



Разработка биостатистической методологии для анализа данных реальной клинической практики при оценке биологически активных добавок

Идриса К.¹, Холопов Р. А.², Драгунс В.², Булкрам М. С.¹

¹ ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Российская Федерация

² Санкт-Петербургский колледж телекоммуникаций им. Э.Т. Кренкеля — структурное подразделение ФГБОУ ВО «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича», Санкт-Петербург, Российская Федерация

³ ООО «Экс Севен Ресеч», Санкт-Петербург, Российская Федерация

Аннотация

Цель. Разработка и валидация гибридной биостатистической методологии для генерации доказательств, основанных на данных реальной клинической практики (RWE) при оценке эффективности и безопасности биологически активных добавок (БАД).

Методы. Предложенная концепция объединяет методы причинно-следственного вывода (IPTW) с ансамблевыми методами машинного обучения (CatBoost, Random Forest) и множественным восполнением пропущенных данных (MICE). Валидация проводилась на двух пилотных неинтервенционных исследованиях (N=249) и высококоразмерном синтетическом наборе данных (N=6000) с уровнем пропусков до 83,33%. Для оценки соотношения польза—риск использовалась многокритериальная Парето-оптимизация.

Результаты. Применение IPTW позволило достичь стандартизированной разницы средних (SMD) <0,10 по всем исходным характеристикам. Регрессионная модель эффективности достигла $R^2 = 0,3596$. Модель классификации безопасности на основе CatBoost после оптимизации порога классификации показала $PR_{AUC} = 0,123$ и F1-меру = 0,218, что позволило выявить скрытые сигналы нежелательных явлений (например, 12,47% для магния 400 мг). Парето-оптимизация обеспечила объективное сравнение профилей БАД без субъективного взвешивания показателей.

Заключение. Разработанная методология повышает согласованность и надёжность пострегистрационной оценки БАД на основе RWD и может быть использована регуляторными органами (FDA, EFSA, Росздравнадзор РФ).

Ключевые слова: доказательства, основанные на данных реальной клинической практики; пищевые добавки; причинно-следственный вывод; машинное обучение; Парето-оптимизация; обнаружение сигналов; биостатистика

Для цитирования: Идриса К., Холопов Р. А., Драгунс В., Булкрам М. С. Разработка биостатистической методологии для анализа данных реальной клинической практики при оценке биологически активных добавок. *Реальная клиническая практика: данные и доказательства*. 2026;6(2):19-32. <https://doi.org/10.37489/2782-3784-myrwd-101>. EDN: ZTDIKD.

Поступила: 19.04.2026. В доработанном виде: 20.05.2026. Принята к печати: 26.05.2026. Опубликовано: 30.06.2026.

Development and validation of a biostatistical framework for consistent evaluation of food supplement efficacy and safety using real-world data

Kiryowa Idrisa¹, Ruslan A. Kholopov², Vitaly Draguns³, Mohamed Said Boulkrane¹

¹ ITMO University, St. Petersburg, Russian Federation

² St. Petersburg College of Telecommunications named after E.T. Krenkel is a structural division of the St. Petersburg State University of Telecommunications named after prof. M.A. Bonch-Bruевич, St. Petersburg, Russian Federation

³ X7 Research LLC, St. Petersburg, Russian Federation

Abstract

Objective. To develop and validate a hybrid biostatistical methodology for generating Real-World Evidence (RWE) to assess the efficacy and safety of food supplements.

Methods. The proposed framework integrates causal inference techniques (Inverse Probability of Treatment Weighting, IPTW) with ensemble machine learning methods (CatBoost, Random Forest) and Multiple Imputation by Chained Equations (MICE). Validation was performed using two pilot non-interventional studies (N=249) and a high-dimensional synthetic dataset (N=6000) with missingness up to 83.33%. Multi-objective Pareto optimization was applied for benefit–risk assessment.

Results. IPTW achieved a standardized mean difference (SMD) <0.10 across all baseline covariates. The efficacy regression model attained an R^2 of 0.3596. The CatBoost-based safety classification model, after threshold optimization, achieved a PR_{AUC} of 0.123 and an F1-score of 0.218, enabling the detection of latent adverse event signals (e.g., 12.47% for Magnesium 400 mg). Pareto optimization provided an objective comparison of supplement profiles without subjective indicator weighting.

Conclusion. The developed methodology improves the consistency and reliability of post-market supplement evaluation using RWD and can support regulatory activities of organizations such as Food and Drug Administration, European Food Safety Authority, and Roszdravnadzor RF.

Keywords: real-world evidence; food supplements; causal inference; machine learning; Pareto optimization; signal detection; biostatistics

For citation: Idrisa K, Kholopov RA, Draguns V, Boulkrane MS. Development and validation of a biostatistical framework for consistent evaluation of food supplement efficacy and safety using real-world data. *Real-World Data & Evidence*. 2026;6(2):19-32. <https://doi.org/10.37489/2782-3784-myrrwd-101>. EDN: ZTDIKD.

Received: 19.04.2026. **Revision received:** 20.05.2026. **Accepted:** 26.05.2026. **Published:** 30.06.2026.

Introduction

The global market for food supplements has expanded rapidly over the past decade, reflecting growing public interest in preventive healthcare and long-term wellness management. Vitamins, minerals, probiotics, botanical extracts, and nutraceutical formulations have become part of routine health practices for millions of consumers worldwide. Despite the scale of this industry, regulatory supervision remains inconsistent across many countries, creating difficulties in assessing product quality, clinical efficacy, and long-term safety [1]. Differences between commercial availability and scientific validation continue to generate concerns regarding reliability of health claims and post-market monitoring standards. Randomized Controlled Trials (RCTs) traditionally remain the primary method for evaluating medical interventions. However, their application to food supplements is associated with several practical and methodological limitations. Such trials require considerable financial resources, lengthy implementation periods, and strict participant selection criteria [2]. In many cases, individuals with comorbidities, irregular lifestyles, or concurrent medication use are excluded from participation. As a result, highly controlled trial conditions often fail to reflect the variability of real consumers and everyday clinical practice. This reduces external validity and complicates translation of trial outcomes into broader population settings.

Against this background, Real-World Data (RWD) has emerged as an important source of evidence for healthcare analytics and regulatory evaluation [3]. RWD includes information collected outside conventional clinical trials, including electronic health records, observational registries, insurance databases, wearable device outputs, and patient-reported outcomes. After appropriate statistical processing, these datasets can generate Real-World Evidence (RWE) capable of supporting practical decision-making and post-market surveillance

activities [4]. Regulatory organizations such as the Food and Drug Administration (FDA) and the European Food Safety Authority (EFSA) have increasingly incorporated RWE into modern evaluation frameworks [5].

At the same time, analysis of RWD presents substantial methodological challenges. Unlike controlled experimental datasets, real-world information is frequently incomplete, heterogeneous, and structurally inconsistent. Missing observations, coding variability, measurement inaccuracies, and differences in reporting standards complicate statistical interpretation. In addition, observational data is highly susceptible to confounding by indication and selection bias, particularly when treatment allocation is influenced by patient condition or physician preference [6]. Conventional regression approaches often fail to capture these complexities adequately, especially in high-dimensional environments with multiple interacting variables. Another significant issue involves the persistence of “clinical rigidity” within observational supplement studies. Even non-interventional investigations frequently rely on restrictive inclusion criteria that artificially simplify patient populations. Individuals with severe symptoms, multiple chronic conditions, or inconsistent treatment adherence are commonly excluded. Consequently, the resulting evidence may not accurately represent real consumer behavior or actual safety risks encountered during routine supplement use [7].

Without standardized analytical frameworks capable of accounting for such variability, conclusions derived from RWD may remain unstable and potentially misleading. Within the thesis, a biostatistical methodology for standardized analysis of high-dimensional Real-World Data was developed and validated. The proposed framework integrates causal inference techniques with ensemble machine learning models to improve consistency of efficacy and safety assessment

Table 1. Comparative characteristics of biostatistical approaches used for Real-World Evidence generation

Biostatistical Approach	Methodological Strength	Methodological Weakness	Primary RWE Application	Ref.
Traditional Regression (Linear, Logistic, ANOVA)	Computationally straightforward; provides easily interpretable baseline effect sizes.	Highly vulnerable to unmeasured confounding; unable to model complex, non-linear interactions.	Basic efficacy and safety assessments in low-complexity datasets.	[8, 11]
Causal Inference (IPTW & PSM)	Balances observed covariates between groups; creates an unbiased “pseudo-population” to mitigate confounding.	Relies on the untestable assumption of “no unmeasured confounding”; sensitive to extreme propensity weights.	Target trial emulation; mathematically adjusting cohorts to simulate an RCT.	[13, 14]
Ensemble Machine Learning (CatBoost, Random Forests)	Captures deep, non-linear interactions without manual feature engineering; highly effective for imbalanced data.	Inherently susceptible to “black-box” amplification of observational bias if data is not causally balanced first.	Predictive modeling; proactive detection of rare adverse event signals.	[8, 15]

for food supplements. Attention was devoted to reducing observational bias and improving reliability of post-market evidence generation. The methodological novelty of the thesis lies in combining Propensity Score Weighting methods with advanced machine learning architectures, including CatBoost and Random Forest algorithms [8, 9]. This integration makes it possible to identify complex non-linear interactions within observational datasets while simultaneously preserving covariate balance between comparison groups.

In contrast to traditional statistical matching procedures, the developed framework supports deeper modeling of heterogeneous clinical patterns and hidden risk relationships. An additional component of the methodology involves the implementation of Multi-Objective Pareto Optimization for benefit-risk evaluation [10, 11].

Instead of relying on arbitrary weighting coefficients, the framework identifies mathematically optimal trade-offs between efficacy and safety indicators. This makes it possible to compare supplement profiles using an objective and scale-invariant analytical approach suitable for regulatory applications. The overall methodology was constructed according to principles of Target Trial Emulation, allowing observational datasets to approximate conditions associated with randomized comparative studies [5, 12].

Materials and methods

To validate the framework, a dual-track data methodology was employed, encompassing a clinical audit of non-interventional pilot studies and an empirical validation on high-dimensional synthetic RWD.

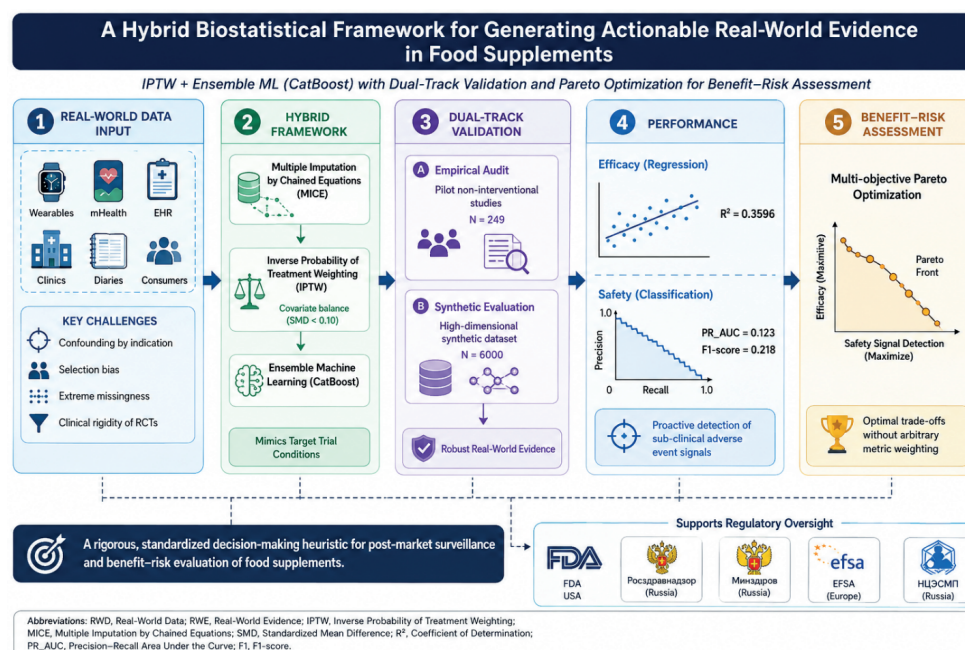


Fig. 1. Workflow of the proposed biostatistical methodology for the transformation of Real-World Data into Real-World Evidence

Data Acquisition and Characterization

Validation of the proposed methodology was carried out using two complementary data sources that reflected both practical clinical observations and large-scale simulated real-world conditions. The first stage included analysis of two non-interventional pilot studies. The Study 1 dataset represented a prospective observational cohort study involving 50 peri- and postmenopausal women receiving Food supplement 1 over a six-month period. The Study 2 dataset included 199 patients with chronic low back pain who received therapy based on the Food supplement 2 during

a 90-day observation interval. These datasets were selected because they reflected typical limitations encountered in routine observational practice, including incomplete records, heterogeneous patient characteristics, and variable treatment adherence.

The second stage involved validation using a high-dimensional dataset containing 6,000 observations designed to reproduce the structural variability of contemporary healthcare information systems. This was artificially created to reproduce the statistical properties and structural complexity of real medical data as shown by the generation algorithm below.

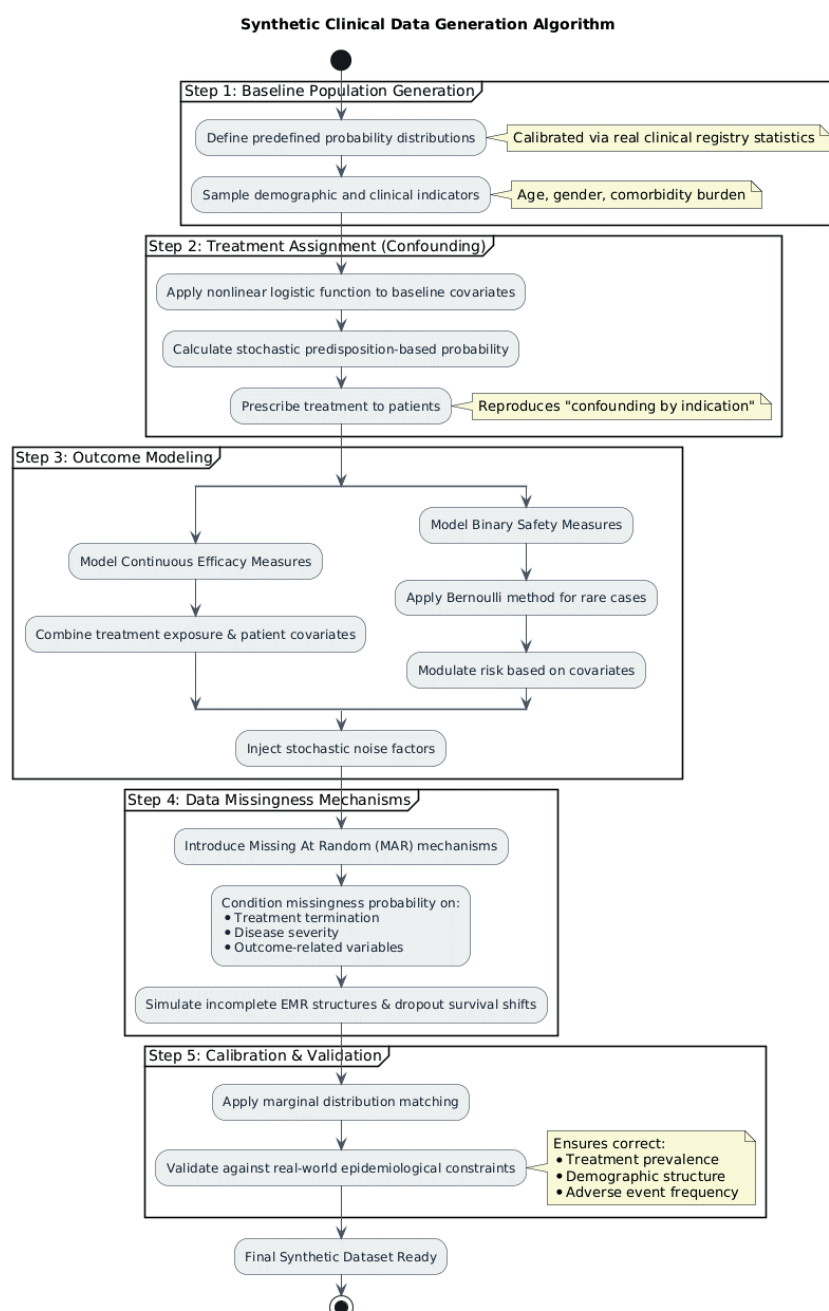


Fig. 2. Conceptual workflow of the synthetic clinical data generation algorithm, incorporating baseline confounding, heterogeneous outcomes, missingness mechanisms, and empirical calibration [16]

The generated dataset combined structured variables, including dosage, treatment duration, demographic characteristics, and clinical outcomes, with semi-structured and unstructured components such as physician notes, adherence descriptions, and socio-behavioral indicators. This made it possible to evaluate the framework under conditions closer to real-world clinical data environments.

Data Preprocessing and Handling Missing Data.

Before analytical modeling, preprocessing procedures were applied to improve consistency and reliability of the collected information. Duplicate records were removed, inconsistent terminology was standardized, and clinical coding systems were harmonized through conversion of ICD-9 and ICD-10 identifiers into phe-code-based groupings [16]. A major challenge involved the large proportion of incomplete observations frequently encountered in Real-World Data. Missing values were processed using Multiple Imputation by Chained Equations (MICE) [17]. Because the datasets originated from observational rather than randomized environments, the missingness mechanism was treated as Missing at Random (MAR). Ten imputation iterations were performed to stabilize parameter estimation and reduce variability of standard errors. The imputation model incorporated demographic variables, treatment-related characteristics, behavioral indicators, and additional derived features associated with text-based clinical descriptions. After completion of the imputation process, predictive outputs from all generated datasets were aggregated according to Rubin's Rules, allowing uncertainty introduced during imputation to be reflected within final estimates and confidence intervals.

The Analytical Framework: Causal Inference. To reduce the influence of selection bias and confounding by indication, propensity scores were estimated using multivariable logistic regression models [9]. Within the framework, the propensity score represented the probability of treatment assignment conditional on observed baseline characteristics.

$$e(X) = P(T = 1|X)$$

After estimation of propensity scores, Inverse Probability of Treatment Weighting (IPTW) was applied to construct a weighted pseudo-population in which treatment assignment became statistically independent from measured covariates [9]. Additional stabilization procedures were introduced to reduce the influence of extreme weights. Asymmetric trimming at the 1st and 99th percentiles was performed to prevent instability during machine learning model training [18]. Balance diagnostics were evaluated using Standardized Mean

Differences (SMDs). Covariate balance was considered acceptable when SMD values remained below 0.10 across baseline variables, indicating successful approximation of target trial conditions within the observational dataset [13].

Machine Learning Component. After causal balancing procedures had been completed, predictive modeling was performed using ensemble machine learning methods trained on the IPTW-weighted dataset. The analytical framework incorporated Gradient Boosting algorithms based on the CatBoost architecture together with Random Forest models [19].

The selected algorithms were capable of identifying non-linear dependencies, hidden feature interactions, and heterogeneous treatment patterns commonly observed in Real-World Data. Unlike traditional parametric models, ensemble tree-based methods do not require extensive manual feature engineering and remain more adaptable when processing complex observational datasets containing mixed variable types and irregular distributions [8].

A separate stage of the analysis focused on adverse-event prediction. Because the proportion of positive safety outcomes remained relatively low, the classification task was characterized by strong class imbalance. Under such conditions, use of the default probability threshold frequently leads to unstable predictive quality and insufficient sensitivity to rare events. To improve analytical performance, classification thresholds were additionally optimized instead of relying on the standard cutoff value of 0.50.

Model assessment emphasized optimization of the F1-score together with analysis of Precision-Recall trade-offs. This approach made it possible to improve detection of weak safety signals and increase stability of adverse-event prediction within highly imbalanced observational datasets.

Multi-Objective KPI Evaluation and Pareto Optimization. Traditional composite indices in clinical evaluation frequently rely on linear scalarization, conventional composite evaluation indices often combine several statistical indicators into a single scalar metric using predefined weighting coefficients [20]. In clinical and biostatistical applications, this approach may introduce subjective bias because different performance measures possess different mathematical properties and scales. For example, regression-based efficacy indicators such as R^2 are frequently combined with classification metrics including PR_{AUC} despite their fundamentally different analytical interpretations.

To reduce dependence on arbitrary weighting procedures, the developed framework incorporated Multi-Objective Pareto Optimization (MOO) [11]. Within the model, supplement interventions were

evaluated across a two-dimensional objective space representing efficacy and safety performance simultaneously.

The efficacy objective was defined as maximization of the regression performance metric:

Objective 1 (Efficacy): Maximize $f_1(x) = R^2$

The safety objective was defined as maximization of the Precision-Recall Area Under Curve:

Objective 2 (Safety): Maximize $f_2(x) = PR_{AUC}$

An intervention was considered Pareto optimal when no alternative configuration could improve efficacy without causing deterioration of the safety profile. This principle allowed identification of mathematically balanced solutions rather than relying on fixed subjective priorities between competing clinical indicators [21].

To support comparative ranking of interventions, the framework additionally calculated the normalized Euclidean distance from each solution to a theoretical

“utopian point” represented by the coordinates (1,1). The inverse transformed distance was then converted into a bounded scalar Key Performance Indicator ranging from 0 to 1 [22]. This made it possible to obtain a scale-invariant metric suitable for comparative regulatory evaluation and benefit-risk assessment.

Results

Phase 1: Critique of the Pilot Study Datasets.

Analysis of the Study 1 and Study 2 revealed significant structural limitations in current RWE generation methodologies. The pilot studies demonstrated profound “clinical rigidity.” By excluding high-risk populations (e. g., patients with severe symptoms or those on antidepressants), they failed to represent the complex consumers who actually utilize these supplements. Safety reporting was entirely reactive, and the use of Complete Case Analysis (deleting dropouts) introduced severe survivorship bias, masking true intolerance rates.

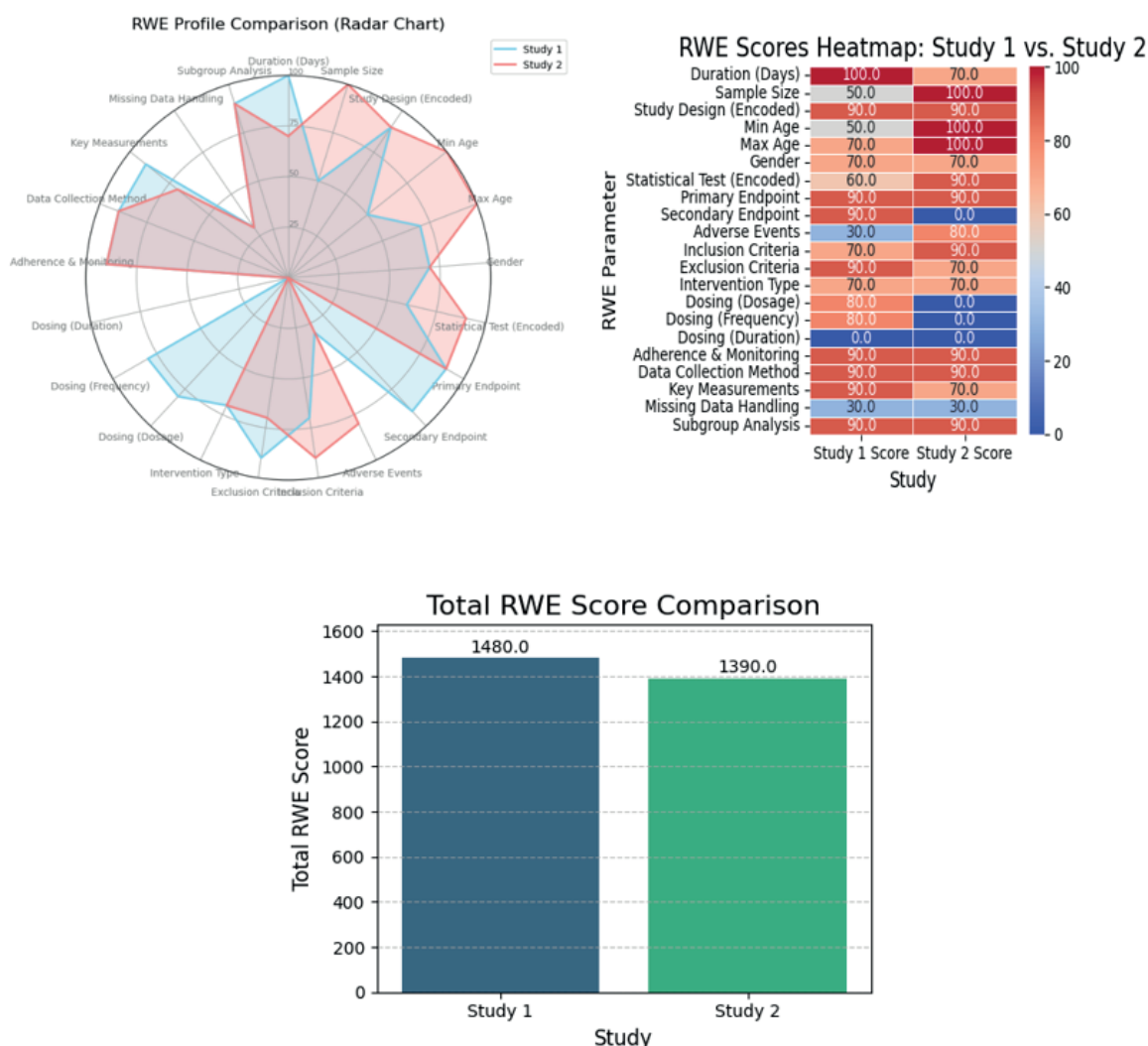


Fig. 3. Comparison of methodological and generalizability limitations in non-interventional supplement studies

Table 2. Methodological limitations were identified during the analysis of non-interventional pilot studies

Parameter	Study 1 (Food supplement 1)	Study 2 (Food supplement 2)	Identified Methodological Gap
Exclusion Criteria	Excluded severe symptoms (Greene > 20) and antidepressants.	Excluded complex pain, participants in other trials.	Clinical Rigidity: Artificially “cleans” the population, destroying external validity.
Missing Data	Complete Case Analysis (Excluded dropouts).	Excluded non-adherent patients.	Survivorship Bias: Deleting dropouts masks product intolerance and inflates efficacy.
Statistical Test	T-tests, ANOVA.	Logistic regression, ANOVA.	Statistical Gap: Fails to account for non-linear, time-varying confounders.

Phase 2: Validation on High-Dimensional RWD Missingness Profile and Imputation Diagnostics.

Evaluation of the synthetic validation dataset (N=6000) revealed a substantial proportion of incomplete observations within several treatment-related variables. The highest level of missingness was identified in the Dose_mg and Duration_Days parameters, where incomplete records accounted for 83.33 % of observations. Such data sparsity reflected conditions commonly encountered in large-scale Real-World Data environments.

To address this issue, Multiple Imputation by Chained Equations (MICE) was applied. Convergence diagnostics based on trace plot analysis demonstrated stable iterative behavior throughout the imputation process. Additional comparison between observed and reconstructed value distributions indicated that the Missing at Random (MAR) assumption remained acceptable for the analyzed dataset. As a result, incomplete patient records associated with treatment discontinuation and irregular follow-up were retained within the analytical pipeline instead of being excluded from subsequent modeling stages.

Causal Balancing Diagnostics. Before application of weighting procedures, several baseline covariates demonstrated Standardized Mean Difference (SMD) values above 0.25, indicating the presence of clinically

significant imbalance between comparison groups and substantial confounding by indication.

After implementation of stabilized Inverse Probability of Treatment Weighting with asymmetric trimming, covariate imbalance was markedly reduced. Across all measured baseline variables, SMD values decreased below the predefined threshold of 0.10. Overlap analysis additionally demonstrated sufficient common support between weighted treatment groups. This made it possible to construct a balanced pseudo-population suitable for subsequent predictive and comparative analysis.

Efficacy Prediction Results. Predictive models were trained using the weighted pseudo-population to estimate changes in Efficacy_Score_Improvement. Among the evaluated approaches, the Linear Regression model supplemented with interaction features associated with dosage and treatment duration demonstrated the most stable predictive performance.

The model achieved an R² value of 0.3596 with a 95 % confidence interval ranging from 0.33 to 0.38. The Mean Absolute Error reached 13.96, indicating moderate predictive deviation under high-dimensional observational conditions. Among evaluated supplement categories, the probiotic formulation based on Bifidobacterium demonstrated the highest average efficacy improvement score, reaching 56.63.

Table 3. Performance metrics of efficacy prediction models within the causally balanced pseudo-population

Model Architecture	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R-Squared (R ²)
Linear Regression (w/ interaction terms)	13.96	17.04	0.3596 (95 % CI: 0.33–0.38)
Random Forest Regressor	14.10	17.25	0.3435

Safety Results: Signal Detection and Threshold Tuning. Predicting rare Adverse Events in RWD presents a severe class-imbalance challenge (positive class rate < 8 %). Consequently, ROC_AUC is misleading. The framework utilized Precision-Recall

Area Under Curve (PR_{AUC}) and the F1-score as definitive metrics.

The CatBoost model with optimized classification thresholds demonstrated the highest predictive performance among the evaluated safety detection

approaches. The model achieved a PR_{AUC} value of 0.1230 together with an optimized F1-score of 0.2180, exceeding the performance of baseline logistic regression models under conditions of severe class imbalance.

The developed analytical approach also enabled identification of latent adverse-event patterns that may remain undetected within conventional passive reporting systems. This made it possible to improve

sensitivity of safety monitoring and strengthen detection of clinically relevant risk signals in observational supplement data.

The analytical model identified elevated adverse-event frequencies for several supplement categories. The highest rate was observed for Magnesium 400 mg at 12.47 %, followed by Magnesium Glycinate at 9.47 % and Vitamin D3 at 8.51 %.

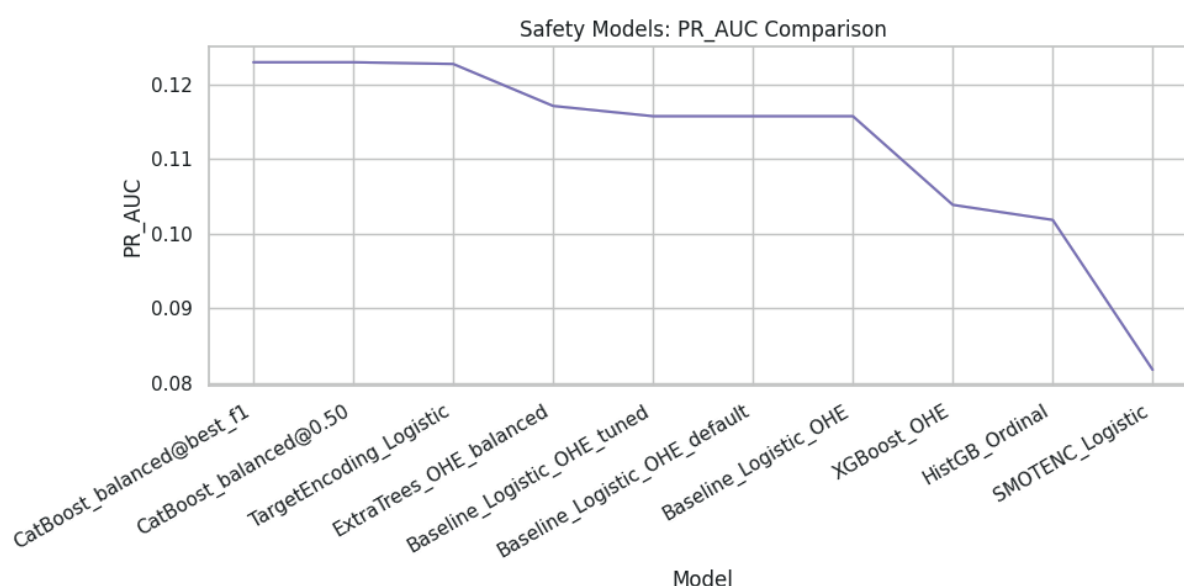


Fig. 4. Comparative evaluation of PR_{AUC} values for machine learning and logistic safety prediction models

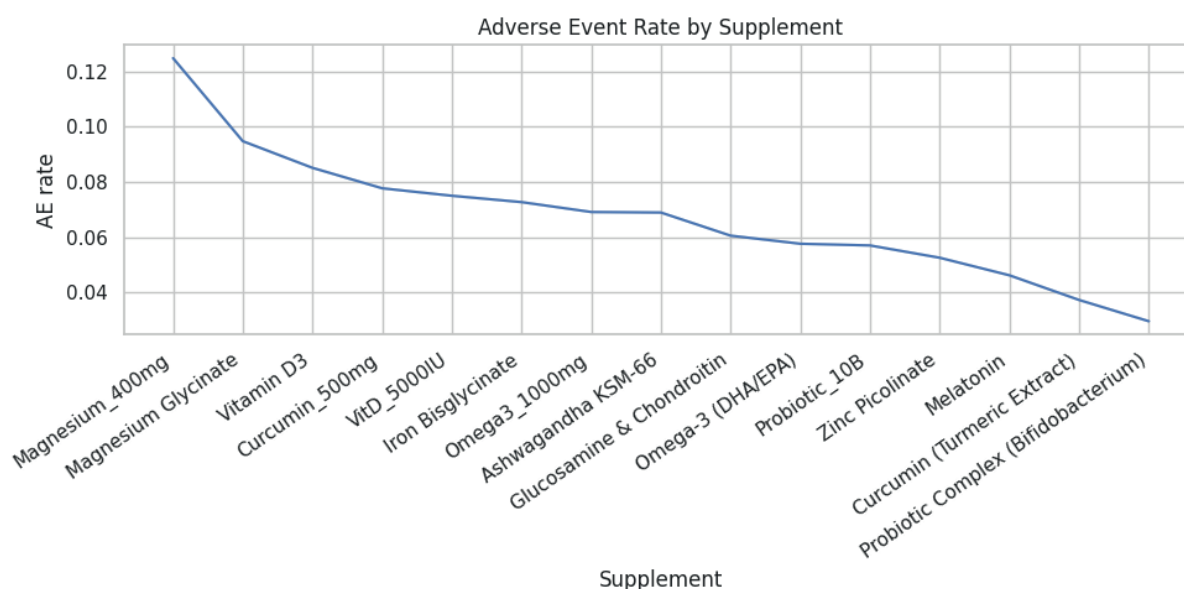


Fig. 5. Detection of high-risk supplement interventions using algorithm-based safety signal analysis

Table 4. Performance comparison of safety prediction models for detection of rare adverse events

Model Architecture	Decision Threshold	ROC _{AUC}	PR _{AUC}	Recall	F1-Score
CatBoost (Balanced)	Tuned (Best F1)	0.6120	0.1230	0.2500	0.2180
CatBoost (Balanced)	Default (0.50)	0.6120	0.1230	0.4347	0.1918
Target Encoding + Logistic	Default (0.50)	0.5915	0.1227	0.5760	0.1693
Baseline Logistic (OHE)	Tuned (Best F1)	0.5858	0.1157	0.3804	0.1928
Baseline Logistic (OHE)	Default (0.50)	0.5858	0.1157	0.5869	0.1674

Multi-objective performance evaluation and Pareto optimization of efficacy — safety trade-offs.

To reduce the limitations associated with fixed scalar weighting and improve interpretation of benefit-risk relationships, multi-objective Pareto optimization was applied within the analytical framework. The optimization procedure mapped efficacy indicators represented by R^2 values against safety performance measured through PR_{AUC} metrics. This approach made it possible to evaluate supplement interventions simultaneously across both clinical effectiveness and safety dimensions without introducing subjective weighting coefficients.

As illustrated in Fig. 5, the optimization procedure eliminated reliance on arbitrary weighting schemes by identifying a well-defined Pareto frontier of optimal supplement interventions across competing efficacy and safety objectives:

Maximum efficacy was observed for Omega3_1000mg, which demonstrated the highest overall clinical improvement score within the evaluated intervention set.

Maximum safety signal sensitivity was associated with Magnesium Glycinate, indicating the strongest capacity for proactive adverse-event detection.

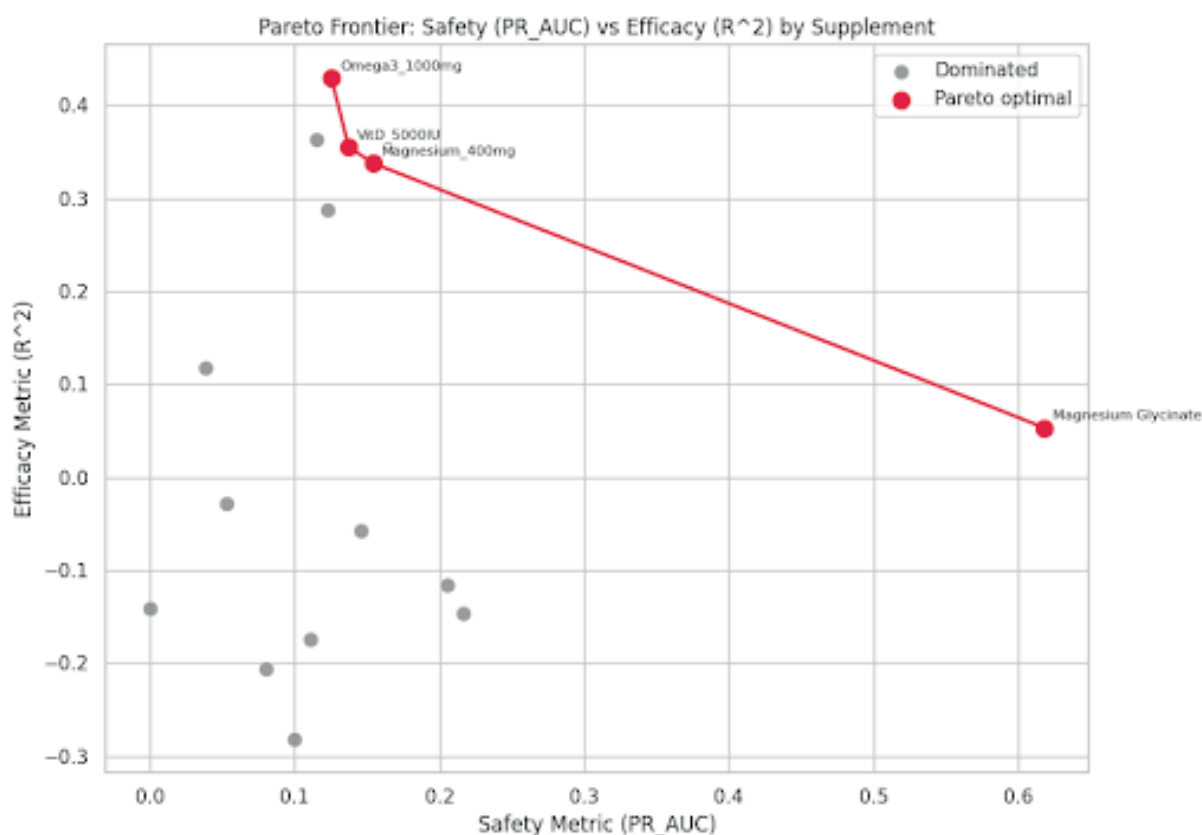


Fig. 6. Multi-objective Pareto frontier illustrating trade-offs between efficacy and safety in supplement intervention evaluation

Balanced trade-off solutions were represented by VitD_5000IU and Magnesium_400mg, which occupied the central region (“elbow point”) of the Pareto frontier, reflecting efficient compromises between therapeutic efficacy and safety monitoring performance.

Discussion and Future Directions

The results obtained were compiled and sent for Human Interpretation by specialists in biological data analysis.

Interpretation of Results and “Clinical Rigidity”. The analysis of the pilot studies highlights a pervasive state of “clinical rigidity” in nutritional science. Although ostensibly designed as non-interventional observational studies, they persist in forcing heterogeneous real-world patients into homogeneous “clean boxes” [6]. By excluding populations with severe symptoms, specific pain etiologies, or those utilizing concomitant medications, these trials create a massive Generalizability Gap. The outputs of such trials represent an idealized physiology, entirely detached from the high-risk, real-world consumer base. Our proposed framework rectifies this by ingesting the totality of the RWD and utilizing IPTW to adjust for such covariates mathematically, rather than excluding them physically [12].

Addressing the Data Continuity Gap. Traditional analytical approaches address messy data and irregular dosing through Complete Case Analysis (data deletion). As seen in the pilot studies, excluding non-adherent patients introduces Survivorship Bias, artificially inflating supplement success rates because only the most tolerant, compliant patients remain [6]. The biostatistical methodology developed herein solves this by instituting Multiple Imputation by Chained Equations (MICE) [17]. Despite the use of MICE, a high proportion of missing values (up to 83.33 % in individual variables) remains a significant source of statistical uncertainty. By computationally estimating missing data from dropouts or Missing at Random (MAR), the framework preserves the statistical reality of product intolerance. The successful application of MICE to data with 83.33 % missingness made it possible to preserve observations that would otherwise have been excluded using listwise deletion. However, the approach does not eliminate the risk of bias completely, especially if assumptions about the structure of omissions are violated [23].

The Statistical Gap: Linear vs. Advanced Inference. The validation dataset proved the necessity of advanced machine learning methodologies over standard parametrics. While a well-engineered Linear Regression captured the continuous baseline efficacy

($R^2=0.3596$), predicting rare safety events required the non-linear dimensional capability of CatBoost [9]. Traditional statistical models like logistic regression fail to account for non-linear, high-dimensional interactions without massive manual feature engineering. CatBoost’s native handling of categorical text and optimized F1 thresholding proved that ensemble tree models are mandatory for proactive signal detection in highly imbalanced real-world datasets [24, 25].

Bridging the Safety Gap: Proactive Monitoring.

Currently, the approach to supplement safety is hindered by a reactive reporting system, relying on physicians to notice and manually document adverse events. This passive method misses cumulative, sub-clinical indicators. Our proposed framework disrupts this paradigm by introducing Proactive Safety Signal Detection [26]. By implementing continuous analysis on raw RWD using CatBoost [25, 27], our framework uncovered hidden safety signals such as the 12.47 % adverse event rate associated with Magnesium 400 mg that traditional “total quantity” manual reporting logs entirely mask.

Methodological Limitations. While highly effective, the proposed framework possesses inherent limitations. First, the primary empirical validation was conducted on a high-dimensional synthetic dataset. Although intricately designed to mimic the multi-modal complexity and structural noise of modern health data [27], authentic real-world registries are characterized by unpredictable anomalies that synthetic generation processes may struggle to fully replicate [4]. Secondly, causal inference methodologies, specifically the IPTW utilized in this framework, operate on the strict, untestable assumption of “no unmeasured confounding” [6]. If critical variables that influence both supplement choice and outcomes are absent from the RWD, residual bias remains. Furthermore, the validity of causal effect estimates relies heavily on the accurate specification of the propensity score model [26]. Consequently, future deployments of this framework must incorporate robust sensitivity analyses (e. g., E-values) to mathematically quantify the potential impact of unmeasured variables on the observed treatment effects [12].

Comparative Validation with Existing Literature. To substantiate the robustness of the proposed framework, it is imperative to benchmark the results obtained against contemporary biostatistical literature. The success of the IPTW in reducing SMDs to <0.10 directly corroborates the target trial emulation principles established by *Cui et al.* [12], who demonstrated that mimicking RCT exchangeability through rigorous propensity score weighting is mathematically

superior to traditional covariate adjustment. Furthermore, the superior performance of the CatBoost architecture in detecting rare adverse events aligns with the findings of *Reps et al.* [26] and *Marzano et al.* [27], who proved that ensemble models significantly outperform standard parametric designs in risk stratification for imbalanced clinical datasets. Finally, the application of the Pareto Frontier objectively maps trade-offs, ensuring that regulatory evaluations are scale-invariant and immune to the *priori* weighting bias heavily critiqued in modern multi-criteria decision analysis literature [28].

Regulatory Implications. The developed framework provides a highly structured, foundational prototype for global oversight bodies seeking to evaluate nutraceuticals using real-world data. It successfully maps the necessary biostatistical pathway to transition from passive, reactive pharmacovigilance to proactive, automated post-market monitoring [5]. By providing a scale-invariant, objective heuristic (the Pareto Frontier and Utopian Point distance), regulatory agencies like the FDA, EFSA, and Roszdravnadzor RF can rank and compare disparate products objectively. This empowers domestic clinical research organizations to quantitatively substantiate product portfolios with mathematically unbiased data, fundamentally shifting post-market surveillance from subjective estimation to rigorous, algorithm-backed science [24].

Future Perspectives. Future iterations of this research should include several technological extensions aimed at improving analytical precision and expanding the applicability of the proposed framework across diverse data environments.

Utilizing direct APIs, the framework can seamlessly ingest Patient-Generated Health Data (PGHD) from wearable devices. This enables continuous acquisition of physiological telemetry in real time and reduces reliance on retrospective self-reporting. As a result, recall bias associated with traditional patient diaries is minimized, while temporal resolution of health-related measurements is significantly improved [29].

The integration of advanced Large Language Models (LLMs) allows systematic processing of unstructured clinical text, including free-form physician notes and narrative medical records. Such models can be used to extract latent safety-related information and convert unstructured descriptions into structured analytical variables suitable for downstream statistical

and machine learning analysis. This makes it possible to improve completeness of safety signal detection and reduce information loss during data preprocessing stages [30].

Further development of the framework should involve deployment across large-scale federated electronic health record (EHR) infrastructures spanning multiple countries. This approach enables evaluation under heterogeneous healthcare systems, differing dietary patterns, and variable genetic backgrounds. Such cross-population validation improves robustness of the methodology and enhances generalizability of results obtained from real-world data analyses [5].

Conclusion

This study developed and validated a hybrid biostatistical framework for evaluating food supplement efficacy and safety using Real-World Data (RWD). The analysis of two non-interventional pilot studies revealed key limitations of conventional statistical approaches in observational settings, particularly related to clinical rigidity and reliance on reactive safety monitoring. Using a high-dimensional validation cohort (N=6000), the framework applied Multiple Imputation by Chained Equations (MICE) to address extreme missingness (83.33 % in key dosage and duration variables), allowing recovery of incomplete records and reducing survivorship bias. Subsequent application of Inverse Probability of Treatment Weighting (IPTW) reduced confounding by indication and improved covariate balance, with all standardized mean differences brought below 0.10 in line with target trial emulation criteria. For outcome modeling, the linear regression approach with interaction terms produced a stable R^2 of 0.3596. In parallel, ensemble machine learning methods (CatBoost) were used for safety signal detection, achieving a PR_{AUC} of 0.123 and an optimized F1-score of 0.218. This enabled identification of previously underreported adverse event patterns, including a 12.47 % rate associated with Magnesium 400 mg, which is not typically captured through conventional reporting systems. To support integrated benefit — risk evaluation, a multi-objective Pareto optimization approach was introduced, removing the need for subjective weighting of performance metrics and enabling identification of optimal trade-offs between efficacy and safety. Overall, the results indicate that the proposed framework improves consistency and robustness in real-world supplement evaluation and supports more structured post-market assessment of efficacy and safety outcomes.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов в связи с данной публикацией.

Участие авторов

Идриса К., Драгунс В., Булкрам М. С. — концептуализация и исследование; Идриса К., Драгунс В. — методология, ресурсы и их подбор; Идриса К., Холопов Р. А. — формальный анализ, визуализация и проверка; Идриса К., Булкрам М. С. — написание первоначального варианта, рецензирование и редактирование; Драгунс В., Булкрам М. С. — научное руководство, интерпретация данных, консультирование и управление.

Благодарность

Авторы хотели бы выразить благодарность компании X7 Research за её вклад в сбор и обработку данных, а также за постоянное и последовательное руководство на протяжении всего процесса.

Финансирование

Исследование выполнено без целевого внешнего финансирования.

СВЕДЕНИЯ ОБ АВТОРАХ

Идриса Кирёва — магистрант, ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Российская Федерация
e-mail: 408079@edu.itmo.ru
ORCID ID: 0009-0007-3178-0424

Холопов Руслан Александрович — студент, Санкт-Петербургский колледж телекоммуникаций им. Э. Т. Кренкеля — структурное подразделение ФГБОУ ВО «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича», Санкт-Петербург, Российская Федерация
e-mail: kholopovruslan@gmail.com
ORCID ID: 0009-0009-5112-3960

Драгунс Виталий — зам. генерального директора ООО «Экс Севен Ресеч», Санкт-Петербург, Российская Федерация
e-mail: vdraguns@x7cpr.com
ORCID ID: 0009-0000-4454-2782

Булкрам Мохамед Саид — доцент, ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Российская Федерация
Автор, ответственный за переписку
e-mail: mboulkrane@itmo.ru
ORCID ID: 0000-0002-1204-5041

ADDITIONAL INFORMATION

Conflict of interests

The authors declare that there is no conflict of interest in connection with this publication.

Authors' participation

Idrisa K., Draguns V., Bulkran M. S. — conceptualization and investigation; Idrisa K., Draguns V. — methodology, resources and their selection; Idrisa K., Kholopov R. A. — formal analysis, visualization and validation; Idrisa K., Bulkran M. S. — writing the original version, review and editing; Draguns V., Bulkran M. S. — scientific supervision, data interpretation, consultancy and management.

Acknowledgement

The authors would like to acknowledge X7 research for their contribution in data acquisition and processing, plus their increased and consistent guidance throughout the process.

Financing

The study was carried out without targeted external funding.

ABOUT THE AUTHORS

Kiryowa Idrisa — MSc, Student, ITMO University, St. Petersburg, Russian Federation
e-mail: 408079@edu.itmo.ru
ORCID ID: 0009-0007-3178-0424

Ruslan A. Kholopov — student, St. Petersburg College of Telecommunications named after E. T. Krenkel is a structural division of the St. Petersburg State University of Telecommunications named after prof. M. A. Bonch-Bruevich, St. Petersburg, Russian Federation
e-mail: kholopovruslan@gmail.com
ORCID ID: 0009-0009-5112-3960

Vitaly Draguns — deputy general director, X7 Research LLC, St. Petersburg, Russian Federation
e-mail: vdraguns@x7cpr.com
ORCID ID: 0009-0000-4454-2782

Mohamed Said Boulkrane — Associate Professor, ITMO University, St. Petersburg, Russian Federation

Corresponding author

e-mail: mboulkrane@itmo.ru
ORCID ID: 0000-0002-1204-5041

References

1. Abdel-Tawab M. Do We Need Plant Food Supplements? A Critical Examination of Quality, Safety, Efficacy, and Necessity for a New Regulatory Framework. *Planta Med.* 2018 Apr;84(6-07):372-393. doi: 10.1055/s-0043-123764.
2. Chodankar D. Introduction to real-world evidence studies. *Perspect Clin Res.* 2021 Jul-Sep;12(3):171-174. doi: 10.4103/picr.picr_62_21.
3. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf.* 2017 Sep;26(9):1033-1039. doi: 10.1002/pds.4297.
4. Justo N, Espinoza MA, Ratto B, et al. Real-World Evidence in Healthcare Decision Making: Global Trends and Case Studies From Latin America. *Value Health.* 2019 Jun;22(6):739-749. doi: 10.1016/j.jval.2019.01.014.
5. FDA UF& DA. Framework for FDA's Real-World Evidence Program. U.S. Department of Health and Human Services; 2018 Feb.
6. Al-Sahab B, Leviton A, Loddenkemper T, et al. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *J Healthc Inform Res.* 2023 Nov 14;8(1):121-139. doi: 10.1007/s41666-023-00153-2.
7. Blonde L, Khunti K, Harris SB, et al. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. *Advances in Therapy.* 2018 Nov;35(11):1763-1774. DOI: 10.1007/s12325-018-0805-y.
8. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *ArXiv Cornell Univ.* 2017 Jun;31:6639-49. DOI: 10.48550/arxiv.1706.09516
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983 Jan;70(1):41-55. DOI: 10.1093/biomet/70.1.41.
10. Cristoforetti R, Süß P, Becher T, Wahl N. Trading robustness: a scenario-free approach to robust Multi-Criteria Optimization for Treatment Planning. *ArXiv Cornell Univ.* 2025 Oct. DOI: 10.48550/arxiv.2510.12519
11. Tsionas EG. Multi-objective optimization using statistical models. *Eur J Oper Res.* 2019 Jan; 276(1):364-78. DOI: 10.1016/j.ejor.2018.12.042.
12. Baumeister SE, Leiva-Escobar I, Demmer RT, et al. Target Trial Emulation in Observational Research-Strengths, Limitations, and Methodological Considerations. *J Periodontol Res.* 2026 Mar 6. doi: 10.1111/jre.70095.
13. Kaufmann NC, Pereira PL. Real-World Evidence: Better than Nothing or the Future? A Perspective on Clinical Evidence Generation in Interventional Oncology. *Cardiovasc Intervent Radiol.* 2026 May; 49(5):854-861. doi: 10.1007/s00270-025-04179-4.
14. Khaowroongrueng V, Kim TE, Park SI, Shin KH. Application of Real-World Evidence to Support FDA Regulatory Decision Making. *AAPS J.* 2025 May 28;27(4):98. doi: 10.1208/s12248-025-01082-1.
15. Lin X, Tarp JM, Evans RJ. Data fusion for efficiency gain in ATE estimation: A practical review with simulations. *arXiv* (Cornell University) [Internet]. Cornell University; 2024. Available from: <http://arxiv.org/abs/2407.01186> DOI: 10.48550/arxiv.2407.01186
16. Yan, C., Yan, Y., Wan, Z. et al. A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun* 2022;13:7609. Doi: 10.1038/s41467-022-35295-1
17. Azur M, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res.* 2011 Feb 24;20(1):40-49. doi: 10.1002/mpr.329.
18. Zang C, Zhang H, Xu J, et al. High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. *Nat Commun.* 2023 Dec 11;14(1):8180. doi: 10.1038/s41467-023-43929-1.
19. Bechny M, Kishi A, Fiorillo L, et al. Novel Digital Markers of Sleep Dynamics: A CausalInference Approach Revealing Age and Gender Phenotypes in Obstructive Sleep Apnea. *Research Square;* 2024. DOI: 10.21203/rs.3.rs-5317875/v1.
20. Yin H, Li XR, Gao Y. Relative Euclidean Distance With Application to TOPSIS and Estimation Performance Ranking. *IEEE Trans Syst Man Cybern Syst.* 2020 Sep;52(2):1052-64. DOI: 10.1109/tsmc.2020.3017814.
21. Mazza A, Chicco G, Russo A, Virjoghe EO. Multi-Objective Distribution Network Reconfiguration Based on Pareto Front Ranking. *Intell Ind Syst.* 2016 Dec;2(4):287-302. DOI: 10.1007/s40903-016-0065-6.
22. Lazzarini B, Pistolesi F. Multiobjective Personnel Assignment Exploiting Workers' Sensitivity to Risk. *IEEE Trans Syst Man Cybern Syst.* 2017 Feb;48(8):1267-82. DOI: 10.1109/tsmc.2017.2665349
23. Dang LE, Gruber S, Lee H, et al. A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science.* 2023 ;7(1):e212. DOI: 10.1017/cts.2023.635.
24. Brantner CL, Chang TH, Nguyen TQ, et al. Methods for Integrating Trials and Non-experimental Data

- to Examine Treatment Effect Heterogeneity. *Stat Sci.* 2023 Nov;38(4):640-654. doi: 10.1214/23-sts890.
25. Huang Y, Li J, Li M, Aparasu RR. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Med Res Methodol.* 2023 Nov 13;23(1):268. doi: 10.1186/s12874-023-02078-1.
 26. Reps JM, Garibaldi JM, Aickelin U, et al. Signaling paediatric side effects using an ensemble of simple study designs. *Drug Saf.* 2014 Mar;37(3):163-70. doi: 10.1007/s40264-014-0137-z.
 27. Marzano L, Darwich AS, Tendler S, et al. A novel analytical framework for risk stratification of real-world data using machine learning: A small cell lung cancer study. *Clin Transl Sci.* 2022 Oct;15(10):2437-2447. doi: 10.1111/cts.13371.
 28. Fatima G, Khan S, Shukla V, et al. Nutraceutical formulations and natural compounds for the management of chronic diseases. *Front Nutr.* 2025 Oct 15;12:1682590. doi: 10.3389/fnut.2025.1682590.
 29. Kadro ZO, Chilcoat A, Hill J, et al. Healthcare Professionals' Perspectives on Improving Dietary Supplement Documentation in the Electronic Medical Record: Current Challenges and Opportunities to Enhance Quality of Care and Patient Safety. *Glob Adv Integr Med Health.* 2023 Dec 19;12:27536130231215029. doi: 10.1177/27536130231215029.
 30. Rita L, Southern J, Laponogov I, et al. Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes. *ArXiv Cornell Univ.* 2024 Sep. DOI: 10.48550/arxiv.2409.08792.