Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37646

Research on Multi-source Data Fusion Quality Risk Prediction Model for Drug Registration Compliance

Zhenzhen Huang[1,*], Huan Gao[2]

[1]*Guangdong-Macao Pharmaceutical Co., Ltd., Zhuhai, China*
[2]*College of Pharmacy, Nanjing University of Chinese Medicine, Nanjing, China*
**Corresponding Author. Email: 330379581@qq.com*

Abstract. Drug registration compliance management faces practical challenges such as complex submission documents, dispersed risk sources, and diversified review stages. Traditional manual review methods have difficulty identifying potential quality risks in a timely manner, especially when cumulative risk is formed by cross source, cross stage, and cross type information. Against this background, a multi source data fusion quality risk prediction model was developed for the drug registration scenario. The model integrated submission and review texts, historical deficiency records, inspection results, manufacturing change information, and regulatory provisions. Unified preprocessing, feature extraction, fusion modeling, and stratified training were then applied to achieve risk level identification. The results showed that the multi source fusion model outperformed single source models in Accuracy, Recall, F1, and AUC. Among them, the fusion XGBoost model achieved the best overall performance and significantly reduced the miss rate of high risk cases. The interpretability analysis further indicated that historical deficiency frequency, regulatory terminology deviation rate, abnormal inspection ratio, and manufacturing change density were the key factors affecting risk judgment. This study provides a computable, verifiable, and early warning oriented technical path for drug registration compliance management, and offers more targeted digital support for enterprise pre review and regulatory assisted assessment.

Keywords: Drug registration compliance, Multi-source data fusion, Quality risk prediction, Machine learning, Intelligent regulation

1. Introduction

Drug registration compliance has gradually shifted from simple dossier submission and formal review toward a model of full process, cross stage, and multi actor quality risk governance, driven by more detailed regulatory requirements, growing dossier complexity, and stronger demands for review efficiency [1]. In practice, submission texts, inspection results, manufacturing changes, historical deficiencies, and regulatory clauses are often scattered across different information carriers, while a single risk signal may appear in different forms at multiple stages, making compliance judgment highly heterogeneous and strongly interconnected [2]. Existing studies have accumulated knowledge on dossier completeness, control of critical quality attributes, identification

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37646

of review deficiencies, and digital support for regulatory review, yet systematic research on quality risk prediction under multi source data interaction remains limited, especially in terms of integrating risk grading, model interpretability, and regulatory applicability [3]. Against this background, the research focus is placed on unified organization of heterogeneous data, cross type feature extraction, construction of risk prediction models, and identification of key influencing factors, thereby extending compliance assessment from static inspection to forward looking early warning.

2. Literature review

2.1. Research on drug registration compliance and quality risk management

Research on drug registration compliance and quality risk management has mainly focused on dossier completeness, consistency of pharmaceutical studies, control of critical processes, and identification of deficiency issues, and these studies have already clarified how registration risks are transmitted across dossier preparation, technical review, and quality management activities [4]. This stream of literature generally understands compliance as the matching process between regulatory requirements and technical evidence, and therefore stresses that stable process parameters, sufficient research data, and traceable critical quality attributes are central to reducing the probability of registration failure [5]. As regulatory practice continues to deepen, the perspective of risk management has gradually entered registration research, and high frequency deficiencies, repeated supplements, and historical inspection outcomes are increasingly treated as important references for judging registration quality, while existing discussions still rely more on normative description and experience based interpretation than on quantitative representation of risk formation, interaction among variables, and integrated assessment of multiple risk signals [6].

2.2. Research on multi-source data fusion and intelligent prediction technologies

Research on multi source data fusion is largely grounded in the view that complex problems cannot be adequately explained from a single data perspective, which is why joint modeling of text, numerical records, behavioral information, and time series has become an important methodological route in intelligent prediction studies [7]. In fields such as healthcare, manufacturing, and supply chains, researchers usually employ feature level fusion, decision level fusion, or deep representation fusion to map information from different channels into a unified analytical space, thereby improving the capacity of models to capture abnormal patterns, latent relationships, and dynamic changes [8]. This methodological logic closely matches the scenario of drug registration compliance, because registration risk rarely comes from one isolated defect and is more often shaped by the combined effects of regulatory wording deviations, signs of quality fluctuation, historical behavioral patterns, and process changes.

2.3. Research on digital pharmaceutical regulation and computer-assisted review

The core concerns of digital pharmaceutical regulation research lie in regulatory process restructuring, intelligent review support, and data driven decision making, reflecting the structural pressure that traditional manual review faces when dealing with large scale registration information in terms of efficiency, consistency, and traceability [9]. Under this trend, knowledge graphs, natural language processing, rule engines, and machine learning have gradually been introduced into tasks such as regulatory retrieval, deficiency extraction, text review, and risk screening, and regulatory

DOI: 10.54254/2755-2721/2026.37646

review activities have extended from static document checking to dynamic integration of heterogeneous information. As digital tools enter the regulatory scenario, review logic increasingly moves from isolated document evaluation toward coordinated analysis of multidimensional evidence, while the boundary between compliance issues and quality issues becomes more closely connected, pushing regulatory attention from post hoc judgment toward process based early warning.

3. Experimental methods

3.1. Multi-source data system construction and preprocessing

The collected data covered five categories of public information required for drug registration compliance assessment, including submission and review texts, historical deficiency records, sampling inspection and testing notices, manufacturing change related announcements, and current regulatory provisions. The time range was set from 2019 to 2025. This range captured the data characteristics formed after the continuous refinement of drug registration supervision in recent years. It also made it possible to observe risk accumulation for the same enterprise or product across multiple registration stages. Public sources mainly included the official website of the National Medical Products Administration, the public sections of the Center for Drug Evaluation of the National Medical Products Administration, supervision and inspection announcements issued by provincial drug regulatory authorities, national drug sampling inspection notices, and current valid regulatory texts such as drug registration measures, technical guidelines for pharmaceutical research, and GMP related standards. Enterprise name, generic drug name, acceptance number, announcement date, and registration matter type were used as joint indices to map information from different sources to the same registration case. The preprocessing procedure included deduplication, missing value handling, standardization of enterprise names and product names, timestamp unification, normalization of regulatory terms, and text cleaning. A multi source database was then established with 2,486 registration cases, including 1,214 low risk cases, 826 medium risk cases, and 446 high risk cases. This database provided a unified foundation for subsequent feature extraction and stratified model training.

3.2. Feature extraction and data fusion modeling

Feature extraction was carried out in three layers according to data form. Structured data were used to derive statistical features, frequency features, and fluctuation features, such as deficiency frequency, abnormal inspection ratio, number of supplementary applications, number of manufacturing site changes, and interval between announcements. These variables reflected historical compliance performance, product quality stability, and the complexity of registration activities. Text data were processed through TF IDF and BERT embeddings. This step extracted semantic signals such as deviation in regulatory terminology, inconsistency in process descriptions, and omission of research details. Time related records were further converted into trend slope and stage fluctuation features, which helped identify whether risk accumulated gradually or appeared in concentrated periods [10]. The abnormal inspection ratio was calculated as the ratio of abnormal items to total tested items, as shown in Equation 1.

$$R_i = \frac{A_i}{T_i} \tag{1}$$

DOI: 10.54254/2755-2721/2026.37646

In the equation, R_i represents the abnormal inspection ratio of sample i, A_i represents the number of abnormal inspection items, and T_i represents the total number of inspection items. Feature level fusion was then adopted. Structured feature vectors, text embedding vectors, and time series derived features were concatenated into a unified input matrix. A weighted fusion function was used to control the contribution of different feature sources, as shown in Equation 2.

$$F_i=\alpha S_i+\beta E_i+\gamma Q_i, \quad \alpha+\beta+\gamma=1 \tag{2}$$

In the equation, F_i denotes the fused comprehensive risk representation, S_i denotes the structured feature vector, E_i denotes the text semantic embedding, and Q denotes the quality time series feature. The weights α, β, and γ correspond to the three components. In the experiment, α=0.35, β=0.40, and γ=0.2. This setting emphasized the dominant role of textual compliance information and historical behavior patterns in registration risk identification.

3.3. Risk prediction model training and evaluation design

The training procedure targeted stratified risk identification using the sample set in Section 3.1 and the fused feature matrix in Section 3.2. The output included three labels: low risk, medium risk, and high risk. Samples were split into training, validation, and test sets at a 7:2:1 ratio, with 1,740, 497, and 249 cases, respectively. Because high risk cases accounted for 17.9 percent, class weighting and five fold cross validation were applied to improve training stability. Three models were compared: XGBoost, Random Forest, and BERT plus MLP. XGBoost tested the effectiveness of fused features, Random Forest assessed tree based robustness, and BERT plus MLP captured semantic information from text data. Grid search was used for parameter tuning. Tree depth, learning rate, and subsampling ratio were optimized for XGBoost [11]. The number of trees and maximum features were adjusted for Random Forest. Learning rate, batch size, and hidden layer dimension were tuned for BERT plus MLP. Model performance was evaluated using Accuracy, Precision, Recall, F1, and AUC, together with the miss rate of high risk cases. SHAP was then applied to rank feature contributions and trace model outputs to major risk sources.

4. Results

4.1. Predictive performance of the multi-source fusion model

The test results showed that multi source fusion outperformed single source models in both overall classification and high risk detection. As shown in Table 1, the text only model achieved an Accuracy of 0.842, a Recall of 0.781, and an F1 of 0.801. The structured feature only model reached an Accuracy of 0.816, a Recall of 0.746, and an F1 of 0.772. These results suggest that single source data can capture part of the risk pattern, but their predictive power remains limited. After integrating submission texts, historical deficiencies, inspection results, and change records, the fusion model based on Random Forest achieved an Accuracy of 0.883 and an AUC of 0.910. The fusion XGBoost model delivered the best performance, with an Accuracy of 0.907, a Recall of 0.881, an F1 of 0.889, and an AUC of 0.934. Its high risk miss rate also dropped to 0.089, compared with 0.214 for the structured only model and 0.187 for the text only model. These findings show that the fused representation identified cumulative cross source risk more effectively.

DOI: 10.54254/2755-2721/2026.37646

Table 1. Performance comparison of the multi source fusion model and baseline models on the test set

Model	Accuracy	Precision	Recall	F1	AUC	High Risk Miss Rate
Structured only	0.816	0.758	0.746	0.772	0.861	0.214
Text only	0.842	0.793	0.781	0.801	0.884	0.187
Fusion Random Forest	0.883	0.861	0.844	0.852	0.910	0.121
Fusion BERT+MLP	0.896	0.872	0.866	0.874	0.922	0.102
Fusion XGBoost	0.907	0.896	0.881	0.889	0.934	0.089

4.2. Analysis of feature contribution and model interpretability

The interpretability analysis clarified the basis of model judgment. Based on the optimal XGBoost model, mean absolute SHAP values identified five major variables: historical deficiency frequency, regulatory terminology deviation rate, abnormal inspection ratio, manufacturing change density, and supplementary application count. As shown in Figure 1, their SHAP contributions were 0.214, 0.187, 0.161, 0.132, and 0.106, respectively, while the remaining features together contributed 0.200. This pattern suggests that high risk cases were more likely to involve repeated historical problems, regulatory wording deviation, and quality fluctuation at the same time. Medium risk cases were more often associated with terminology inconsistency and frequent supplementary applications. Combined with the classification results in Section 4.1, the strong predictive performance did not depend on one dominant variable. It resulted from the coordinated accumulation of several medium to high contribution features, which is consistent with the weighted fusion setting in Section 3.2.

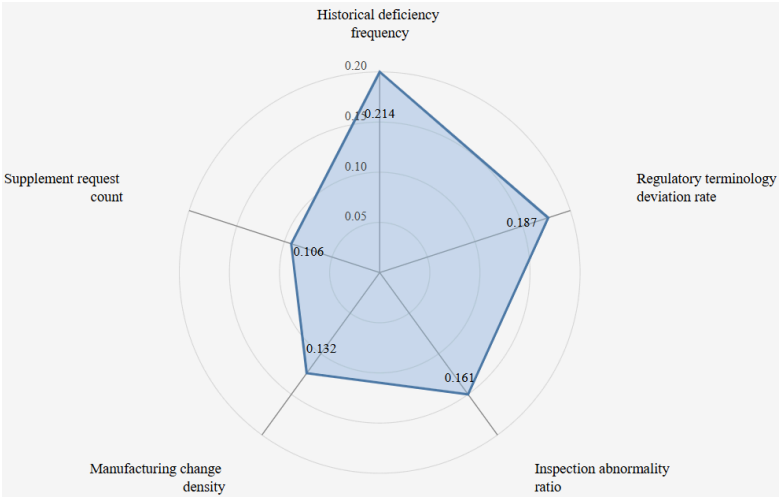

Figure 1. Contribution distribution of key risk features

5. Discussion

The results indicate that drug registration risk has a clear multi source cumulative structure. A single text source or a single structured record can explain only part of the risk pattern and cannot fully reflect the actual formation process. After submission texts, historical deficiencies, inspection results, and manufacturing changes were incorporated into one framework, the model showed a marked improvement in identifying high risk cases. This suggests that registration risk is more often

DOI: 10.54254/2755-2721/2026.37646

shaped by the joint accumulation of regulatory deviation, historical problems, and quality fluctuation. The interpretability analysis further showed that the dominant variables differ across risk levels, while high risk cases are more likely to present repeated deficiencies and concentrated abnormal indicators at the same time. Drug registration compliance assessment is therefore moving from static document checking toward dynamic evidence integration, and data fusion models can better support graded screening and focused review.

6. Conclusion

An intelligent analytical framework based on multi source heterogeneous data fusion was developed for quality risk prediction in drug registration compliance. Publicly available data were used to complete sample construction, feature fusion, model training, and result interpretation. The experimental results showed that the multi source fusion scheme outperformed single source models in both overall classification and high risk identification, indicating that complex registration risk is more effectively detected through cross source joint modeling. Feature contribution analysis further confirmed that historical deficiencies, regulatory terminology deviation, inspection abnormalities, and manufacturing changes are the core variables affecting risk judgment. Overall, this study provides a clear technical route for drug registration compliance management and establishes a methodological basis for future risk early warning research using larger scale real world business data.

References

[1] Patil, R. S., Kulkarni, S. B., & Gaikwad, V. L. (2023). Artificial intelligence in pharmaceutical regulatory affairs. *Drug Discovery Today, 28*(9), 103700.
[2] Vora, L. K., et al. (2023). Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics, 15*(7), 1916.
[3] Noorain, et al. (2023). Artificial intelligence in drug formulation and development: Applications and future prospects. *Current Drug Metabolism, 24*(9), 622–634.
[4] Verweij, S., et al. (2024). A natural language processing approach towards harmonized communication of uncertainties identified during the European Medicine Authorization Process. *Clinical Pharmacology & Therapeutics, 115*(4), 871–880.
[5] Nene, L., et al. (2024). Evolution of drug development and regulatory affairs: The demonstrated power of artificial intelligence. *Clinical Therapeutics, 46*(8), e6–e14.
[6] Fu, L., et al. (2025). The applications and advances of artificial intelligence in drug regulation: A global perspective. *Acta Pharmaceutica Sinica B, 15*(1), 1–14.
[7] Warraich, H. J., Tazbaz, T., & Califf, R. M. (2025). FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *Jama, 333*(3), 241–247.
[8] Matheny, M. E., et al. (2024). Enhancing postmarketing surveillance of medical products with large language models. *JAMA Network Open, 7*(8), e2428276.
[9] Dimitsaki, S., Natsiavas, P., & Jaulent, M.-C. (2024). Applying AI to structured real-world data for pharmacovigilance purposes: Scoping review. *Journal of Medical Internet Research, 26*, e57824.
[10] Algarvio, R. C., et al. (2025). Artificial intelligence in pharmacovigilance: A narrative review and practical experience with an expert-defined Bayesian network tool. *International Journal of Clinical Pharmacy, 47*(4), 932–944.
[11] Niazi, S. K. (2025). Regulatory perspectives for AI/ML implementation in pharmaceutical GMP environments. *Pharmaceuticals, 18*(6), 901.