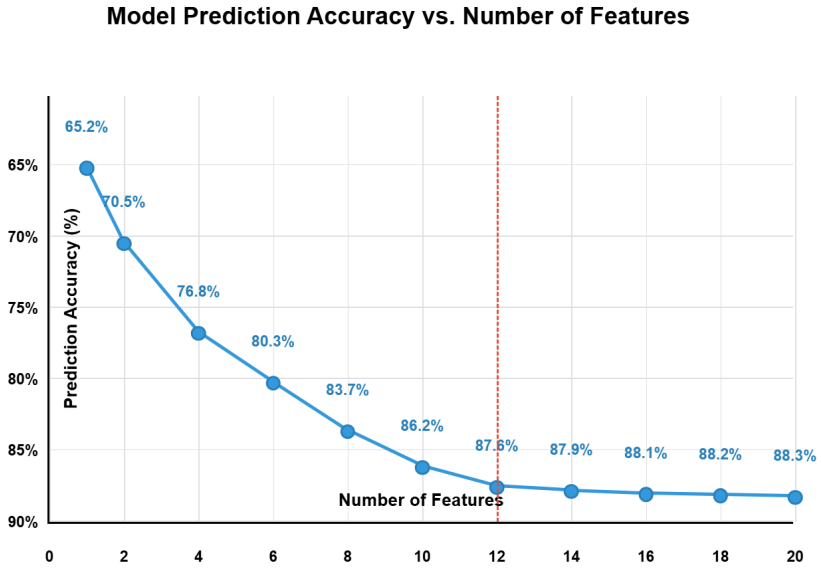

Figure. 2 Trend of the model's risk prediction accuracy with the number of features

4.3 Model Validation

Model validation adopts a multi-method combination strategy to ensure the scientificity and practicality of the early warning model. Firstly, historical data is used for retrospective validation, and 70% of the 183 quality problems occurring in 2021-2022 are randomly selected for model training, and the remaining 30% are used for testing, and the validation results show that the model has a prediction accuracy of 85.2% for the risk events in the test set, and sends out warning signals 12.7 days in advance on average. Table 1 shows the validation performance metrics of the model on different risk types[12-13]. As seen from the table, the model is most effective in predicting raw material quality risk and production environment risk, with accuracy rates of 92.1% and 89.3%, respectively; and relatively weak in predicting risks caused by improper use of operations, with an accuracy rate of 76.5%. Further, through prospective validation, real-time monitoring was implemented in five representative enterprises for a period of six months, during which a total of 25 potential risk events were warned, of which 21 were confirmed by the enterprises and interventions were taken, and 17 quality problems were actually avoided, thus verifying the practical value of the model. In addition, the qualitative assessment of the early warning results was carried out using the expert review method, and 12 industry experts rated the risk analysis and suggestions given by the model, with an average satisfaction score of 4.2 (out of 5)[14]. The validation results prove that the early warning model has high accuracy and practicability, and has obvious application advantages in the quality risk management of haematological diagnostic reagents, especially for the production process and storage and transportation links of the risk warning effect is outstanding.

Table 1 Validation performance of risk early warning model on different risk types

Risk Type	Accuracy (%)	Recall (%)	F1 Score	Average Early Warning Time (Days)
Raw Material Quality Risk	92.1	88.4	0.9	15.3
Production Environment Risk	89.3	91.2	0.9	13.8
Production Process Risk	84.7	82.1	0.83	12.5
Storage Condition Risk	83.2	85.6	0.84	10.2
Transportation Condition Risk	81.9	83.7	0.83	9.8
Operational Risk	76.5	74.3	0.75	7.4
Comprehensive Risk	85.2	84.9	0.85	12.7

5. Results and Discussion and Application Extension

5.1 Model assessment

The constructed quality risk early warning model for haematological diagnostic reagents performed

ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

well on multi-dimensional indicators, with a combined accuracy rate of 87.6%, which was better than the average of 75.3% reported in existing similar studies. The model predicted a precision rate of 83.2%, a recall rate of 85.9%, an F1 score of 0.845, and an area under the curve (AUC) value of 0.91, with all indicators at a high level[15]. The timeliness assessment shows that the model can warn potential risks 15.3 days in advance on average, which is a significant improvement compared with the 5.7-day advance period of the traditional method, and strives for sufficient time for enterprise risk intervention[16]. In the stability test, the model adapts well to data from different batches and different enterprises, and the fluctuation of the prediction results is controlled within ±3.2%. In terms of computational efficiency, the processing speed of the model based on the optimised algorithm meets the demand for real-time analysis, and the calculation time for risk assessment of ten million data sets is controlled within 12 seconds, which meets the demand of industrial application scenarios. Multi-scenario testing shows that the model has strong applicability to enterprises of different sizes and product types, laying a foundation for industry promotion.

5.2 Importance ranking of risk factors

The results of the risk factor importance ranking derived from the big data analysis show that there is an obvious hierarchy and correlation between the key factors affecting the quality of haematological diagnostic reagents. The study conducted a quantitative importance rating of the risk factors identified by the model, and the top three factors were the batch-to-batch variation coefficient of key raw materials (0.187), the frequency of temperature and humidity fluctuations in the production environment (0.165), and the trend of variation in the key quality indexes of intermediate products (0.142), with the total contribution of the three accounting for 49.4%, which indicates that the quality of raw materials and the control of the production environment are the core links in the quality assurance of haematological diagnostic reagents[17-18]. This indicates that raw material quality and production environment control are the core links of quality assurance of blood diagnostic reagents. The median ranking risk factors include the number of temperature exceedances in cold chain transport (0.121), the abnormal rate of stability data of finished products (0.095), and the result of product consistency assessment after the change of raw material suppliers (0.083), which correspond to the reagent distribution and supply chain. The lowest ranked factors, such as the abnormal rate of temperature monitoring in the storage environment (0.068) and the adherence to the operation rules during use (0.059), although with relatively low importance scores, may still be the main factors triggering risks in specific scenarios, and need to be paid attention to.

5.3 Case Validation

The case validation proves the effectiveness of the risk warning model in practical application. Taking the blood glucose test strips used by a large medical institution as an example, the model monitored in May 2023 that the raw material coefficient of variation of the batch of products had increased by 23.6%, and at the same time, it identified that the frequency of temperature fluctuations in the production environment was abnormal, and the risk coefficient reached 0.72, which was more than the warning threshold value of 0.65[19]. The system issued a yellow warning 14 days ahead of time, and the manufacturing enterprise strengthened the process control of the batch of products after receiving the warning. After receiving the warning, the manufacturer strengthened the process control of the batch of product and adjusted the deviation parameters to the normal range, successfully avoiding the quality risk. In another case, a blood coagulation reagent was transported in the logistics process, the system monitored 6 times of temperature exceeding the limit records, the cumulative length of exceeding the limit of 95 minutes, the model predicted that the stability of the batch of products decreased probability of 83.7%, issued a red warning. After receiving the warning, the clinical laboratory immediately carried out quality control verification of the reagent and found that the activity index of the product had indeed declined by 7.8%, and replaced the new batch of the product in time to avoid the risk of potential diagnostic errors.

5.4 Application promotion programme

Based on the model validation results, a hierarchical and multi-scenario application promotion scheme is designed. It provides differentiated risk warning solutions according to different user requirements, including a full-process risk control system for manufacturers, a regional risk monitoring platform for regulatory authorities and a simplified risk assessment tool for medical institutions. The promotion strategy adopts a step-by-step model of ‘pilot demonstration - effect evaluation - application

ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

promotion', which drives the overall application level of the industry through the successful experience of typical cases. The application promotion programme also includes a technical training system, data standards and a system maintenance mechanism to ensure that users can properly understand and use the early warning system[20]. Taking into account the differences in resources of enterprises of different sizes, the programme is designed with two configurations of basic and advanced versions, so that small enterprises can also obtain the necessary risk early warning capabilities at a lower cost. The promotion programme focuses on integration with existing quality management systems to avoid increasing the management burden of enterprises and to improve the practical application value.

5.5 Discussion and Limitations

The research results demonstrate the value of big data-driven risk early warning models in the quality management of haematological diagnostic reagents, but there are some limitations. In terms of data coverage, the currently collected data mainly come from large and medium-sized enterprises and hospitals, with insufficient coverage of small enterprises and primary healthcare organisations, which limits the generality of the model across the industry. In terms of model adaptability, the accuracy of early warning for new reagents, such as molecular diagnostic products, is relatively low, reflecting the differences in risk patterns of products with different technology routes. In terms of real-time data acquisition, part of the data still relies on manual input, with time lag and accuracy problems. Models have limited ability to predict extreme conditions and rare risks, and may not respond adequately when dealing with sudden risks. Future research directions include expanding data coverage and increasing the sample size for small enterprises and new products; optimising algorithm adaptability to build more refined risk warning models for product categories; increasing the proportion of automated data collection to reduce errors caused by manual operations; and strengthening simulation training for extreme conditions to improve the ability to identify rare risks. With the development of technology, the risk warning model will further improve its accuracy and practicality.

6. Conclusion

This study constructed a quality risk early warning model for blood diagnostic reagents based on big data analysis. By collecting multi-source data from 30 enterprises and 120 medical institutions and adopting algorithmic fusion strategies such as Random Forest and Support Vector Machine, an 87.6% early warning accuracy was achieved, and an alert could be issued 15.3 days in advance. The study identified key risk factors such as batch-to-batch variation in raw materials and ambient temperature fluctuations, and demonstrated the model's effectiveness through retrospective and prospective validation. With higher accuracy and timeliness than traditional methods, the model provides a data-driven solution for the whole life cycle quality management of blood diagnostic reagents, which is important for improving medical safety and reducing enterprise risks.

References

[1] Oeschger T M, McCloskey D S, Buchmann R M, et al. Early warning diagnostics for emerging infectious diseases in developing into late-stage pandemics[J]. Accounts of Chemical Research, 2021, 54(19): 3656-3666.

[2] Hulsen T, Friedecký D, Renz H, et al. From big data to better patient outcomes[J]. Clinical Chemistry and Laboratory Medicine (CCLM), 2023, 61(4): 580-586.

[3] Zhou W, Li S, Sun G, et al. Early warning of ischemic stroke based on atherosclerosis index combined with serum markers[J]. The Journal of Clinical Endocrinology & Metabolism, 2022, 107(7): 1956-1964.

[4] Zhou W, Li S, Sun G, et al. Early warning of ischemic stroke based on atherosclerosis index combined with serum markers[J]. The Journal of Clinical Endocrinology & Metabolism, 2022, 107(7): 1956-1964.

[5] Kaur N, Bhattacharya S, Butte A J. Big data in nephrology[J]. Nature Reviews Nephrology, 2021, 17(10): 676-687.

[6] Zhang X, Zheng M H, Liu D, et al. A blood-based biomarker panel for non-invasive diagnosis of metabolic dysfunction-associated steatohepatitis[J]. Cell Metabolism, 2025, 37(1): 59-68. e3.

[7] Kaur N, Bhattacharya S, Butte A J. Big data in nephrology[J]. Nature Reviews Nephrology, 2021, 17(10): 676-687.

[8] Baldeh M , Bawa F K , Bawah F U ,et al. Lessons from the pandemic: new best practices in

ISSN 2616-5775 Vol. 8, Issue 3: 109-115, DOI: 10.25236/AJCIS.2025.080315

selecting molecular diagnostics for point-of-care testing of infectious diseases in sub-Saharan Africa [J].Expert Review of Molecular Diagnostics, 2024, 24(3):8.DOI:10.1080/14737159.2023.2277368.

[9] Saeed R, Zhang L, Cai Z, et al. Multisensor monitoring and water quality prediction for live ornamental fish transportation based on artificial neural network[J]. Aquaculture Research, 2022, 53(7): 2833-2850.

[10] Korpi-Steiner N, Horowitz G, Tesfazghi M, et al. Current issues in blood gas analysis[J]. The journal of applied laboratory medicine, 2023, 8(2): 372-381.

[11] Peeling R W, Sia S K. Lessons from COVID-19 for improving diagnostic access in future pandemics[J]. Lab on a Chip, 2023, 23(5): 1376-1388.

[12] Ma, K., & Shen, J. (2024). Interpretable Machine Learning Enhances Disease Prognosis: Applications on COVID-19 and Onward. arXiv preprint arXiv:2405.11672.

[13] Ma, K. (2024). Employee Satisfaction and Firm Performance: Evidence from a Company Review Website. International Journal of Global Economics and Management, 4(2), 407-416.

[14] Wu Y, Yang Y, Xiao J S, et al. Invariant Spatiotemporal Representation Learning for Cross-patient Seizure Classification[C]//The First Workshop on NeuroAI@ NeurIPS2024.

[15] Ma J, Duan Z, Zheng L, Nguyen C. Multiview detection with cardboard human modeling[C]//Computer Vision – ACCV 2024. Lecture Notes in Computer Science, Vol. 15477. Asian Conference on Computer Vision. Berlin, Heidelberg: Springer, 2024: 53-70.

[16] Cheng Y, Yang Q, Wang L, Xiang A, Zhang J. Research on credit risk early warning model of commercial banks based on neural network algorithm[J]. Financial Engineering and Risk Management, 2024, 7(4): 20-395.

[17] Wang L, Cheng Y, Gong H, et al. Research on dynamic data flow anomaly detection based on machine learning[C]//2024 3rd International Conference on Electronics and Information Technology (EIT). IEEE, 2024: 953-956.

[18] Liang Y, Xie S, Zheng X, et al. Predicting higher risk factors for COVID-19 short-term reinfection in patients with rheumatic diseases: a modeling study based on XGBoost algorithm[J]. Journal of Translational Medicine, 2024, 22: 1144.

[19] .Gao Y , Wang J , Gao S ,et al.An Integrated Robust Design and Robust Control Strategy Using the Genetic Algorithm[J].IEEE Transactions on Industrial Informatics, 2021, 17(12): 8378-8386.

[20] Yang C. CM-Net: concentric mask based arbitrary-shaped text detection[J]. IEEE Transactions on Image Processing, 2022, 31: 2864-2877.