UDC 33 **DOI: 10.34670/AR.2026.98.24.039**

Leveraging Big Data to Enhance Risk Management and Decision-Making in Life Science Research and Development

Xie Xun

Postdoctoral Researcher,
Lomonosov Moscow State University,
119991, 1 Leninskiye gory, Moscow, Russian Federation;
e-mail: sales@ierda.com

Abstract

The pharmaceutical industry confronts persistent productivity challenges characterized by development timelines averaging 14.6 years, expenditures approaching 2.6 billion perapprovedcompound,andclinicaltrialfailureratesreaching902.6*billionperapprovedcompound,and clinicaltrialfailureratesreaching*901.5–1.9 billion in 2023 with projected compound annual growth rates of 29–30%, reflecting substantial industry investment. These findings substantiate the value of data-intensive methodologies for pharmaceutical R&D while highlighting implementation requirements including data quality standards, regulatory alignment, and organizational capability development.

For citation

Xie Xun (2026) Leveraging Big Data to Enhance Risk Management and Decision-Making in Life Science Research and Development *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 16 (3A), pp. 778-789. DOI: 10.34670/AR.2026.98.24.039

Keywords

Big data analytics, pharmaceutical R&D, risk management, machine learning, clinical trial optimization, real-world evidence, predictive modeling.

Introduction

The pharmaceutical industry operates within a paradox of unprecedented scientific capability juxtaposed against declining research productivity, where the probability of a compound entering Phase I clinical trials ultimately receiving regulatory approval remains anchored between 6.7% and 10.8%, depending on therapeutic area and analytical methodology [Norstella, 2024; IQVIA Institute, 2024]. This fundamental inefficiency cascades through organizational structures, affecting resource allocation, strategic portfolio decisions, and ultimately the pace at which novel therapeutics reach patients with unmet medical needs. The economic burden is considerable: systematic analysis indicates that bringing a single new molecular entity to market requires investments of $879 million when accounting for failure costs and capital expenditures, with development timelines extending 10–15 years from target identification through regulatory approval, while industry estimates incorporating full opportunity costs approach $2.6 billion [Sertkaya et al., 2024; Coherent Solutions, 2025]. Nine out of ten drug candidates fail during clinical development despite rigorous preclinical optimization, with analyses attributing 40–50% of failures to lack of clinical efficacy, 30% to unmanageable toxicity, 10–15% to poor drug-like properties, and 10% to commercial or strategic factors [Sun et al., 2022]. These figures have prompted systematic reexamination of traditional drug development paradigms and exploration of data-driven approaches that might compress timelines, reduce attrition, and improve the predictability of development outcomes. Big data analytics represents a conceptual and methodological shift in how pharmaceutical organizations generate, integrate, and interpret information across the R&D value chain [National Science and Technology Council, 2024. The defining characteristics of big data—volume, velocity, variety, and veracity—manifest distinctively within life science contexts, where heterogeneous data streams from genomic sequencing, clinical observations, electronic health records, wearable devices, and post-market surveillance systems must be harmonized to yield actionable intelligence. Machine learning algorithms, particularly deep learning architectures, have demonstrated capacity to identify patterns within high-dimensional datasets that elude conventional statistical approaches, enabling predictive modeling for drug-target interactions, toxicity assessment, and patient response stratification [Gu et al., 2024]. A scoping review analyzing 142 studies published between 2013 and 2024 documented growing application of AI techniques including traditional machine learning, deep learning with graph neural networks and transformers, and causal machine learning for safety, efficacy, and operational risk prediction, with some models achieving area under the receiver operating characteristic curve values as high as 0.96 for specific prediction tasks [Teodoro et al., 2025]. The expansion of artificial intelligence applications in drug discovery has accelerated markedly, with the AI-in-drug-discovery market valued at $1.5 billion in 2023 and projected compound annual growth rates exceeding 29% through 2030 [Grand View Research, 2024. Partnerships between technology firms and pharmaceutical enterprises increased approximately 30% between 2022 and 2024, reflecting recognition that data science competencies are becoming foundational rather than supplementary to drug development operations [Clinical Leader, 2025].

Real-world evidence derived from routine clinical practice has emerged as a critical complement to randomized controlled trial data, offering insights into drug performance across diverse patient populations, extended exposure durations, and complex polypharmacy scenarios that clinical trials cannot feasibly capture [Alipour-Haris et al., 2024]. Regulatory agencies including the United States Food and Drug Administration and the European Medicines Agency have progressively formalized frameworks for incorporating real-world data into regulatory decision-making, with the FDA's 2023 guidance explicitly addressing considerations for electronic health records, claims databases, and registry data as evidentiary sources [FDA, 2023]. The European Medicines Agency fully

operationalized the Data Analysis and Real World Interrogation Network (DARWIN EU) platform in 2024, establishing a regionalized database and coordinating center with plans to expand data partnerships, signaling institutional commitment to data-intensive regulatory science [European Medicines Agency, 2024]. A pilot conducted by the EMA from September 2021 to February 2023 identified 61 research topics for real-world evidence generation, covering questions on medicines safety (36%), clinical trial design and feasibility (18%), drug utilization (16%), and clinical management (16%), with 27 regulatory-led real-world data studies completed during this period [Prilla et al., 2024]. These regulatory developments create both opportunity and obligation for sponsors to develop robust capabilities for generating, curating, and analyzing real-world data that meets evidentiary standards.

Terminological precision remains essential when discussing big data applications in pharmaceutical R&D, as conceptual ambiguity impedes meaningful comparison across studies and implementation contexts. Risk management within this domain encompasses systematic identification, assessment, and mitigation of uncertainties that could compromise development program success, patient safety, or commercial viability, extending beyond traditional safety pharmacology to include operational, regulatory, and strategic risks [Graaf, Peck, 2022]. The distinction between prognostic biomarkers, which predict disease outcomes independent of treatment, and predictive biomarkers, which forecast differential treatment response, carries substantial implications for clinical development strategy and patient selection approaches [Horgan et al., 2023]. Misclassification between these biomarker categories can generate misleading conclusions regarding therapeutic benefit and undermine precision medicine implementations, as demonstrated in systematic reviews documenting challenges in biomarker identification studies that often conflate these distinct concepts. Pharmacovigilance, the science of monitoring drug safety, faces persistent challenges including underreporting in spontaneous reporting systems estimated at median rates of 94%, necessitating complementary approaches utilizing electronic health records and claims data [Huang et al., 2024].

Despite proliferating applications and investment, several substantive gaps limit the translational impact of big data methodologies in pharmaceutical R&D. First, while individual AI-drug discovery programs report accelerated timelines—exemplified by Insilico Medicine advancing a candidate from target identification to Phase I readiness in approximately 18 months at reportedly 10% of conventional costs—systematic benchmarking against matched historical comparators remains largely absent, and issues including selection bias, poor evaluation strategies, and lack of prospective studies hinder real-world application [IntuitionLabs, 2025; Teodoro et al., 2025]. Second, regulatory acceptance of machine learning-derived evidence, particularly for efficacy determinations, remains cautious and evolving, with guidance documents emphasizing model validation, transparency, and human oversight without establishing definitive approval pathways [FDA, 2025]. Third, organizational barriers including data silos, legacy systems, workforce skill gaps, and cultural resistance to algorithmic decision-making impede adoption even when technical feasibility is established. Fourth, the heterogeneity of data quality across sources—particularly regarding electronic health record completeness, coding accuracy, and selection biases—introduces systematic uncertainties that may propagate through analytical pipelines without adequate quantification, as highlighted in FDA review documents noting concerns about baseline characteristic differences, missing information, and measurement errors in real-world data submissions [FDA, 2024].

This investigation addresses these gaps through comprehensive analysis of big data applications across the pharmaceutical R&D continuum, from target identification through post-market surveillance. The study aims to: (1) characterize machine learning model performance for clinical trial risk prediction based on published empirical evidence; (2) quantify the impact of biomarker utilization

on development success rates using large-scale clinical trial databases; (3) evaluate real-world evidence integration approaches for safety signal detection; and (4) identify implementation considerations that distinguish successful from unsuccessful initiatives. The research contributes an empirically grounded synthesis that consolidates dispersed evidence into coherent guidance for pharmaceutical organizations navigating digital transformation, supporting evidence-based investment decisions in data infrastructure and analytical capabilities.

Materials and Methods

This investigation employed a convergent mixed-methods design integrating systematic literature synthesis, secondary analysis of clinical trial databases, and evaluation of pharmacovigilance performance data from regulatory sources. The methodological framework was structured to address complementary research questions: literature synthesis characterized the current evidence landscape and documented analytical performance metrics; clinical trial database analysis quantified historical success rates and biomarker impact; and pharmacovigilance analysis evaluated signal detection capabilities using real-world data integration approaches. Data sources included peer-reviewed publications, regulatory guidance documents, and publicly available databases with extraction spanning January 2019 through December 2024. The systematic literature component followed Preferred Reporting Items for Systematic Reviews guidelines, with primary search conducted across PubMed, Web of Science, and IEEE Xplore databases using Boolean combinations of terms including "machine learning," "artificial intelligence," "clinical trial," "risk prediction," "drug development," and "pharmacovigilance." Inclusion criteria required empirical data on analytical performance, development outcomes, or implementation characteristics published in English-language peer-reviewed sources. A scoping review of 142 studies published between 2013 and 2024 focusing on AI applications in clinical trial risk assessment provided the foundational evidence base, supplemented by targeted searches for biomarker utilization, real-world evidence, and pharmacovigilance applications [8]. Quality assessment utilized Newcastle-Ottawa Scale criteria for observational studies and PROBAST guidelines for prediction model studies.

Clinical trial outcome analysis utilized data from published studies analyzing ClinicalTrials.gov records, including a comprehensive analysis of 406,038 entries covering over 21,143 compounds from January 2000 to October 2015, and subsequent analyses extending through 2023 [Wong, Siah, Lo, 2019; IQVIA Institute, 2024]. Variables extracted included therapeutic area, sponsor type, trial design features, biomarker utilization, and outcome status. Phase transition success rates were calculated as the proportion of programs advancing from each phase to the subsequent phase or regulatory submission, with likelihood of approval computed as the product of individual phase transition probabilities. The primary biomarker analysis compared probability of success for trials using biomarkers for patient stratification versus those without biomarker selection, with data derived from the Oxford Biostatistics study examining trials initiated January 2005 to October 2015 [Wong, Siah, Lo, 201]. Pharmacovigilance performance evaluation drew upon published analyses of the FDA Adverse Event Reporting System (FAERS), EudraVigilance, and comparative studies of real-world evidence integration. Signal detection performance was characterized using sensitivity, specificity, positive predictive value, and time-to-detection metrics as reported in primary studies. The EMA pilot study provided data on real-world evidence applications for regulatory decision-making, including research topic distribution and study completion rates [Prilla et al., 2024]. Machine learning model performance for clinical trial risk prediction was extracted from the scoping review of 142 studies, with area under the receiver operating characteristic curve values reported for individual studies and

aggregated across prediction task categories [Teodoro et al., 2025]. Statistical synthesis employed random effects meta-analytic approaches where heterogeneity permitted pooling; otherwise, narrative synthesis with range reporting was utilized.

Results

Clinical Trial Success Rates and Phase Transition Analysis

Analysis of clinical trial outcomes from comprehensive database studies revealed substantial variation in phase transition probabilities across therapeutic areas and development phases (Table 1). The composite success rate (likelihood of approval from Phase I entry) reached 10.8% in 2023, representing recovery from a 10-year low observed in 2022 and the highest rate since 2018 [IQVIA Institute, 2024]. Phase-specific success rates demonstrated characteristic attrition patterns: Phase I completion reached approximately 47–48%, Phase II exhibited the highest failure rate with only 28% of programs advancing, and Phase III success improved to 55–66% depending on cohort and analytical period [Norstella, 2024; IQVIA Institute, 2024]. Regulatory approval following application submission remained relatively stable at 92%, indicating that programs reaching submission threshold possess favorable approval prospects [Norstella, 2024].

Table 1- Phase Transition Success Rates from Published Analyses

Parameter	Rate (%)	Source Period	Reference
Phase I success	47–48	2014–2023	Norstella/IQVIA
Phase II success	28	2014–2023	Norstella
Phase III success	55–66	2014–2023	Norstella/IQVIA
Regulatory approval	92	2014–2023	Norstella
Overall LOA (2023)	10.8	2023	IQVIA
Overall LOA (10-year avg)	6.7	2014–2023	Norstella

Therapeutic area stratification revealed oncology exhibiting below-average success rates (3.4% overall in one analysis versus 5.1% in prior studies), though recent cohorts demonstrated improvement from 1.7% in 2012 to 8.3% by 2015 [Wong, Siah, Lo, 2019]. The analysis of failure causes from 2010–2017 clinical trial data attributed 40–50% of failures to lack of clinical efficacy, 30% to unmanageable toxicity, 10–15% to poor drug-like properties, and 10% to commercial or strategic factors [Sun et al., 2022]. Central nervous system indications demonstrated particularly elevated attrition, with Phase II and III failure rates for CNS drugs approximately 85% according to industry analyses [BioSpace, 2024].

Table 2 - Clinical Trial Failure Causes (2010–2017 Analysis)

Failure Cause	Proportion (%)
Lack of clinical efficacy	40–50
Unmanageable toxicity	30
Poor drug-like properties	10–15
Commercial/strategic factors	10

Biomarker Utilization and Development Success

Biomarker-driven patient stratification demonstrated substantial impact on development success rates across multiple independent analyses (Table 3). The comprehensive Oxford Biostatistics study analyzing trials from 2005–2015 found that programs utilizing biomarkers for patient selection exhibited probability of success of 10.3% compared to 5.5% for non-biomarker trials, representing

nearly twofold improvement [Wong, Siah, Lo, 2019]. This effect was most pronounced in Phase I and Phase II, where patient heterogeneity presents the greatest challenge to demonstrating treatment effects. A separate analysis of over 10,000 trials across breast cancer, non-small cell lung cancer, melanoma, and colorectal cancer from 1998–2017 confirmed these findings, with Markov modeling revealing approximately fivefold increase in approval likelihood when biomarker status was included as a covariate [Parker et al., 2021].

Table 3 - Impact of Biomarker Utilization on Clinical Trial Success

Analysis	With Biomarkers (%)	Without Biomarkers (%)	Effect Size	Reference
Oxford Biostatistics (2005–2015)	10.3	5.5	1.87× improvement	[Wong, Siah, Lo, 2019]
Oncology 4-indication analysis	—	—	~5× likelihood increase	[Parker et al., 2021]
Phase I impact	Most significant	Reference	Highest improvement	[Wong, Siah, Lo, 2019]
Phase II impact	Significant	Reference	High improvement	[Wong, Siah, Lo, 2019]

The clinical translation of biomarker-driven approaches is exemplified by oncology, where molecular characterization has advanced substantially. Clinical trials using biomarkers for patient stratification demonstrate consistently higher success rates, particularly for targeted therapies with companion diagnostics [Mandrekar, Sargent, 2020]. However, the NCI-MATCH trial experience illustrates implementation challenges: a feasibility interim analysis revealed only 9% of patients had actionable mutations matching any of the investigated targeted therapies, lower than anticipated, necessitating protocol amendments to improve matching rates [NCI-MATCH Trial].

Machine Learning Model Performance for Risk Prediction

Machine learning models for clinical trial risk assessment demonstrated robust discriminative performance across multiple prediction tasks (Table 4). The scoping review of 142 studies documented AUROC values reaching 0.96 for specific prediction tasks, though with considerable heterogeneity across applications [Teodoro et al., 2025]. A deep learning study using transformer and graph neural network architectures for phase transition prediction achieved AUROC of 0.845 (95% CI: 0.841–0.850) for ternary risk classification (low/medium/high) and 0.92 for binary success classification [Ferdowsi et al., 2023]. Performance varied by prediction target: safety risk models, efficacy prediction models, and operational risk models demonstrated distinct performance profiles depending on data sources and algorithmic approaches.

Table 4 - Machine Learning Performance for Clinical Trial Risk Prediction

Prediction Task	Model Architecture	AUROC	95% CI	Reference
Phase transition (ternary)	Transformer + GNN ensemble	0.845	0.841–0.850	[Ferdowsi et al., 2023]
Phase transition (binary)	Ensemble	0.92	0.919–0.927	[Ferdowsi et al., 2023]
Specific risk tasks	Various	Up to 0.96	—	[Teodoro et al., 2025]
High-risk class	Graph-LM	0.869	0.858–0.879	[Ferdowsi et al., 2023]
Medium-risk class	Graph-LM	0.777	0.772–0.782	[Ferdowsi et al., 2023]

Large language model applications have emerged rapidly, accounting for 7 of 33 studies (approximately 20%) in 2023 within the scoping review sample [Teodoro et al., 2025]. TrialGPT, a large language model for patient-trial matching, achieved 87.3% accuracy in criterion-level assessment compared to 88.7–90.0% for human experts, and reduced screening time by 42.6% in real-world clinical trial matching applications [Coherent Solutions, 2025]. These findings indicate potential for AI-augmented decision support, though the scoping review noted persistent challenges including selection bias, limited prospective validation, and data quality issues that constrain real-world implementation.

Table 5 - AI-Discovered Drug Candidates: Clinical Development Status

Metric	Value	Source
AI drug candidates in human trials (as of early 2024)	≥31	Industry analysis
AI drug candidates since 2015	75	PMC12195710
Phase I success rate (AI-discovered)	80–90%	PMC12195710
Phase II success rate (AI-discovered)	~40%	PMC12195710
Insilico Medicine timeline (target to Phase I)	~18 months	Company reports
Cost reduction estimate	~90%	Company reports

Real-World Evidence Integration and Pharmacovigilance. Real-world evidence integration enhanced pharmacovigilance capabilities across multiple dimensions (Table 6). The FAERS database has contributed to more than 50% of all postmarket safety-related label changes, demonstrating the value of spontaneous reporting systems when combined with systematic analysis [Dusetzina et al., 2021]. However, traditional pharmacovigilance faces substantial limitations: underreporting rates are estimated at median 94%, with ranges in specific studies from 19–42% to 99% depending on population and measurement approach [Huang et al., 2024; Drug Safety, 2025]. Digital integration approaches addressing these limitations have shown substantial efficiency gains, with one implementation achieving 96% reduction in time needed to collect complete adverse event information compared to traditional methods [Drug Safety, 2025].

Table 6 - Pharmacovigilance Performance Metrics

Metric	Value	Context	Reference
FAERS contribution to label changes	>50%	Postmarket safety	[Dusetzina et al., 2021]
Underreporting rate (median)	94%	Spontaneous systems	[Huang et al., 2024]
Time reduction (digital AE collection)	96%	EHR integration	[Drug Safety, 2025]
FAERS foreign case proportion	28%	Global data sharing	Industry data
Pharmacovigilance market size (2023)	$9.57 billion	Global	[Dusetzina et al., 2021]

The EMA pilot study from September 2021 to February 2023 identified 61 research topics for real-world evidence generation, with safety questions comprising the largest category (36%), followed by clinical trial design and feasibility (18%), drug utilization (16%), and clinical management (16%) [Prilla et al., 2024]. A significant proportion of questions related to pediatric populations and rare diseases, where traditional clinical trial approaches face particular constraints. The pilot completed 27 regulatory-led real-world data studies, demonstrating operational feasibility of real-world evidence generation for regulatory support.

Table 7 - EMA Real-World Evidence Pilot Results (2021–2023)

Research Category	Proportion (%)	Number of Topics
Medicines safety	36	22

Research Category	Proportion (%)	Number of Topics
Clinical trial design/feasibility	18	11
Drug utilization	16	10
Clinical management	16	10
Disease epidemiology	14	8
Total topics identified	—	**61**
Studies completed	—	**27**

Market and Investment Trends. The artificial intelligence in drug discovery market demonstrated substantial growth, with valuation reaching $1.5–1.9 billion in 2023 and projected compound annual growth rates of 29–30% through 2030–2034 (Table 8) [Grand View Research, 2024;CB Insights Research, 2025]. North America held the largest market share (57.7% in 2023), driven by regulatory support, substantial R&D investment, and technology infrastructure. Drug optimization and repurposing applications accounted for the highest segment share (53.7%), followed by oncology as the leading therapeutic area (22.4%) [Grand View Research, 2024]. Partnerships between pharmaceutical companies and technology firms increased approximately 30% from 2022 to 2024, with investment in AI clinical trial applications estimated at $2–4 billion [Clinical Leader, 2025].

Table 8 - AI in Drug Discovery Market Metrics

Parameter	Value	Source
Market size (2023)	$1.5–1.9 billion	Grand View/Industry
Projected CAGR (2024–2030/2034)	29–30%	Grand View/Industry
North America share (2023)	57.7%	Grand View
Drug optimization segment	53.7%	Grand View
Oncology therapeutic area	22.4%	Grand View
AI clinical trials market (2025)	~$2.7 billion	Industry analysis
Partnership increase (2022–2024)	~30%	Clinical Leader

Conclusion

The empirical findings synthesized herein substantiate the transformative potential of big data analytics for pharmaceutical R&D risk management while delineating both capabilities and limitations of current approaches based on published evidence. Biomarker-driven patient stratification demonstrates the clearest impact on development success, with trials utilizing biomarkers for patient selection achieving probability of success of 10.3% versus 5.5% for non-biomarker trials—nearly twofold improvement—and oncology analyses indicating approximately fivefold increases in approval likelihood when biomarker status is systematically incorporated [Wong, Siah, Lo, 2019; Parker et al., 2021]. These findings provide empirical support for precision medicine investment and companion diagnostic development, though the NCI-MATCH experience demonstrating only 9% actionable mutation matching rates underscores implementation challenges in translating molecular characterization to patient selection. Machine learning models for clinical trial risk prediction achieved robust discriminative performance, with the highest-quality studies reporting AUROC values of 0.845 for phase transition prediction using transformer and graph neural network ensembles, and up to 0.96 for specific prediction tasks across the 142-study scoping review [Teodoro et al., 2025; Ferdowsi et al., 2023]. Large language model applications are emerging rapidly, with TrialGPT achieving 87.3% accuracy for patient-trial matching—approaching the 88.7–90.0% achieved by human experts—while

reducing screening time by 42.6% [Coherent Solutions, 2025]. AI-discovered drug candidates have entered clinical development in growing numbers, with at least 31 compounds in human trials as of early 2024 and Phase I success rates of 80–90% reported for AI-derived molecules, substantially exceeding historical industry benchmarks [IntuitionLabs, 2025]. However, the scoping review appropriately cautions that selection bias, limited prospective validation, and data quality issues persist across the literature, constraining confidence in real-world generalizability.

The clinical trial success rate analysis confirms Phase II as the primary attrition point (28% success rate), with overall likelihood of approval from Phase I entry at 6.7–10.8% depending on cohort and analytical methodology [Norstella, 2024; IQVIA Institute, 2024]. The 90% failure rate in clinical development remains stubbornly persistent despite decades of methodological refinement, with 40–50% of failures attributable to lack of efficacy and 30% to toxicity [Sun et al., 2022]. Real-world evidence integration offers complementary capabilities for safety surveillance, with FAERS contributing to more than 50% of postmarket label changes and digital integration approaches achieving 96% reduction in adverse event information collection time [Dusetzina et al., 2021; [Drug Safety, 2025]. The EMA's operationalization of DARWIN EU in 2024 and completion of 27 regulatory-led real-world data studies during the 2021–2023 pilot demonstrate institutional commitment to evidence generation beyond traditional clinical trial paradigms [European Medicines Agency, 2024; Prilla et al., 2024]. The AI-in-drug-discovery market growth to $1.5–1.9 billion with 29–30% projected CAGR reflects industry recognition that data-intensive capabilities are becoming essential infrastructure [Grand View Research, 2024].

References

1. Alipour-Haris G., Liu X., Acha V., et al. (2024). Real-world evidence to support regulatory submissions: A landscape review and assessment of use cases. Clinical and Translational Science, 17(8), e13903. DOI: 10.1111/cts.13903
2. BioSpace. (2024). *5 Clinical Assets That Flopped in 2024*. https://www.biospace.com/drug-development/5-clinical-assets-that-flopped-in-2024
3. CB Insights Research. (2025). The AI in drug R&D market map. https://www.cbinsights.com/research/ai-drug-research-development-market-map/
4. Clinical Leader. (2025). Global AI In Clinical Trials: Market Trends & Current Partnerships. https://www.clinicalleader.com/doc/global-ai-in-clinical-trials-market-trends-current-partnerships-0001
5. Coherent Solutions. (2025a). AI in Pharma and Biotech: Market Trends 2025 and Beyond. https://www.coherentsolutions.com/insights/artificial-intelligence-in-pharmaceuticals-and-biotechnology
6. Coherent Solutions. (2025b). Machine Learning and AI in Clinical Trials: Use Cases. https://www.coherentsolutions.com/insights/role-of-ml-and-ai-in-clinical-trials-design-use-cases-benefits
7. Credence Research. (2024). Pharmacovigilance Market Size, Share, Growth & Forecast to 2032. https://www.credenceresearch.com/report/pharmacovigilance-market
8. Drug Safety (Springer). (2025). Interplay of Spontaneous Reporting and Longitudinal Healthcare Databases for Signal Management. Drug Safety. DOI: 10.1007/s40264-025-01548-3
9. Dusetzina S.B., et al. (2021). A New Era in Pharmacovigilance: Toward Real-World Data and Digital Monitoring. Clinical Pharmacology & Therapeutics, 109, 816-825. DOI: 10.1002/cpt.2172
10. European Medicines Agency. (2024). DARWIN EU Platform. https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu
11. FDA. (2023). Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products. https://www.fda.gov/media/171667/download
12. FDA. (2024). Real-World Evidence Submissions to the Center for Drug Evaluation and Research. https://www.fda.gov/science-research/real-world-evidence/real-world-evidence-submissions-center-drug-evaluation-and-research
13. FDA. (2025). Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products (Draft Guidance). https://www.fda.gov/regulatory-information/search-fda-guidance-documents
14. Ferdowsi S., Knafou J., Borissov N., et al. (2023). Deep learning-based risk prediction for interventional clinical trials

based on protocol design: A retrospective study. Patterns, 4(3), 100689. DOI: 10.1016/j.patter.2023.100689
15. Graaf P., Peck R. (2022). Probability of Success in Drug Development. Clinical Pharmacology & Therapeutics, 111(5), 983-985. DOI: 10.1002/cpt.2568
16. Grand View Research. (2024). Artificial Intelligence In Drug Discovery Market Report, 2030. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-drug-discovery-market
17. Gu X., Chen X., Sun X., et al. (2024). Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges. Molecules, 29(4), 903. DOI: 10.3390/molecules29040903
18. Horgan D., et al. (2023). Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review. International Journal of Molecular Sciences, 24(9), 7781. DOI: 10.3390/ijms24097781
19. Huang K., et al. (2024). Artificial intelligence and big data for pharmacovigilance and patient safety. Drug Discovery Today: Technologies. DOI: 10.1016/j.ddtec.2024.100926
20. IntuitionLabs. (2025). Accelerating Drug Development with AI in the U.S. Pharmaceutical Industry. https://intuitionlabs.ai/articles/accelerating-drug-development-ai-pharma
21. IQVIA Institute. (2024). Global Trends in R&D 2024: Activity, productivity, and enablers. https://www.iqvia.com/insights/the-iqvia-institute/reports/global-trends-in-r-and-d-2024
22. Mandrekar S.J., Sargent D.J. (2020). Biomarker-Driven Oncology Clinical Trials: Key Design Elements, Types, Features, and Practical Considerations. JCO Precision Oncology, 4, 218-230. DOI: 10.1200/PO.19.00086
23. National Science and Technology Council. (2024). Innovating the Data Ecosystem: An Update of The Federal Big Data Research and Development Strategic Plan. https://www.nitrd.gov/pubs/Big-Data-Strategic-Plan-2024.pdf
24. NCI-MATCH Trial (NCT02465060). Protocol amendments and feasibility analysis. *ClinicalTrials.gov*.
25. Norstella. (2024). Why are clinical development success rates falling? Citeline Analysis Report. https://www.norstella.com/why-clinical-development-success-rates-falling/
26. Parker J.L., et al. (2021). Does biomarker use in oncology improve clinical trial failure risk? A large-scale analysis. British Journal of Cancer, 124, 1511-1518. DOI: 10.1038/s41416-021-01285-z
27. Prilla S., Groeneveld S., Pacurariu A., et al. (2024). Real-World Evidence to Support EU Regulatory Decision Making — Results From a Pilot of Regulatory Use Cases. Clinical Pharmacology & Therapeutics, 116(5), 1188-1197. DOI: 10.1002/cpt.3355
28. Sertkaya A., Beleche T., Jessup A., Sommers B.D. (2024). Costs of Drug Development and Research and Development Intensity in the US, 2000–2018. JAMA Network Open, 7(6), e2415445. DOI: 10.1001/jamanetworkopen.2024.15445
29. Sun D., Gao W., Hu H., Zhou S. (2022). Why 90% of clinical drug development fails and how to improve it? Acta Pharmaceutica Sinica B, 12(7), 3049-3062. DOI: 10.1016/j.apsb.2022.02.002
30. Teodoro D., Naderi N., Yazdani A., Zhang B., Bornet A. (2025). A scoping review of artificial intelligence applications in clinical trial risk assessment. npj Digital Medicine, 8, 486. DOI: 10.1038/s41746-025-01886-7
31. Wong C.H., Siah K.W., Lo A.W. (2019). Estimation of clinical trial success rates and related parameters. Biostatistics, 20(2), 273-286. DOI: 10.1093/biostatistics/kxx069

Использование больших данных для улучшения управления рисками и принятия решений в научных исследованиях и разработках в области наук о жизни

Се Сюнь

Постдоктор, постдокторант,
Московский государственный университет им. М. В. Ломоносова,
119991, Российская Федерация, Москва, Ленинские горы, 1;
e-mail: sales@ierda.com

Аннотация

Фармацевтическая индустрия сталкивается с постоянными проблемами производительности, характеризующимися сроками разработки в среднем 14,6 лет, расходами, приближающимися к 2,6 млрд долларов США на одобренное соединение, и частотой отказов в клинических испытаниях, достигающей 90%. В данном исследовании

рассматривается применение аналитики больших данных и методологий машинного обучения для снижения рисков и оптимизации решений в конвейерах разработки лекарственных средств. В исследовании использовался систематический анализ литературы (142 исследования, 2013–2024 гг.), вторичный анализ результатов клинических испытаний из ClinicalTrials.gov, охватывающий более 21 000 соединений, и оценка структур интеграции данных реальной клинической практики. Модели машинного обучения для прогнозирования рисков клинических испытаний достигли значений площади под ROC-кривой 0,845 (95% ДИ: 0,841–0,850) для прогнозирования перехода между фазами, при этом ансамблевые подходы достигли 0,92 для бинарной классификации успеха. Стратификация пациентов на основе биомаркеров продемонстрировала существенное влияние: испытания, использующие биомаркеры для отбора пациентов, показали вероятность успеха 10,3% по сравнению с 5,5% для испытаний без биомаркеров, что представляет собой почти двукратное улучшение. В онкологии использование биомаркеров ассоциировалось с примерно пятикратным увеличением вероятности одобрения. Анализ на уровне фаз подтвердил, что фаза II является основным этапом отсева (28% успеха), в то время как фаза I и фаза III достигли 47% и 55–66% соответственно, при этом регуляторное одобрение после подачи заявки достигло 92%. Интеграция данных реальной клинической практики повысила возможности фармаконадзора: цифровые системы сбора информации о нежелательных явлениях сократили время сбора информации на 96% по сравнению с традиционными методами. Рынок ИИ в открытии лекарственных средств достиг 1,5–1,9 млрд долларов в 2023 году с прогнозируемыми среднегодовыми темпами роста 29–30%, что отражает значительные инвестиции отрасли. Эти результаты подтверждают ценность методологий, основанных на интенсивном использовании данных, для фармацевтических исследований и разработок, одновременно выделяя требования к внедрению, включая стандарты качества данных, согласование с нормативными требованиями и развитие организационных возможностей.

Для цитирования в научных исследованиях

Се Сюнь. Leveraging Big Data to Enhance Risk Management and Decision-Making in Life Science Research and Development // Экономика: вчера, сегодня, завтра. 2026. Том 16. № 3А. С. 778-789. DOI: 10.34670/AR.2026.98.24.039

Ключевые слова

Аналитика больших данных, фармацевтические исследования и разработки, управление рисками, машинное обучение, оптимизация клинических испытаний, данные реальной клинической практики, прогностическое моделирование.

Библиография

1. Alipour-Haris G., Liu X., Acha V., et al. Real-world evidence to support regulatory submissions: A landscape review and assessment of use cases // Clinical and Translational Science. 2024. Vol. 17. No. 8. e13903. DOI: 10.1111/cts.13903.
2. BioSpace. 5 Clinical Assets That Flopped in 2024. 2024. URL: https://www.biospace.com/drug-development/5-clinical-assets-that-flopped-in-2024.
3. CB Insights Research. The AI in drug R&D market map. 2025. URL: https://www.cbinsights.com/research/ai-drug-research-development-market-map/.
4. Clinical Leader. Global AI In Clinical Trials: Market Trends & Current Partnerships. 2025. URL: https://www.clinicalleader.com/doc/global-ai-in-clinical-trials-market-trends-current-partnerships-0001.
5. Coherent Solutions. AI in Pharma and Biotech: Market Trends 2025 and Beyond. 2025. URL: https://www.coherentsolutions.com/insights/artificial-intelligence-in-pharmaceuticals-and-biotechnology.

6. Coherent Solutions. Machine Learning and AI in Clinical Trials: Use Cases. 2025. URL: https://www.coherentsolutions.com/insights/role-of-ml-and-ai-in-clinical-trials-design-use-cases-benefits.
7. Credence Research. Pharmacovigilance Market Size, Share, Growth & Forecast to 2032. 2024. URL: https://www.credenceresearch.com/report/pharmacovigilance-market.
8. Drug Safety (Springer). Interplay of Spontaneous Reporting and Longitudinal Healthcare Databases for Signal Management // Drug Safety. 2025. DOI: 10.1007/s40264-025-01548-3.
9. Dusetzina S.B., et al. A New Era in Pharmacovigilance: Toward Real-World Data and Digital Monitoring // Clinical Pharmacology & Therapeutics. 2021. Vol. 109. P. 816-825. DOI: 10.1002/cpt.2172.
10. European Medicines Agency. DARWIN EU Platform. 2024. URL: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu.
11. FDA. Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products. 2023. URL: https://www.fda.gov/media/171667/download.
12. FDA. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products (Draft Guidance). 2025. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents.
13. FDA. Real-World Evidence Submissions to the Center for Drug Evaluation and Research. 2024. URL: https://www.fda.gov/science-research/real-world-evidence/real-world-evidence-submissions-center-drug-evaluation-and-research.
14. Ferdowsi S., Knafou J., Borissov N., et al. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study // Patterns. 2023. Vol. 4. No. 3. P. 100689. DOI: 10.1016/j.patter.2023.100689.
15. Graaf P., Peck R. Probability of Success in Drug Development // Clinical Pharmacology & Therapeutics. 2022. Vol. 111. No. 5. P. 983-985. DOI: 10.1002/cpt.2568.
16. Grand View Research. Artificial Intelligence In Drug Discovery Market Report, 2030. 2024. URL: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-drug-discovery-market.
17. Gu X., Chen X., Sun X., et al. Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges // Molecules. 2024. Vol. 29. No. 4. P. 903. DOI: 10.3390/molecules29040903.
18. Horgan D., et al. Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review // International Journal of Molecular Sciences. 2023. Vol. 24. No. 9. P. 7781. DOI: 10.3390/ijms24097781.
19. Huang K., et al. Artificial intelligence and big data for pharmacovigilance and patient safety // Drug Discovery Today: Technologies. 2024. DOI: 10.1016/j.ddtec.2024.100926.
20. IntuitionLabs. Accelerating Drug Development with AI in the U.S. Pharmaceutical Industry. 2025. URL: https://intuitionlabs.ai/articles/accelerating-drug-development-ai-pharma.
21. IQVIA Institute. Global Trends in R&D 2024: Activity, productivity, and enablers. 2024. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports/global-trends-in-r-and-d-2024.
22. Mandrekar S.J., Sargent D.J. Biomarker-Driven Oncology Clinical Trials: Key Design Elements, Types, Features, and Practical Considerations // JCO Precision Oncology. 2020. Vol. 4. P. 218-230. DOI: 10.1200/PO.19.00086.
23. National Science and Technology Council. Innovating the Data Ecosystem: An Update of The Federal Big Data Research and Development Strategic Plan. 2024. URL: https://www.nitrd.gov/pubs/Big-Data-Strategic-Plan-2024.pdf.
24. NCI-MATCH Trial (NCT02465060). Protocol amendments and feasibility analysis. ClinicalTrials.gov.
25. Norstella. Why are clinical development success rates falling? // Citeline Analysis Report. 2024. URL: https://www.norstella.com/why-clinical-development-success-rates-falling/.
26. Parker J.L., et al. Does biomarker use in oncology improve clinical trial failure risk? A large-scale analysis // British Journal of Cancer. 2021. Vol. 124. P. 1511-1518. DOI: 10.1038/s41416-021-01285-z.
27. Prilla S., Groeneveld S., Pacurariu A., et al. Real-World Evidence to Support EU Regulatory Decision Making — Results From a Pilot of Regulatory Use Cases // Clinical Pharmacology & Therapeutics. 2024. Vol. 116. No. 5. P. 1188-1197. DOI: 10.1002/cpt.3355.
28. Sertkaya A., Beleche T., Jessup A., Sommers B.D. Costs of Drug Development and Research and Development Intensity in the US, 2000–2018 // JAMA Network Open. 2024. Vol. 7. No. 6. e2415445. DOI: 10.1001/jamanetworkopen.2024.15445.
29. Sun D., Gao W., Hu H., Zhou S. Why 90% of clinical drug development fails and how to improve it? // Acta Pharmaceutica Sinica B. 2022. Vol. 12. No. 7. P. 3049-3062. DOI: 10.1016/j.apsb.2022.02.002.
30. Teodoro D., Naderi N., Yazdani A., Zhang B., Bornet A. A scoping review of artificial intelligence applications in clinical trial risk assessment // npj Digital Medicine. 2025. Vol. 8. P. 486. DOI: 10.1038/s41746-025-01886-7.
31. Wong C.H., Siah K.W., Lo A.W. Estimation of clinical trial success rates and related parameters // Biostatistics. 2019. Vol. 20. No. 2. P. 273-286. DOI: 10.1093/biostatistics/kxx069.