ABSTRACT

for the dissertation of **Madina Utegenovna Suleimenova** entitled
“Machine learning in predicting the risk of developing cardiovascular disease**”**
submitted for the degree of **Doctor of Philosophy (PhD)**
in the educational program **8D06114 – Artificial Intelligence in Medicine**

Relevance of the research topic. Cardiovascular diseases are among the leading causes of mortality and disability both in Kazakhstan and worldwide and have a substantial impact on life expectancy. According to WHO data, Kazakhstan belongs to the regions with a high risk of cardiovascular disease (CVD). In Kazakhstan, life expectancy declined to 70.23 years in 2021 due to increased morbidity and mortality, whereas in 2020 it was 71.37 years. Over the past two years, average life expectancy decreased by 2.95 years, returning to the level of 2013. These indicators demonstrate the need to identify markers that play a decisive role in predicting early cardiovascular aging and preventing adverse outcomes. The use of artificial intelligence (AI) methods, such as machine learning and modeling, makes it possible to identify individual risks and predict the probability of CVD development, which is becoming an important part of modern medicine because it accelerates analysis, forecasting, and other processes.

Machine learning offers a flexible framework for prediction; however, its clinical implementation requires rigorous validation and interpretation. In our study, model quality was assessed not only by ROC-AUC, but also comprehensively, taking into account internal validation without data leakage, calibration assessment, clinical utility based on Decision Curve Analysis, and interpretation using coefficients, permutation importance, and SHAP. A significant methodological challenge arises when the development sample is small: without careful verification, performance may appear high, feature selection may be unstable, and model robustness may remain uncertain.

Assessment of CVD risk is complicated by multimorbidity, heterogeneity of metabolic profiles, and population-specific baseline risk, which may affect both discrimination and, above all, calibration of probabilistic predictions. Many traditional risk scores were developed on mixed-age cohorts and may demonstrate limited transportability when applied to older patients or new settings.

Thus, in routine practice, the usefulness of a prognostic tool is determined not only by its ability to stratify patients according to risk level, but also by how well the predicted probabilities are calibrated for threshold-based decision-making and how interpretable the model’s behavior is from a clinical perspective.

Based on the above, the aim and objectives of this study were formulated.

Aim of the dissertation. To develop an intelligent system for predicting the risk of cardiovascular disease in individuals aged 65 years and older using a reproducible end-to-end process and machine learning models.

Objectives of the study. To achieve this aim, the following objectives were set:

To investigate machine learning models for determining the prognostic risk of disease development in medicine.

To form a development sample of individuals aged 65+ with clinical and biochemical indicators and perform data preprocessing in the form of a reproducible pipeline without data leakage.

To conduct a comparative analysis and train machine learning models using nested cross-validation and obtain out-of-fold (OOF) probabilities.

To evaluate model quality in terms of discrimination and probabilistic predictions, assess probability calibration, interpret predictions, and determine the stability of feature importance using permutation importance.

To justify decision thresholds and risk category boundaries using Decision Curve Analysis (DCA) and to construct a Low/Intermediate/High scale on the development sample.

To perform external validation of the developed scale on an independent cohort, assess model transportability, and carry out probability recalibration.

Object and subject of the study. The object of the study is tabular multimodal data, including clinical and biochemical data, from individuals aged 65 years and older in Almaty. The subject of the study is machine learning models for predicting the risk of cardiovascular disease based on clinical and biochemical data.

Research methods. The study involved collection and preprocessing of multimodal data, correlation analysis, and construction of predictive models (Elastic Net Logistic Regression, Random Forest), as well as internal and external validation (n = 695) with hyperparameter optimization. Model quality was assessed using ROC-AUC, PR-AUC, Brier score, LogLoss, and calibration; decision thresholds were justified using DCA; and factors were interpreted using odds ratios, permutation importance, and SHAP.

Scientific novelty:

• A reproducible end-to-end process was proposed for constructing an interpretable probabilistic risk scale on tabular multimodal data, combining standardized preprocessing without data leakage, unbiased assessment of predictive performance, and verification of the correctness of probabilistic predictions.

• For the first time, a probabilistic CVD risk scale for individuals aged 65+ was proposed, based on a combination of clinical and biochemical markers, for practical patient stratification and decision support. The tool provides the risk probability p and a category (Low/Intermediate/High) for patient routing.

Main provisions submitted for defense:

The proposed reproducible end-to-end machine learning process predicts the risk of CVD development on a small sample with an accuracy of 93%.

The predictive ability of the developed model and the quality of probabilistic prediction (Brier/LogLoss) were confirmed by external validation on an independent cohort (n = 695), taking into account differences in baseline population risk (ROC-AUC ≈ 0.835). The interpretable probabilistic scale CARDIO RISK KZ is applicable for assessing the risk of cardiovascular disease in individuals aged 65 years and older.

Practical and scientific-technical significance. The developed interpretable probabilistic scale using machine learning models makes it possible to assess the potential risks of cardiovascular disease development in groups of patients aged 65 years and older living in Almaty who have similar clinical and biochemical data. The use of this scale will support effective preventive and therapeutic decisions within a personalized approach.

Degree of reliability and approbation of the results.

The dissertation topic has resulted in 3 articles published in journals indexed in Scopus, 2 articles published in scientific journals recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan, as well as 2 certificates of entry in the State Register of Rights to Copyright-Protected Objects.

Articles published in journals indexed in Scopus:

Suleimenova, M., Abzaliyev, K., Abzaliyeva, S., & Nassyrova, N. (2026). An Interpretable and Externally Validated Model for Cardiovascular Disease Risk Assessment in Older Adults. *Applied Sciences*, *16*(8), 3903. https://doi.org/10.3390/app16083903

Abzaliyev, K., Suleimenova, M., Chen, S., Mansurova, M., Abzaliyeva, S., Manapova, A., Kurmanova, A., Bugibayeva, A., Sundetova, D., Bitemirova, R., Baizhigitova, N., Abdykassymova, M., & Sagalbayeva, U.

Predicting Cardiovascular Aging Risk Based on Clinical Data Through the Integration of Mathematical Modeling and Machine Learning. Applied Sciences, 2025, 15(9), 5077. https://doi.org/10.3390/app15095077 (Scopus: Q2)

Abzaliyev, K., Suleimenova, M., Abzaliyeva, S., Mansurova, M., Shomanov, A., Bugibayeva, A., Tolemisova, A., Kurmanova, A., & Nassyrova, N. Multidimensional Visualization and AI-Driven Prediction Using Clinical and Biochemical Biomarkers in Premature Cardiovascular Aging. Biomedicines, 2025, 13(10), 2482. https://doi.org/10.3390/biomedicines13102482 (Scopus: Q1)

Articles published in journals included in the list of scientific publications recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan for publishing the main results of scientific work:

Suleimenova, M., Chen Siming, Shomanov, A., Abzaliyev, K., Kurmanova, A., Abzaliyeva, S., Abdykassymova, M., Sagalbayeva, U., Bitemirova, R., Sundetova, D., Bugibayeva, A. AI in Identifying Aging Markers. Phthisiopulmonology, No. 01(43), 2024, pp. 130–139. https://www.doi.org/10.26212/2227-1937.2024.50.73.018

Suleimenova M.U., Mansurova M.E. Artificial Intelligence as a Tool for Investigating Aging Markers. Bulletin of the National Engineering Academy of the Republic of Kazakhstan, No. 2, 2024.

Certificate of entry in the State Register of Rights to a copyright-protected object for a computer program entitled:

"CARDIO RISK KZ: Intelligent System for Calculating Individual Risk of Premature Cardiovascular Aging." No. 64683, 2025.

"Program for Probabilistic Assessment of Cardiovascular Disease Risk in Individuals Aged 65 Years and Older." No. 68445, 2026.

Personal contribution of the researcher.

In accordance with the aims and objectives of the dissertation, all tasks were completed. Based on the clinical and biochemical data of individuals aged over 65, a reproducible pipeline was developed, and Elastic Net logistic regression and Random Forest models were built without data leakage. The models were evaluated using nested cross-validation, and ROC-AUC, PR-AUC, Brier, LogLoss, and calibration metrics were calculated. Based on Decision Curve Analysis, risk thresholds (t_low = 0.23, t_high = 0.40) were determined and a Low / Intermediate / High scale was constructed. The developed scale was tested on an independent external cohort (n = 695), and the results were implemented in the form of a software calculator.

Scope and structure of the dissertation.

The dissertation is written in Kazakh. The work consists of an introduction, 4 chapters, and a conclusion. The total volume of the dissertation is 86 pages and includes 24 figures, 10 tables, and 53 references.

The introduction substantiates the relevance of the study, demonstrates the need to develop clinically applicable and interpretable machine learning models for predicting the risk of cardiovascular disease in individuals aged 65 years and older, and presents the aim, objectives, and methods of the dissertation, the scientific novelty of the results obtained, the theoretical and practical significance of the work, and the list of publications.

The first chapter provides a literature review on the current state of research into cardiovascular diseases, cardiovascular aging, and risk factors in individuals over 65 years old. It also analyzes traditional cardiovascular risk assessment scales and machine learning methods, highlighting their advantages and limitations. The relevance of developing new interpretable and clinically convenient approaches to risk assessment for the elderly population is substantiated.

The second chapter describes the materials and methods of the study. It covers the preparation of the development cohort (n = 100) and the external validation cohort (n = 695), preprocessing, selection of clinical and laboratory indicators, correlation analysis, and multidimensional visualization methods. In addition, it presents the stages of constructing machine learning models, nested cross-validation, obtaining out-of-fold predictions, applying interpretation methods, determining risk thresholds based on Decision Curve Analysis, conducting external validation, and developing the digital prototype.

The third chapter presents the results of developing and evaluating models for predicting cardiovascular disease risk. Logistic regression, Random Forest, and other approaches were compared, after which the final model for clinical use was selected as the most interpretable and robust. Model quality was assessed using ROC-AUC, PR-AUC, Brier score, and LogLoss; calibration results were analyzed; and risk thresholds t_low = 0.23 and t_high = 0.40 were justified using DCA. In addition, the results of independent external validation are presented, and the CARDIO RISK KZ intelligent system for assessing cardiovascular disease risk in individuals aged 65 years and older is proposed.

The fourth chapter discusses the study results. The scientific and practical significance of the developed cardiovascular disease risk prediction model is substantiated, and the choice of clinical logistic regression as the final model is explained. The roles of family history, body mass index, physical activity, and diabetes mellitus are identified as the main predictors. The chapter also describes the possibilities of applying the developed CARDIO RISK KZ scale in clinical practice, as well as the study limitations and future directions for development.

The conclusion presents the main results and conclusions of the dissertation.