
БАСТЫ МАҚАЛАЛАР

UDC: 001.8:519.2:614:61**METHODS FOR ASSESSING THE CUMULATIVE EFFECT OF
MULTIPLE FACTORS ON THE MANIFESTATION OF THE EFFECT IN
BIOMEDICAL RESEARCH****Sabirov Z.B¹, Otarov E.Z¹, Alekseev A.V¹, Grebeneva O.V¹,
Shadetova A.Z¹.**

¹NC JSC «National Centre Occupational Health and Diseases» (100017, Republic of Kazakhstan, Karaganda, Mustafina str. 15, e-mail: info@naoncgt.kz)

1.Sabirov Z.B., PhD, Head of the Scientific Laboratory of Occupational Pathology of NC JSC National Centre Occupational Health and Diseases, e-mail: audacious_zap@mail.ru

2.Otarov E.Z., e-mail: otarov_kgmu@mail.ru

3.Alekseev A.V., e-mail: alekseev@qmu.kz

4.Grebeneva O.V., e-mail: ol_grebeneva@bk.ru

5.Shadetova A.Z., e-mail: alma7722@mail.ru

In biomedical research, it is often necessary to quantitatively assess the combined effect of multiple factors-both quantitative and qualitative-on a specific biomedical outcome. This article discusses approaches to data normalization, determination of factor weights, information aggregation, and model development for evaluating the cumulative influence of factors. Both expert-based and data-driven methods are examined, along with their application to health prediction, risk assessment, and decision-making in biomedicine.

Key words: medical and biological research, integral indicator, multivariate analysis, statistical models.

1. Introduction. Modern biomedical research increasingly requires a comprehensive assessment of the combined effects of multiple heterogeneous factors on human health. Such challenges are particularly relevant in studies of disease risk prediction, the impact of occupational and environmental exposures, and the identification of diagnostic and prognostic biomarkers. A distinctive feature of these investigations is the heterogeneity of input data: the influencing factors may be quantitative (e.g., blood glu-

cose concentration) or qualitative/dichotomous (e.g., presence of harmful habits, working conditions, sex).

Effective integration of such diverse variables, as well as consideration of their potential interactions, demands the use of appropriate statistical and mathematical methods. These approaches enable researchers to move beyond the analysis of individual factors and toward the development of multidimensional models that reflect the cumulative and potentially synergistic effects of various influences on biomedical outcomes.

2. Classification and Representation of Factors. The factors influencing the studied outcome can be classified as follows:

Quantitative continuous variables: measured on an interval or absolute scale (e.g., age, body weight).

Quantitative discrete variables: for instance, the number of disease exacerbations per year.

Qualitative (categorical) variables: such as the presence or absence of a particular trait (dichotomous), or blood type categories.

To be included in a statistical or mathematical model, all variables must be transformed into a numerical format. Common approaches for this transformation include binarization, one-hot encoding, and normalization procedures [1].

3. Methods for Factor Normalization. To ensure comparability across different factors, it is necessary to bring them to a common scale. The following approaches are commonly applied:

Min–max normalization:

$$x_i^{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Z-transformation (standardization):

$$x_i^{norm} = \frac{x_i - \mu}{\sigma}$$

Quantile normalization: commonly used for high-dimensional biological data, such as transcriptomics [1].

Binarization: applied to dichotomous factors (0/1).

The choice of normalization method depends on the nature of the data and the goals of the analysis.

4. Determining the Importance Weights of Factors. To combine multiple factors into an aggregate assessment, it is necessary to assign importance weights w_i , reflecting the contribution of each feature to the studied effect. The following approaches are commonly used:

4.1. Expert-Based Methods

Direct ranking and normalization: experts assign weights based on their professional judgment.

Pairwise comparison method (Analytic Hierarchy Process, AHP): used to assess the relative importance of factors [2].

Delphi method: an iterative survey process designed to achieve consensus among experts [3].

4.2. Data-Driven Methods

Linear and logistic regression: model coefficients can be interpreted as weights [4].

Decision trees and ensemble methods (Random Forest, XGBoost): importance is estimated based on each factor's contribution to reducing model error [5].

Feature selection methods (LASSO, PCA): applied to reduce dimensionality and identify the most relevant factors [6].

The choice of method depends on the availability of data and the objectives of the study.

5. Factor Aggregation and Construction of a Composite Index.

5.1. Linear Model

The most common method of integration is a weighted linear combination:

$$E = \frac{\sum_{i=1}^n w_i x_i^{norm}}{\sum w_i}$$

Where $E \in [0,1]$ — is the resulting quantitative estimate of the effect.

5.2. Nonlinear Models

Logistic function:

$$E = \frac{1}{1 + e^{-\sum w_i x_i}}$$

Logistic function: Frequently used when the outcome is binary and a probabilistic interpretation is required. Neural networks: Capable of capturing complex, nonlinear interactions between factors. Bayesian networks: Particularly useful when data have an inherent probabilistic structure. In clinical research, especially in situations with limited or incomplete data, Bayesian methods offer powerful tools for risk estimation and disease prediction.

For example, a study based on a diabetic patient database employed Bayesian regression to predict the likelihood of complications based on factors such as blood glucose levels, age, and genetic predisposition. In cases with missing or sparse data, such methods can incorporate uncertainty and generate more accurate, individualized predictions [7–8].

In small-sample studies, Bayesian logistic regression can be used to estimate the probability of breast cancer based on variables such as age, family history, and biopsy results. In the context of rare diseases or limited data availability, Bayesian approximation methods help to account for uncertainty in model parameters and enhance predictive accuracy.

A notable advantage of Bayesian models is their capacity to handle missing data by incorporating all possible scenarios and variations. For instance, in modeling chronic disease development in patients with incomplete medical test results, Bayesian methods allow for more accurate projections. They are also valuable in adjusting for data limitations in clinical trials, where small sample sizes and missing observations are common.

These models can capture probabilistic dependencies between variables (e.g., age, sex, cholesterol level), enabling more precise and tailored risk assessments. The 2019 study “A Bayesian Network for Predicting Lung Cancer Risk” demonstrated how a Bayesian network approach can be applied to estimate lung cancer risk based on multiple factors including age, smoking status, family history, and medical background [9].

Other research has explored the use of Bayesian hierarchical models to integrate data from multiple clinical centers, accounting for variations in data collection methods and quality across sites. Such models are especially useful for analyzing multi-center clinical studies with heterogeneous or incomplete datasets [10].

The choice of model depends on the complexity of the problem and the availability of data. The next step involves evaluating model performance.

6. Model Evaluation and Interpretation of Results. Evaluating the quality of a model that aggregates multiple factors into a single composite indicator is essential to ensure its reliability and applicability in biomedical research. Without proper validation, it is impossible to determine whether the resulting model adequately reflects the cumulative impact of the contributing factors on the outcome of interest.

6.1. Model Performance Metrics.

If the outcome variable is continuous (e.g., a health status score ranging from 0 to 1), commonly used evaluation metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). These metrics quantify the model's accuracy and assess how well it explains the observed data.

For binary outcomes (e.g., presence or absence of disease), performance is assessed using sensitivity (the ability to correctly identify positive cases), specificity (the ability to correctly identify negative cases), accuracy, and the Area Under the ROC Curve (AUC). AUC provides a measure of the trade-off between sensitivity and specificity, and is particularly important in screening and early diagnostic tasks [11].

6.2. Model Validation Methods.

To assess the robustness and generalizability of the model, various validation techniques are employed. The most widely used is cross-validation, in which the dataset is partitioned into several subsets; each subset is used in turn as a test set while the remaining subsets are used for training. This method helps prevent overfitting and improves the objectivity of the performance assessment.

Another approach is the bootstrap method, which involves repeated random sampling with replacement. This technique provides an estimate of the variability and stability of the model's performance metrics. A simpler alternative is the train-test split, in which the dataset is divided into a training and a test set; however, this method is more sensitive to random variation and less robust [12].

6.3. Interpretability and Trust in the Model.

Interpretability is of paramount importance in biomedical applications, as the model's outputs must be understandable to healthcare professionals making decisions. Modern techniques for model interpretation include SHAP (SHapley Additive ex Planations), which is based on cooperative game theory, and LIME (Local Interpretable

Model-agnostic Explanations), which constructs local surrogate models to explain predictions.

In addition, partial dependence plots (PDPs) are used to visualize the marginal effect of individual features on the predicted outcome while holding other variables constant. These tools enhance both the transparency and the usability of predictive models, which is especially critical when integrating such models into clinical decision-support systems [13].

Ensuring interpretability is particularly important in medical and biological contexts, where understanding the contribution of each factor is crucial. For example, these approaches can be applied in:

Environmental risk assessment based on pollutant concentrations, chronic disease prevalence, and lifestyle factors;

Treatment outcome prediction using clinical and laboratory data;

Occupational risk evaluation in the presence of harmful workplace exposures.

7. Conclusion. A comprehensive assessment of the influence of multiple factors on biomedical outcomes requires an integration of statistical and expert-based approaches. Standardized normalization procedures, flexible aggregation models, and data-driven learning methods allow the analytical framework to be tailored to specific research objectives. Among the most promising directions are approaches that combine machine learning techniques with model interpretability, ensuring both predictive power and transparency in biomedical applications.

Reference

1.Zhang X., Lee J., Goh W. W. B. An investigation of how normalisation and local modelling techniques confound machine learning performance in a mental health study // Heliyon. – 2022. – T.8. – №. 5. https://scholar.google.com/scholar?output=instlink&q=info:RV1mT0muf8kJ:scholar.google.com/&hl=ru&as_sdt=0,5&scillfp=9109182181854836428&oi=lle

2.Salomon V. A. P., Gomes L. F. A. M. Consistency improvement in the analytic hierarchy process // Mathematics. – 2024. – T.12. – №.6. – C. 828. <https://www.mdpi.com/2227-7390/12/6/828/pdf>

3.Best R. W. et al. An Overview of the Delphi Method's Origin, Modifications, and Use to Augment Instrument Development and Data Collection: A Research Note //

Journal of International Agricultural and Extension Education. – 2025. – T.32. – №.1. – C.4. <https://newprairiepress.org/cgi/viewcontent.cgi?article=1497&context=jiaee>

4.Dey D. et al. The proper application of logistic regression model in complex survey data: a systematic review // BMC Medical Research Methodology. – 2025. – T.25. – №.1. – C.15. <https://link.springer.com/content/pdf/10.1186/s12874-024-02454-5.pdf>

5.Johnson R. A. quantile-forest: A python package for quantile regression forests // Journal of Open Source Software. – 2024. – T.9. – №.93. – C.5976. <https://joss.theoj.org/papers/10.21105/joss.05976.pdf>

6.Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society Series B: Statistical Methodology. – 1996. – T.58. – №.1. – C.267-288. https://www.ccs.neu.edu/home/eelhami/courses/EE290A/LASSO_Tibshirani.pdf

7.Keeney R. L., Raiffa H. Decisions with multiple objectives: preferences and value trade-offs. – Cambridge university press, 1993. https://books.google.com/books?hl=ru&lr=&id=1oEa-BiARWUC&oi=fnd&pg=PR11&dq=7.%09Keeney+R.+L.,+Raiffa+H.+Decisions+with+multiple+objectives:+preferences+and+value+trade-offs.+—+Cambridge+university+press,+1993.&ots=cEDKV0sn-x&sig=6ZuyQg2d9m_xLb6dPAg45zN5F-Y

8.Bishop C. M., Nasrabadi N. M. Pattern recognition and machine learning. – New York : springer, 2006. – T. 4. – №. 4. – C. 738. http://crowley-coutaz.fr/jlc/Courses/20_20/PRML/ENSI3.PRML.S6.Encoders.pdf

9.Todaro S. et al. Correlation between extinction pattern and $\delta^{13}\text{C}$ fluctuations across the Triassic Jurassic boundary in shallow water settings: a proxy for the present day acidification processes // "Geosciences for the environment, natural hazard and cultural heritage" Congresso SGI-SIMP 2018-Abstract book. – [https:// www. socgeol. it/files/download/pubblicazioni/Abstract Book/Abstract Catania_ok. pdf](https://www.socgeol.it/files/download/pubblicazioni/Abstract%20Book/Abstract%20Catania_ok.pdf), 2018. – C. 20-20. <https://iris.unipa.it/handle/10447/298218>

10.Bedrick E. J. Data Reduction Prior to Inference: Are There Consequences of Comparing Groups Using at-Test Based on Principal Component Scores? // Biometrics. – 2020. – T.76. – №.2. – C.508-517. <https://academic.oup.com/biometrics/article-abstract/76/2/508/7452939>

11. James G. et al. An introduction to statistical learning. – New York : springer, 2013. – Т.112. – №.1. <https://thuvienso.hoasen.edu.vn/bitstream/handle/123456789/10495/Contents.pdf?sequence=1>

12. Tibshirani R. J., Efron B. An introduction to the bootstrap // Monographs on statistics and applied probability. – 1993. – Т.57. – №.1. – С.1-436. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5765871610321f83521c283980e47755fe793d1c>

13. Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions // Advances in neural information processing systems. – 2017. – Т. 30. <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>

МЕДИЦИНАЛЫҚ-БИОЛОГИЯЛЫҚ ЗЕРТТЕУЛЕРДЕГІ ӘСЕРДІҢ КӨРІНІСІНЕ КӨПТЕГЕН ФАКТОРЛАРДЫҢ ЖИЫНТЫҚ ӘСЕРІН БАҒАЛАУ ӘДІСТЕРІ

**Сабилов Ж. Б.¹, Отаров Е. Ж.¹, Алексеев А. В.¹, Гребенева О. В.¹,
Шадетова А. Ж.¹.**

¹«Еңбек гигиенасы және кәсіптік аурулар ұлттық орталығы» КеАҚ (100017, Қазақстан Республикасы, Қарағанды қ., Мұстафин к. 15, e-mail: info@naoncgt.kz)

1. Сабилов Ж.Б., PhD, «Еңбек гигиенасы және кәсіптік аурулар ұлттық орталығы» КеАҚ кәсіптік патология ғылыми зертханасының меңгерушісі, e-mail: audacious_zap@mail.ru

2. Отаров Е.Ж., e-mail: otarov_kgmu@mail.ru

3. Алексеев А.В., e-mail: alekseev@qmu.kz

4. Гребенева О.В., e-mail: ol_grebeneva@bk.ru

5. Шадетова А.Ж., e-mail: alma7722@mail.ru

Тұжырым

Медициналық-биологиялық зерттеулерде көптеген факторлардың – сандық та, сапалық та – белгілі бір биомедициналық әсерге жиынтық ықпалын сандық тұрғыдан бағалау жиі қажет болады. Бұл мақалада деректерді қалыпқа келтіру, факторлардың салмақтарын анықтау, ақпаратты біріктіру және факторлардың жиынтық әсерін бағалау үшін модельдер құру тәсілдері қарастырылады. Сараптамалық және деректерге негізделген әдістер, сондай-ақ олардың денсаулық жағ-

дайын болжау, тәуекелді бағалау және биомедицинадағы шешім қабылдау міндеттеріндегі қолданылуы талқыланады.

Түйінді сөздер: медициналық-биологиялық зерттеулер, интегралдық көрсеткіш, көпөлшемді талдау, статистикалық модельдер.

МЕТОДЫ ОЦЕНКИ СУММАРНОГО ВЛИЯНИЯ МНОЖЕСТВА ФАКТОРОВ НА ПРОЯВЛЕНИЕ ЭФФЕКТА В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

**Сабиров Ж.Б.¹, Отаров Е.Ж.¹, Алексеев А.В.¹, Гребенева О.В.¹,
Шадетова А.Ж.¹.**

¹НАО «Национальный центр гигиены труда и профессиональных заболеваний» (100017, Республика Казахстан, г.Караганда, ул.Мустафина 15, e-mail: info@naoncgt.kz)

1. Сабиров Ж.Б., PhD, заведующий научной лаборатории профессиональной патологии НАО «Национальный центр гигиены труда и профессиональных заболеваний», e-mail: audacious_zap@mail.ru
2. Отаров Е.Ж., e-mail: otarov_kgmu@mail.ru
3. Алексеев А.В., e-mail: alekseev@qmu.kz
4. Гребенева О.В., e-mail: ol_grebeneva@bk.ru
5. Шадетова А.Ж., e-mail: alma7722@mail.ru

Резюме

В медико-биологических исследованиях часто требуется количественно оценить совокупное воздействие множества факторов - как количественных, так и качественных - на определённый биомедицинский эффект. В данной статье рассматриваются подходы к нормализации данных, определению весов факторов, агрегированию информации и построению моделей для оценки суммарного влияния факторов. Обсуждаются как экспертные, так и основанные на данных методы, а также их применение в задачах прогнозирования состояния здоровья, оценки риска и принятия решений в биомедицине.

Ключевые слова: медико-биологические исследования, интегральный показатель, многомерный анализ, статистические модели.